SPATIAL
REGRESSION ANALYSIS
ON THE PC:
SPATIAL STATISTICS
USING MINITAB

Institute of Mathematical Geography

by

Daniel A. Griffith

Department of Geography
Syracuse University

February 1, 1989

ISBN: 1-877751-24-3

THIS WORKBOOK IS DEDICATED TO

MICHELE GRIFFITH,

WHO HAS BROUGHT SO MUCH JOY TO MY LIFE,

ESPECIALLY DURING THOSE DIFFICULT YEARS

WHEN I WAS INITIATING MY CAREER.

PREFACE

Applied statistical and econometric analysis in Regional Science and Geography frequently deals with data collected for aggregate spatial units of observation. Typically these data are affected by a variety of measurement problems, resulting in spatial dependence and spatial heterogeneity. However, most empirical work with spatial data series fails to take this complication into account, even though awareness of problems caused by spatial structure and spatial dependence, and their impact on the validity of traditional statistical methods, is not recent. In fact, an important series of methodological developments has occurred in Regional Science and Geography based on the need to deal with the special nature of spatial data sets. Despite these substantial methodological advances, the actual application of appropriate spatial techniques in situations where they are likely to be relevant has been rather limited, even within the academic community of regional scientists and geographers. This point has been emphasized in a survey, reported in Anselin and Griffith (1988), of articles in journals that typically are receptive to the publication of analytical developments in spatial statistics and spatial econometrics: only about 7% of those studies examining spatial data series actually took spatial effects into account. This finding suggests that further methodological developments alone are not sufficient to affect the dissemination of spatial techniques to both empirical academic work and to the applied practice of professional geography and regional science. A concerted effort must be forged in order to bring about an increased availability of readily implementable software, which can be easily integrated within currently existing and widely used packages. Moreover, although methodological results achieved in the fields of spatial statistics and spatial econometrics have been substantial, the dissemination from the research community to the applied world has been virtually nonexistent.

The recognition of spatial effects has resulted in a large number of specialized analytical techniques and spawned the separate fields of spatial statistics and spatial econometrics. For definitional purposes, we consider the distinguishing characteristic for spatial statistics to be its data-driven orientation, whereas spatial econometrics is viewed as being essentially model-driven. In contrast to this, there is an almost total ignorance of spatial effects in the mainstream statistical and econometric literature. For example, in most recent statistics and econometrics textbooks hardly any mention of the spatial effects problem can be found. This lack of consideration of the complications caused by spatial effects is also reflected in the practice of statistics and econometrics, in that spatial techniques are absent from standard regression packages, such as SAS, MINITAB, SPSSX, or BMDPP commonly used by regional scientists and economic geographers.

As a consequence, there has been an almost total lack of diffusion of the spatial statistical and spatial econometric techniques to empirical work in Regional Science and Geography. Even though the type of data used in this empirical work, such as aggregate observations for states, counties, or census tracts, is likely to reflect spatial structure and to be subject to various spatial spill-over effects, the spatial autocorrelation and spatial heterogeneity which may result are almost completely ignored.

Collaborative research addressing this problem was proposed to and has been funded by the Geography and Regional Science Program of the National Science Foundation (Research Grants # SES-8722086 and # SES-8721875). It is an integrative effort between a spatial statistician and a spatial econometrician, and in part has the distinct purpose of transcending theoretical work by emphasizing the development of accessible software and its dissemination to the profession. This joint venture maintains that the integration of new software within existing commercially available packages and GIS systems will result in an increased accessibility of the proper spatial techniques to the professional Geography and Regional Science community, and should allow for an improvement in the general quality of modeling efforts that can be carried out. It also is hoped that products from this collaboration would considerably increase the level of sophistication and the degree of accuracy of the technical analyses achieved by applied spatial analysts. Accordingly, one of the broad objectives of this research endeavor is to develop and implement a set of tools for the improved dissemination and accessibility of methodological results for use in applied work in empirical Regional Science, professional Geography and spatial analysis of Geographic Information Systems. In essence, this should result in a set of procedures and techniques that the applied spatial analyst can effectively utilize to deal with the problems of spatial dependence and spatial heterogeneity. These procedures and techniques are made available in a user-friendly format and presented through a series of discussion papers, of which the first two are Spatial Regression Analysis on the PC: Spatial Statistics Using MINITAB (#1, by Griffith), and Spatial Regression Analysis on the PC: Spatial Econometrics Using GAUSS (#2, by Anselin). These two publications are designed to present translations of methodological tools into computer code that can be easily and quickly integrated into existing, widely available computer packages. This integration is in the form of macros and subroutines, which essentially remain transparent to the user. The specific microcomputer package treated in Discussion Paper #1 is MINITAB-PC (its code can easily be uploaded to a mainframe housing MINITAB), whereas the designated microcomputer package of GAUSS is treated in Discussion Paper #2.

Wide dissemination of the computer code constructed under the sponsorship of this research project is to be achieved through the availability of this series of discussions papers, and the hosting of workshops. These discussion papers have been designed to accompany a workshop that introduces users to hands-on experience with microcomputer versions of the code. They are suited for use as laboratory manuals for upper-level undergraduate or graduate courses in geographical data analysis or applied regression analysis, too. The workshop was first held at the Association of American Geographers (AAG) Baltimore meeting (March, 1989), under the joint auspices of the AAG Mathematical Models & Quantitative Methods Specialty Group and the NSF National Center for Geographic Information and Analysis (NCGIA) outreach program. In addition, this Workshop also has been supported by the MINITAB demonstration program. The discussion papers list the different sets of computer code, describe how to integrate this code into the two packages, and provide test examples for determining whether or not correct implementation has occurred (interested persons can obtain digital copies of the computer code either through the workshop or by completing the mail-order acquisition form found at the end of this discussion paper).

Although the routines described in these two discussion papers have been tested, both before and during the workshop, and used by their respective authors and by people elsewhere, no warranties, expressed or implied, are made by University of California/Santa Barbara, Syracuse University, or the authors that the computer code or documentation are free of error. Furthermore, this software is not warranted for correctness, accuracy, functioning of the macros and related routines, or fitness for a task. Users rely on the results of the routines solely at their own risk; no responsibility is assumed in connection therewith.

It is hoped that this pair of discussion papers will begin to provide some specific advice and guidance to practitioners, couched in user-friendly commercial software. The first step in analyzing a spatial data series should be to assess the sources, nature and degree of prevailing spatial effects. The software outlined in these discussion papers is intended to facilitate this, and when needed, to allow the implementation of the proper modeling techniques.

Daniel A. Griffith
Syracuse, New York

and

Luc Anselin
Santa Barbara, California

February 1, 1989

TABLE OF CONTENTS

TABLE OF CONTENTS, continued

CHAPTER 1.
INTRODUCTION, BACKGROUND, AND CAVEATS

This Discussion Paper presents computer code that exploits the MINITAB commercial software package in order to substantially reduce the tedium of numerically intensive, complex spatial statistical calculations, as well as convert the accompanying massive matrix manipulations into transparent operations. Its preparation is part of an ongoing collaborative effort to develop, popularize, and disseminate a comprehensive set of user-friendly procedures for undertaking spatial statistical analyses. These modified statistical techniques are necessary because, on the one hand, spatial statistics addresses situations involving geographic data series for which standard statistical tests and estimation procedures often are invalid, and, on the other hand, no present standard commercial statistical package provides these specialized routines in a user-friendly format or environment. In part the impetus for this research came from the success with which results for a simultaneous autoregressive (SAR) model can be teased out of standard commercial statistical software [see Griffith (1988)]. Findings for previously published as well as new numerical examples are included; the data for these examples appear in Appendices 2-A, 2-B, and 5-A. Prominent statistical issues for this approach to geographic data analysis include model specification, model estimation, and model diagnostics. Questions concerning model specification will not be addressed here; the single model of concern is the SAR specification. Maximum likelihood estimation will be employed, with appropriate adjustments being made by substituting degrees of freedom for sample size in order to secure unbiased estimates. And, diagnostics will be focused on; especially the violation of independent errors in ordinary least squares (OLS) regression solutions, and the assumption of normally distributed errors will be scrutinized.

1.1. The MINITAB commercial software package

Interactive MINITAB is a user-friendly, widely available, inexpensive, and general purpose statistical analysis package. It encompasses the entire range of customary univariate statistical techniques, and recent updates have begun to include some of the favored multivariate procedures. It has a PC version that performs quite efficiently, is very powerful, and is attractively flexible. It operates on IBM-compatible machines, with DOS version 2.0 or later, requires a minimum 512K RAM (there must be at least 450K available after system configuration), is best used with a 10 Megabyte hard disk and one double-sided, double-density diskette drive (a minimum of two double-sided, double-density diskette drives are needed if a hard disk is unavailable), a monochrome or color graphics monitor (80 character width), and preferably a math coprocessor chip. Once it is installed on a machine, running it usually is initiated with the following command (↵ denotes a carriage return):

```
C>MINITAB ↵
```

where C> is the system prompt; customized paths may alter this command to some degree.

MINITAB stores, manages, and manipulates data in three different forms. Data may be stored in constants, which are denoted by K1, K2, ..., K100 (a maximum of 100), in vectors or columns, which are denoted by C1, C2, ..., C100 (a maximum of 100), or in matrices, which are denoted by M1, M2, ..., M15 (a maximum of 15). A list of consecutively numbered constants, columns, or matrixes may be abbreviated by using a dash. Besides these machine-specific limitations, MINITAB-PC has the following restrictions:

```
Worksheet size:    16,000
Maximum width of "A" format variables:  4 with READ
                                       80 with SET
Maximum number of characters per line with formatted
  READ:  256
  WRITE: 256
Defaults for:
  output width: 79
  output height: 24
Maximum number of open files (macros plus outfile):  6
To temporarily exit Minitab to DOS, type:
  MTB > SYSTEM
To return to your Minitab session type "EXIT".
Minitab supports DOS 2.0 file paths up to 30 characters.
```

These parameters constrain the size of problems that can be handled by this PC software package; mainframe versions possess less binding limitations and hence are capable of handling much larger problems. Furthermore, the output height of 24 means that after each 24 lines are displayed on the CRT screen, the prompt CONTINUE? appears; the simplest response to this prompt is a carriage return.

MINITAB commands are entered after the prompt MTB>. MINITAB embraces about 180 different commands. A command always begins with a command word, which usually is followed by a list of arguments; an argument is either a column C_, a constant K_, a matrix M_ (the underline signifies the position of an appropriate numeral), a number or a file name (inside single quotation marks). When columns, constants, or matrices are reused, all previous contents are erased and the new contents are inserted. For shorthand purposes, only the first four letters of a command and/or argument are needed. Annotations may be included by preceding remarks with the MINITAB command NOTE. Each command must start on a new line, with continuations indicated by including an ampersand (&) as the last character of the line that is to be continued (MINITAB automatically inserts ampersands when writing output to files). Subcommands are available for some MINITAB commands; these subcommands are used to evoke special options. Commas cannot appear within a number; only consecutive numerals can be constructed, except for the appearance of a decimal point.

A single user-defined macro file is necessary for implementation of the spatial statistical analysis software outlined in this Discussion Paper. This macro stipulates the problem size, variable definitions, and data input, and as such requires proficiency with only a few MINITAB commands. In order to maintain consistency with the test materials presented in this workbook, this file is expected to be named START.___ (the underline signifies a three-alphanumeric extension; see Section 2.1 for several examples). The first file line should contain the MINITAB command NOECHO, which turns off the echo printing of MINITAB commands; this command suppresses the display of data that are read, and MINITAB commands that are executed, on the CRT screen. The next two lines should contain initial problem size parameters, and involves using the LET command, which allows MINITAB constants, columns, and matrices to be assigned values; LET K1 = the number of areal units, and LET K2 = the number of X variables. The software described in this Discussion Paper is designed for problems where K1 ≤ 50, and K2 ≤ 10. The fourth line should contain the READ statement for the binary connectivity matrix M1, which will have dimensions K1-by-K1; its format is READ 'file name' K1 K1 M1, and it expects the first K1 entries to appear in free-format style on the first line of the file in question (clearly this file will contain at least K1 lines). In most cases the fifth line should be LET K3 = K2 + 1, and increments the number of predictor variables by one in order to take into account the regression intercept term; in analysis of variance this declaration is LET K3 = K2, while it may remain undefined for the trend surface routines. The sixth line should contain the READ statement for the data file, defining that variable to be designated Y as C1, and those variables to be designated the Xs as C2-CK3; READ 'file name' C1 C2-CK3. As Chapter 2 illustrates, considerable latitude is available here, with the designation of variables being made in whatever order is compatible with their appearance in the data file. The last line of this file is END, which terminates execution of the macro. Once all of the MINITAB macros have been executed (see Chapter 6 for a summary of the necessary macro sequences for various spatial statistics routines), MINITAB is terminated by entering the command STOP.

Output appearing on the CRT screen also can be captured in a file by entering OUTFILE 'file name' before the execution of macros. This MINITAB command is extremely valuable for collecting results from a final analysis; it creates an ASCII file that can be easily read into many word processing software packages.

1.2. The Moran Coefficient and the Geary Ratio

One of the first diagnostics utilized in these spatial statistics routines is a test for spatial autocorrelation in regression residuals. The Moran Coefficient (MC) has been selected for this test. Griffith (1987, pp. 48-49) shows that this index can be calculated using the following set of

MINITAB commands, where there are K1 areal units, the connectivity matrix is housed in MINITAB matrix M1, and the geographic data is housed in MINITAB column C1:

```
CENTER C1 C75;
LOCATION.
MULT M1 C75 C76
LET K2 = SUM(C75*C76)
LET K3 = SUM(C75**2)
SET C77
K1(1)
END
MULT M1 C77 C78
LET K4 = SUM(C78)
LET K5 = (K1/K4)*(K2/K3)
PRINT K5
END
```

The MINITAB constant K5 is MC.

Another spatial autocorrelation index option is the Geary Ratio (GR), which entails squared paired comparisons calculations. But for any geographic variable, the paired comparison $(x_i - x_j)$ can be rewritten such that

$$[(x_i - \bar{x}) - (x_j - \bar{x})]^2 = (x_i - \bar{x})^2 - 2(x_i - \bar{x})(x_j - \bar{x}) + (x_j - \bar{x})^2 \quad .$$

This middle term appears in the numerator of MC. Therefore, GR may be calculated with the set of MINITAB commands

```
CENTER C1 C75;
LOCATION.
MULT M1 C75 C76
LET K2 = SUM(C75*C76)
LET K3 = SUM(C75**2)
SET C77
K1(1)
END
MULT M1 C77 C78
LET K4 = SUM(C78)
LET K5 = (K1/K4)*(K2/K3)
LET K6 = SUM(C78*(C75**2))
LET K7 = ((K1-1)/(2*K4))*(2*K6/K3) - ((K1-1)/K1)*K5
PRINT K7
END
```

The MINITAB constant K7 is GR. Consequently, in order to calculate GR, one must calculate MC; moreover, GR is a function of MC. In addition, Cliff and Ord (1981) suggest that the statistical properties of MC are better behaved than are their GR counterparts.

The MINITAB code presented in this Discussion Paper assumes that the matrix M1, or the geographic connectivity matrix, is binary (0,1) and symmetric. If one wishes to go beyond the classical statistical distribution theory of MC, and use other forms of a binary connectivity matrix (only zeroes and ones are present here), then several slight adjustments to the ensuing MINITAB code will be required. For mathematical purposes, this matrix will be denoted as **C**.

1.3. Types of autoregressive models

Various types of autocorrelation models have been used successfully with geographic data series. One class of these models spotlights error; the error component of a model is spatially autocorrelated. Conceptually the simplest structure for this autocorrelation is a moving average (MA) model, whose error covariance matrix may be written as $(\mathbf{I} - \rho\mathbf{C})\sigma^2$. The second simplest structure found in the literature is a conditional autoregressive (CAR) model, whose error covariance matrix may be written as $(\mathbf{I} - \rho\mathbf{C})^{-1}\sigma^2$. Extensive use of either of these models for regression analysis purposes requires a decomposition of the matrix $(\mathbf{I} - \rho\mathbf{C})$; two possibilities are an eigenfunction decomposition, and a Cholesky decomposition. Findings reported by Griffith (1988c) imply that final regression analysis results are conditional with respect to the selection of a particular decomposition. Thus, there is a need for further meticulous study of these two models before implementation procedures for them are disseminated. A second class of these models is characterized by the autoregressive response specification, which presumably will be the topic treated in one of the next releases in this Discussion Paper series.

One model whose estimation theory is sufficiently developed to support deployment in the area of applied spatial statistics is the simultaneous autoregressive (SAR) model. The error covariance matrix for this autocorrelation structure may be written as $[(\mathbf{I} - \rho\mathbf{W})^t(\mathbf{I} - \rho\mathbf{W})]^{-1}\sigma^2$. Here the previous connectivity matrix **C** is replaced with a stochastic matrix **W** (all elements of this matrix should be non-negative, and the row sums of this matrix should be unity), frequently called a weights matrix. Unless user intervention occurs (see the Upton and Fingleton example of Chapter 5), this stochastic matrix is automatically calculated from the initial connectivity matrix; all SAR model routines presented subsequently assume this latter matrix. Griffith (1988) presumes the use of this matrix in model specifications; in one instance, however, Upton and Fingleton (1985) employ a weights matrix that fails to be stochastic. In order to accommodate their example, primarily for illustrative purposes, the spatial statistics software of this Discussion Paper has been generalized beyond that of Griffith (1988); unfortunately, this generalization increases the numerical intensity of the procedures, and hence increases the required computer execution time. Briefly, this extension means that the regression intercept variable must be written as $(\mathbf{I} - \rho\mathbf{W})\mathbf{1}$ rather than $(1 - \rho)\mathbf{1}$.

Algebraic expansion of the matrix product $(\mathbf{I} - \rho\mathbf{W})^t(\mathbf{I} - \rho\mathbf{W})$ to $[\mathbf{I} - \rho(\mathbf{W}^t + \mathbf{W}) + \rho^2\mathbf{W}^t\mathbf{W}]$ emphasizes the close relationship between CAR and SAR model specifications. If ρ is quite small, then ρ^2 will be very close to zero, and this expansion becomes analogous to a CAR specification of the form $[\mathbf{I} - \rho(\mathbf{W}^t + \mathbf{W})]$. This linkage implies that if $\hat{\rho}$ for a SAR model is small, then a CAR specification should be explored.

As is well known [see Griffith (1988b)], estimation of the SAR model requires maximum likelihood techniques. This optimization problem reduces to one of

$$\text{MIN:}\quad \det|(\mathbf{I} - \rho\mathbf{W})^t(\mathbf{I} - \rho\mathbf{W})|^{-1/n}(\mathbf{Y} - \mathbf{X}\beta)^t(\mathbf{I} - \rho\mathbf{W})^t(\mathbf{I} - \rho\mathbf{W})(\mathbf{Y} - \mathbf{X}\beta) \quad ,$$

where $\det|(\mathbf{I} - \rho\mathbf{W})^t(\mathbf{I} - \rho\mathbf{W})|^{-1/n} = [\prod_{i=1}^{i=n} (1 - \rho\hat{\lambda}_i)]^{-2/n}$ is the Jacobian term, with λ_i being the i-th eigenvalue of matrix **W**. Extraction of the eigenvalues of matrix **W** combined with the minimization of a nonlinear function are what make this maximum likelihood estimation calculation so numerically intensive. By construction the principal eigenvalue, λ_1, always equals unity; this result constrains the spatial autocorrelation parameter such that $|\rho| < 1$, and is the main reason why Ord (1975) contends that the use of a stochastic version of the connectivity matrix leads to a natural interpretation for the spatial autocorrelation parameter. Because this minimization problem involves the multiplication of the Jacobian term, which is a constant for any value of $\hat{\rho}$, times a summation, this constant can be distributed over the sum of squares term represented by the matrix product $(\mathbf{Y} - \mathbf{X}\beta)^t(\mathbf{I} - \rho\mathbf{W})^t(\mathbf{I} - \rho\mathbf{W})(\mathbf{Y} - \mathbf{X}\beta)$. After Griffith (1988a), this Jacobian term may be rewritten as

$$1/(\exp\{[\sum_{i=1}^{i=n} \ln(1 - \rho\lambda_i)]/n\})^2 \quad ,$$

which is a mean as well as a squared number. This rendition of the Jacobian term allows the minimization problem to be translated into a regression problem involving the following model expression:

$$(\mathbf{I} - \rho\mathbf{W})\mathbf{Y}/\exp\{[\sum_{i=1}^{i=n} \ln(1 - \rho\hat{\lambda}_i)]/n\} = (\mathbf{I} - \rho\mathbf{W})\mathbf{X}\beta/\exp\{[\sum_{i=1}^{i=n} \ln(1 - \rho\hat{\lambda}_i)]/n\} + \xi \quad .$$

This is the estimating equation employed in this Discussion Paper.

At times the nonlinear estimation procedure for calculating $\hat{\rho}$ fails to converge. Convergence may not be attained when the geographic data series under study is nonstationary (this is not the only reason for convergence failure, though); one reason for the presence of nonstationarity is a spatial drift in the parameters of the regression model. Probably the simplest form of this drift is a spatially varying regression intercept term. Even if no spatial autocorrelation is latent in the geographic data series, if the intercept term drifts, and this drift is ignored in the model specification, then the regression residuals may exhibit autocorrelation. Such a drift can be removed by introducing sundry powers of absolute coordinate measures into the matrix **X** of predictor variables. In essence, if $\rho = 1$, then an areal unit value is exactly the average of the values taken on by those areal units to which it is juxtaposed. This scenario is precisely what a linear trend surface describes! Inclusion of the trend surface terms will remove the detected spatial autocorrelation in the residuals, and hence in the absence of spatial autocorrelation will move $\hat{\rho}$ close to zero; the new geographic distribution of residuals now is stationary.

1.4. Statistical properties of OLS versus SAR

Classical statistics has provided thorough documentation of the desirable properties of the ordinary least squares (OLS) solution for a regression model. In an ideal situation in which the Gauss-Markov Theorem holds, the quality of the OLS regression coefficients **b** is such that they are the best (minimum variance) linear unbiased estimators; moreover, they are consistent (they differ from the true parameter values by a decreasingly very small amount as n becomes large), sufficient (they extract all of the relevant information about their respective population parameters that is contained in the original sample), unbiased (the mean of their sampling distribution equals the population parameter), and equivalent to the maximum likelihood estimators for a random effects model having normally distributed errors. The covariance of the estimators **b** is given by $(\mathbf{X}^t\mathbf{X})^{-1}\sigma^2$. The maximum likelihood estimate of the error variance σ^2 is asymptotically unbiased; its OLS estimate is unbiased.

Spatial statistics has found that if the OLS solution is used for a SAR model, then the regression coefficient estimators **b** are unbiased, but no longer consistent or sufficient. Mardia and Marshall (1984) show that the SAR estimators are weakly consistent, and Ord (1975) shows that the estimator for ρ is consistent. Griffith (1988a) shows that these SAR parameter estimates are sufficient. Upton and Fingleton (1985) note that the estimator for the spatial autocorrelation parameter will be biased, strictly due to the nature of the nonlinear estimation procedure involved. The covariance of the estimators **b** is given by $[\mathbf{X}^t(\mathbf{I} - \rho\mathbf{W})^t(\mathbf{I} - \rho\mathbf{W})\mathbf{X}]^{-1}\sigma^2$, which is the asymptotic covariance matrix, and is independent of the estimates of parameters ρ and σ^2; the covariance matrix for these latter two estimators is such that they are not independent of each other. The maximum likelihood estimate of the error variance σ^2 in the SAR model is asymptotically unbiased; it can be converted to an unbiased estimator using the same degrees of freedom principle as for conventional OLS estimation. The OLS mean square error for a SAR model tends to be severely biased, beyond what can be accounted for merely by making an adjustment with degrees of freedom.

1.5. What the SAR model estimation procedure accomplishes

Two prominent points advanced in the preceding section highlight why the specification of a SAR model can be worthwhile. First, the OLS mean square error estimate of σ^2 may be too large. When this value is inflated, it can result in (1) failure to reject the null hypothesis about a regression parameter when in fact that null hypothesis should be rejected (the Type I error probability is incorrect), and (2) underestimation of the R^2 value for the overall regression model. Second, the diagonal terms of the covariance matrix $[\mathbf{X}^t(\mathbf{I} - \rho\mathbf{W})^t(\mathbf{I} - \rho\mathbf{W})\mathbf{X}]^{-1}$ for **b** may be either too large or too small, since the spatial linear operator could either inflate or deflate (depending upon the nature of the latent spatial autocorrelation) individual regression coefficient standard errors. If these values are inflated, then again the Type I error probability may lead to an incorrect failure to reject H_0; the compound effect of a double inflation can be dramatic. If these values are deflated, then the null hypothesis about a regression parameter may be rejected when in fact it should not be. It is extremely unlikely that an inflation due to the incorrect OLS estimate of σ^2 would be exactly offset by a deflation due to the incorrect OLS estimate that substitutes the matrix $(\mathbf{X}^t\mathbf{X})^{-1}$ for $[\mathbf{X}^t(\mathbf{I} - \rho\mathbf{W})^t(\mathbf{I} - \rho\mathbf{W})\mathbf{X}]^{-1}$. Obviously the Type II error probabilities are impacted upon by these inflations and deflations, too. The magnitude of these effects increases in severity as $|\rho|$ approaches unity.

When spatial autocorrelation is present in the residuals, then the Gauss-Markov theorem no longer holds. Consequently, many standard diagnostics lose interpretability. Numbers are obtained for an OLS solution, but many of them become either meaningless or purely descriptive. Moreover, results become sample specific, for the method of least squares can be used to estimate the parameters of a linear regression model regardless of the form of the statistical distribution of residual errors. This OLS solution erodes the soundness of the inferential basis, though, if it is mistakenly determined when a SAR model specification should be posited.

CHAPTER 2.
OLS REGRESSION
WITH A TEST FOR SPATIAL AUTOCORRELATION

One of the first steps in a regression analysis is to obtain and then evaluate the traditional ordinary least squares (OLS) regression solution. In MINITAB this task is accomplished using the command REGRESS. Regression residuals are to be normally distributed, and are to lack spatial autocorrelation. These residuals are captured here using the MINITAB subcommand RESIDS with REGRESS. A Moran Coefficient (MC) is used to test for the presence of non-zero spatial autocorrelation, based upon the assumption of normality. A t-statistic is calculated for this observed MC value, and its corresponding degrees of freedom are determined. The t-statistic is preferred here for two reasons. First, usually the sample size in question is rather small (almost always less than 100; frequently less than 30). Second, the estimate s_{MC} is being used, rather than σ_{MC}. Critical values for this statistic, using a two-tailed testing situation and a critical region of 5%, are as follows:

degrees of freedom	critical value	degrees of freedom	critical value	degrees of freedom	critical value
1	± 12.71	14	± 2.14	27	± 2.05
2	± 4.30	15	± 2.13	28	± 2.05
3	± 3.18	16	± 2.12	29	± 2.04
4	± 2.78	17	± 2.11	30	± 2.04
5	± 2.57	18	± 2.10	35	± 2.03
6	± 2.45	19	± 2.09	40	± 2.02
7	± 2.36	20	± 2.09	50	± 2.01
8	± 2.31	21	± 2.08	60	± 2.00
9	± 2.26	22	± 2.07	80	± 1.99
10	± 2.23	23	± 2.07	100	± 1.98
11	± 2.20	24	± 2.06	200	± 1.97
12	± 2.18	25	± 2.06	500	± 1.96
13	± 2.16	26	± 2.06		

These tabulated t-statistic results have been gleaned from standard statistical tables, and are presented here with the threshold values for larger degrees of freedom (i. e., all critical values between 50 and 59 degrees of freedom equal 2.01; convergence on the normal distribution value occurs at 500 degrees of freedom).

Results for the randomization assumption are avoided here because, as Cliff and Ord (1981) note,

> ... under assumption R[andomization] we are assuming, falsely, that the sample residuals are uncorrelated under H_0, whereas by using assumption N(ormality) we are allowing for the correlation among the sample residuals under H_0. The amount and direction of the bias introduced when we make assumption R ... (p. 211)

A modified Shapiro-Wilk statistic is calculated to test for normality of the regression residuals, allowing one to assess how reasonable invoking this normality assumption is. This statistic is produced by first converting the residuals to normal scores (these are not z-scores) using the MINITAB command NSCORES, and then correlating the original residual values with their corresponding normal score values using the MINITAB command CORR. Critical values for this statistic, using a one-tail testing situation (a two-tailed test is inappropriate, since the population correlation is hypothesized to be unity) and a critical region of 5%, are as follows:

sample size	critical value	sample size	critical value	sample size	critical value
4	0.8734	20	0.9503	50	0.9764
5	0.8804	25	0.9582	60	0.9799
10	0.9180	30	0.9639	75	0.9835
15	0.9383	40	0.9715		

These tabulated correlation coefficient counterparts to the Shapiro-Wilk statistic have been taken from the MINITAB Reference Manual (1988, p. 63).

The Eire data (by counties) analyzed by Cliff and Ord (1981) is employed here for benchmarking purposes. Selected agricultural production data for the Mayaguez Region of Puerto Rico are used for illustrative purposes, too. All source code macros are housed on a diskette that should be located in Drive A, while the MINITAB software package is to be housed on a hard disk.

2.1. The Moran Coefficient for Regression Residuals

An initial macro file needs to be constructed that defines the sample size, the number of regressor variables, the connectivity matrix file, and the data set file (see Section 1.1 for a more complete discussion of this file). The code housed in this file has the following structure:

```
noecho
let k1 = sample size
let k2 = number of independent variables
read 'a:connectivity matrix file name' k1 k1 m1
let k3 = k2 + 1
read 'a:data set file name' c1 c2-ck3
end
```

The first MINITAB READ command accesses the connectivity matrix file from the diskette in Drive A; this two-dimensional array is stored in matrix M1. The second MINITAB READ command accesses the data set file from the diskette in Drive A; variable Y is stored in column C1, whereas the set of X variables is stored in columns C2-CK3. The corresponding files for the Eire data analyzed by Cliff and Ord, and the Puerto Rican data, respectively are a:START.TST and a:START.PR, and appear as follows:

```
a:START.TST

noecho
let k1 = 26
let k2 = 1
read 'a:eireconn.tst' k1 k1 m1
let k3 = k2 + 1
read 'a:eiredata.tst' c1 c2-ck3
end
```

```
a:START.PR

noecho
let k1 = 16
let k2 = 7
read 'a:mayaguez.con' k1 k1 m1
let k3 = k2 + 1
read 'a:mayaguez.dat' c50 c1 c2-ck3
end
```

One should notice that the second variable in the Puerto Rican data set is the Y variable, and that the first variable in this data set is being eliminated from the analysis.

The concern here is with spatially autocorrelated regression residuals. A standard OLS regression is obtained with the sequence of MINITAB command lines

```
LET K3 = K2 + 1
REGRESS C1 K2 C2-CK3;
RESIDS C50.
```

The column C50 contains the residuals, say vector **e**, that are to be analyzed. The Moran Coefficient for these regression residuals may be defined, using matrix notation, as follows:

$$\mathrm{MC} = (n/\mathbf{1}^t\mathbf{C1})*(\mathbf{e}^t\mathbf{Ce}/\mathbf{e}^t\mathbf{e}) \quad ,$$

where the vector **1** is n-by-1 and has all unity entries, and the superscript "t" denotes the operation of matrix transpose. The expected value of this coefficient is given by

$$\mathrm{E(MC)} = -\{n/[(n-k)*\mathbf{1}^t\mathbf{C1}]\}*\mathrm{tr}[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{CX}] \quad ,$$

where k equals one more than the number of regressor variables (this increment is necessary to include the intercept term), and tr denotes the matrix operation of "trace" (summing the diagonal entries). Matrix **X** consists of the set of regressor variables together with an initial vector of ones, which is the variable associated with the intercept term. The variance of this coefficient is given by

$$\mathrm{VAR(MC)} = \{n^2/[(\mathbf{1}^t\mathbf{C1})^2*(n-k)*(n-k+1)]\}*(S_1 + 2*\mathrm{tr}\{[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{CX}]^2\}$$
$$- \mathrm{tr}[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{C}+\mathbf{C}^t)^2\mathbf{X}] - 2*\{\mathrm{tr}[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{CX}]\}^2/(n-k)) \quad ,$$

where S_1 is defined as

$$(1/2) \sum_{i=1}^{i=n} \sum_{i=1}^{i=n} (w_{ij} + w_{ji})^2 \quad .$$

A t-statistic can be obtained for the observed MC by calculating

$$t \;=\; \{MC - E(MC)]/\sqrt{VAR(MC)} \quad ,$$

which should be distributed with (n - k) degrees of freedom.

The vector of ones, which will be column C49 here, to be included with the predictor variables in order to construct matrix **X** is obtained with the following MINITAB commands:

```
SET C49
K1(1)
END
```

Next, matrix **X**, which will be M2 here, is constructed with the MINITAB command

```
COPY C49 C2-CK3 M2
```

Matrix $\mathbf{X}^t$, which will be M3 here, is produced with the MINITAB command

```
TRANS M2 M3
```

The inverse matrix $(\mathbf{X}^t\mathbf{X})^{-1}$, which will be M5 here, is obtained with the following sequence of MINITAB commands:

```
MULT M3 M2 M4
INVERT M4 M5
```

Construction of matrix $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{CX}$, which will be M8 here, is achieved with the following sequence of MINITAB commands:

```
MULT M1 M2 M6
MULT M3 M6 M7
MULT M5 M7 M8
```

The trace of matrix $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{CX}$, or matrix M8, is obtained with the MINITAB commands

```
DIAG M8 C48
SUM(C48)
```

The sum of the connectivity matrix entries, $\mathbf{1}^t\mathbf{C1}$, is computed with the MINITAB commands

```
MULT M1 C49 C47
SUM(C47)
```

And, the numerator of the Moran Coefficient is calculated with the MINITAB commands

```
MULT M1 C50 C46
SUM(C50*C46)
```

while the denominator of this statistic is obtained with the MINITAB command

```
SUM(C50**2)
```

which is the sum of the squared residuals.

These various MINITAB commands allow MC to be defined as follows:

```
LET K4 = (K1/SUM(C47))*SUM(C50*C46)/SUM(C50**2)
```

In addition, these foregoing sets of MINITAB commands allow E(MC) to be defined as follows:

```
LET K5 = (-K1/(K1-K3))*SUM(C48)/SUM(C47)
```

The standard error of MC, requiring the calculation of VAR(MC), involves additional computations. First, the trace term $\mathrm{tr}[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t(\mathbf{C} + \mathbf{C}^t)^2\mathbf{X}]$ is calculated with the sequence of MINITAB commands

```
TRANS M1 M9
ADD M1 M9 M10
MULT M10 M10 M11
MULT M3 M11 M12
MULT M12 M2 M13
MULT M5 M13 M14
DIAG M14 C44
SUM(C44)
```

Next, the trace term $\mathrm{tr}\{[(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{C}\mathbf{X}]^2\}$ is computed with the sequence of MINITAB commands

```
MULT M8 M8 M15
DIAG M15 C43
SUM(C43)
```

And, the term S_1 is calculated with the sequence of MINITAB commands

```
LET K6 = 50 + K1
COPY M10 C51-CK6
LET K7 = 51
EXEC 'a:SQUARE.MCV' K1
RSUM C51-CK6 C42
SUM(C42)
```

This last set of code involves the macro a:SQUARE.MCV, which appears as follows:

```
a:SQUARE.MCV

let ck7 = ck7**2
let k7 = k7 + 1
end
```

These foregoing MINITAB commands are housed in the file a:CLASSIC.REG, which when executed yields an OLS regression, an MC test for spatial autocorrelation amongst the regression residuals, and a Shapiro-Wilk test for normality of the residuals. This set of code appears as follows:

```
a:CLASSIC.REG

trans m1 m9
sub m1 m9 m3
print m3
let k3 = k2 + 1
NOTE c1 is Y; c2-ck2 are the predictors
regress c1 k2 c2-ck3;
resids c50.
set c49
k1(1)
end
copy c49 c2-ck3 m2
trans m2 m3
mult m3 m2 m4
invert m4 m5
mult m1 m2 m6
mult m3 m6 m7
mult m5 m7 m8
diag m8 c48
mult m1 c49 c47
mult m1 c50 c46
NOTE Moran Coefficient (MC) calculated for residuals; printed as k4
let k4 = (k1/sum(c47))*sum(c50*c46)/sum(c50**2)
NOTE Expected value of MC for residuals
let k5 = (-k1/(k1-k3))*sum(c48)/sum(c47)
add m1 m9 m10
mult m10 m10 m11
mult m3 m11 m12
mult m12 m2 m13
mult m5 m13 m14
diag m14 c44
mult m8 m8 m15
diag m15 c43
let k6 = 50 + k1
copy m10 c51-ck6
let k7 = 51
exec 'a:square.mcv' k1
rsum c51-ck6 c42
let k8 = (k1**2)/((sum(c47)**2)*(k1-k3)*(k1-k3+2))
NOTE Variance of MC for residuals
let k8 = k8*(sum(c42)/2+2*sum(c43)-sum(c44)-2*(sum(c48)**2)/(k1-k3))
```

```
NOTE t-score calculated for MC; printed as k9, df printed as k10
let k9 = (k4 - k5)/sqrt(k8)
let k10 = k1 - k3
print k4,k9,k10
NOTE Shapiro-Wilk test for normality performed on residuals
nscores c50 c45
corr c50 c45
end
```

The Moran Coefficient is printed as K4, its t-statistic is printed as K9, and the number of degrees of freedom for this t-statistic are printed as K10. One should note that the first three lines of MINITAB commands generate a display of the differene between matrix **C** and $\mathbf{C}^t$, so that one can determine whether or not this matrix is symmetric. Moreover, matrix **C** is transposed, and the difference between this matrix and its transpose is calculated; the code involved here is as follows:

```
TRANS M1 M9
SUB M1 M9 M3
PRINT M3
```

2.2. Benchmark output for the Eire data

The results reported in Cliff and Ord (1981) are duplicated here by first executing the MINITAB command

```
EXEC 'a:START.TST' ↵
```

which results in the CRT display

```
26 ROWS READ
26 ROWS READ
```

This first response indicates that 26 rows of data have been read from the file a:EIRECONN.TST, while this second response indicates that 26 rows of data have been read from the file a:EIREDATA.TST.

Next, the regression analysis is completed by executing the MINITAB command

```
EXEC 'a:CLASSIC.REG' ↵
```

which results in the following CRT screen display for assessing the symmetry of the connectivity matrix

```
MATRIX M3
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
```

```
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

Since every entry in matrix M3 is zero, then matrix **C** and its transpose must have identical cell entries. During the displaying of the above matrix, and in fact throughout the displaying of all CRT MINITAB results, remember that the prompt "CONTINUE?" will appear periodically; a carriage return (↵) is the correct response to this prompt.

Next the OLS regression results are displayed on the CRT, together with a brief note.

```
cl is Y; c2-ck3 are the predictors


The regression equation is
C1 = 133 - 0.0103 C2


Predictor        Coef        Stdev    t-ratio        p
Constant       133.45        11.54      11.57    0.000
C2          -0.010340     0.002565      -4.03    0.000


s = 13.31        R-sq = 40.4%     R-sq(adj) = 37.9%


Analysis of Variance

SOURCE        DF          SS          MS         F        p
Regression     1      2880.7      2880.7     16.25    0.000
Error         24      4254.7       177.3
Total         25      7135.4


Unusual Observations
Obs.      C2        C1        Fit  Stdev.Fit  Residual   St.Resid
  5     7500     75.00      55.90       8.42     19.10      1.85 X
  6     3078    142.00     101.62       4.24     40.38      3.20R
 16     6815     71.00      62.98       6.77      8.02      0.70 X
```

```
R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

These outcomes agree with those findings reported in Cliff and Ord (1981, p. 209); in particular one should compare the a, b, and R^2 values. These standard OLS calculations are supplemented here with the spatial autocorrelation and normality test results, producing the CRT screen display

```
Moran Coefficient (MC) calculated for residuals; printed as k4
Expected value of MC for residuals
Variance of MC for residuals
t-score calculated for MC; printed as k9, df printed as k10
K4       0.190785
K9       2.17649
K10      24.0000
Shapiro-Wilk test for normality performed on residuals
Correlation of C50 and C45 = 0.965
```

The Moran Coefficient and t-statistic outcomes agree with those findings reported in Cliff and Ord (1981, Table 8.2, p. 211). The Shapiro-Wilk statistic implies that invoking the normality assumption is reasonable to do (this observed value lies closer to the hypothesized value of one than does its associated critical value of approximately 0.9593).

The second data analysis replication undertaken here is for the logarithmic form of the Eire model analyzed by Cliff and Ord. After completing the first regression analysis, the following MINITAB commands were executed:

```
LET C1 = LOGTEN(C1)
LET C2 = LOGTEN(C2)
```

These two commands transformed the original Eire data by converting them into their $\log_{10}$ versions. Then the MINITAB command executed was

```
EXEC 'a:CLASSIC.REG' ↵
```

which repeats the same CRT screem display for assessing the symmetry of the connectivity matrix as seen above. This execution also produces OLS regression results, and displays them on the CRT screen, together with a brief note.

```
c1 is Y; c2-cke are the predictors

The regression equation is
C1 = 4.19 - 0.621 C2

Predictor       Coef      Stdev   t-ratio        p
Constant      4.1937     0.4450      9.43    0.000
C2           -0.6211     0.1225     -5.07    0.000

s = 0.05613     R-sq = 51.7%     R-sq(adj) = 49.7%
```

```
Analysis of Variance

SOURCE        DF        SS          MS          F        p
Regression     1    0.081008    0.081008    25.71    0.000
Error         24    0.075610    0.003150
Total         25    0.156618

Unusual Observations
Obs.     C2        C1        Fit  Stdev.Fit  Residual   St.Resid
  5     3.88    1.8751    1.7869    0.0318    0.0882      1.91 X
  6     3.49    2.1523    2.0271    0.0207    0.1252      2.40R
 12     3.70    1.7782    1.8962    0.0138   -0.1181     -2.17R
 16     3.83    1.8513    1.8127    0.0271    0.0386      0.78 X

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

These outcomes also agree with those findings reported in Cliff and Ord (1981, p. 209); in particular one should compare the a, b, and R^2 values. Again these standard OLS calculations are supplemented here with the spatial autocorrelation and normality test results, producing the CRT display

```
Moran Coefficient (MC) calculated for residuals; printed as k4
Expected value of MC for residuals
Variance of MC for residuals
t-score calculated for MC; printed as k9, df printed as k10
K4       0.130061
K9       1.67558
K10      24.0000
Shapiro-Wilk test for normality performed on residuals

Correlation of C50 and C45 = 0.992
```

The MC and t-statistic outcomes agree with those findings reported in Cliff and Ord (1981, Table 8.2, p. 211). The Shapiro-Wilk statistic once again implies that invoking the normality assumption is reasonable to do (this observed value lies considerably closer to the hypothesized value of one than does its associated critical value of approximately 0.9593).

Consequently, the benchmarks used here demonstrate that the calculations done with MINITAB code are correct.

2.3. Illustrative output for the Puerto Rican data

Results for selected agricultural production density data from the Mayaguez Agricultural Administration Region of Puerto Rico (see Figure 2.1) are produced here for illustrative purposes, and are obtained by first executing the MINITAB command

```
EXEC 'a:START.PR' ↵
```

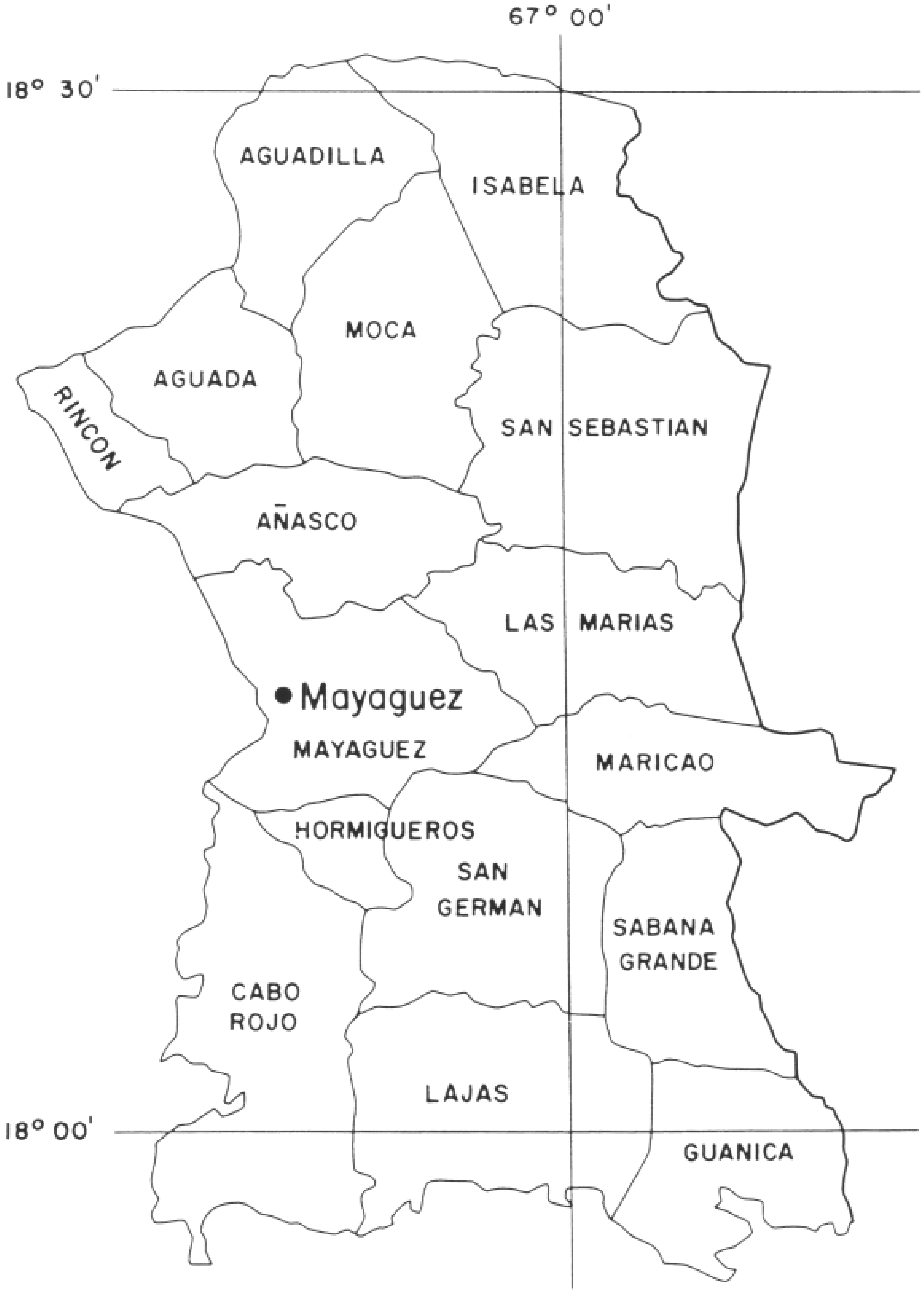


FIGURE 2.1

MAYAGUEZ AGRICULTURAL ADMINISTRATION REGION OF PUERTO RICO which results in the CRT display

```
16 ROWS READ
16 ROWS READ
```

This first response indicates that 16 rows of data have been read from the file a:MAYAGUEZ.CON, while this second response indicates that 16 rows of data have been read from the file a:MAYAGUEZ.DAT.

Next, the regression analysis is completed by executing the MINITAB command

```
EXEC 'a:CLASSIC.REG'  ↵
```

which results in the following CRT screen display for assessing the symmetry of the connectivity matrix:

```
MATRIX M3
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

Since every entry in matrix M3 is zero, then matrix **C** and its transpose must have identical cell entries, and so matrix **C** is symmetric.

Next the OLS regression results are displayed on the CRT, together with a brief note.

```
c1 is Y; c2-ck3 are the predictors


The regression equation is
C1 = 3.20 - 0.026 C2 + 0.119 C3 + 0.477 C4 +0.000937 C5 + 0.677 C6 - 1.61 C7 - 1.91 C8


Predictor        Coef       Stdev    t-ratio       p
Constant        3.199       3.561       0.90   0.395
C2            -0.0265      0.1035      -0.26   0.805
C3             0.1188      0.1408       0.84   0.423
C4             0.4769      0.1534       3.11   0.014
```

```
C5         0.0009373  0.0004807       1.95    0.087
C6            0.6766     0.5055       1.34    0.218
C7            -1.610      1.554      -1.04    0.331
C8            -1.910      1.140      -1.68    0.132

s = 3.301       R-sq = 78.3%     R-sq(adj) = 59.4%

Analysis of Variance
SOURCE         DF          SS          MS          F        p
Regression      7      315.42       45.06       4.14    0.032
Error           8       87.17       10.90
Total          15      402.59

SOURCE         DF      SEQ SS
C2              1        6.55
C3              1       15.07
C4              1        1.50
C5              1      218.25
C6              1       36.55
C7              1        6.93
C8              1       30.58
```

These results differ from those for the Cliff-Ord Eire data in that none of the observations seems to be uncharacteristic of the sample (presently the author is working on the problem of properly interpreting the battery of regression diagnostics in the presence of non-zero spatial autocorrelation). These OLS regression results will be referred to in later chapters, for comparative purposes.

Again these standard OLS calculations are supplemented here with the spatial autocorrelation and normality test results, producing the CRT display

```
Moran Coefficient (MC) calculated for residuals; printed as k4
Expected value of MC for residuals
Variance of MC for residuals
t-score calculated for MC; printed as k9, df printed as k10
K4       -0.205364
K9       -0.649001
K10      8.00000
Shapiro-Wilk test for normality performed on residuals

Correlation of C50 and C45 = 0.982
```

These findings suggest that there is no spatial autocorrelation present in the regression residuals, since the relevant critical value of the t-statistic is -2.31. Furthermore, these residuals imply that their parent population frequency distribution conforms to a normal distribution, given the relevant Shapiro-Wilk statistic critical value of about 0.90; this specific outcome reinforces the appropriateness of the normality assumption for testing MC.

APPENDIX 2-A.

EIRE DATA FROM CLIFF AND ORD

County	Variables				Connectivity Matrix																									
	Y	log(Y)	X	log(X)	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	97	1.9868	3664	3.5640	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
B	69	1.8388	5000	3.6990	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1	0	0	0	0	0	1	0	0
C	78	1.8921	4321	3.6356	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0
D	90	1.9542	4118	3.6147	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	1	1	0	0	0
E	75	1.8751	7500	3.8751	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0
F	142	2.1523	3078	3.4883	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
G	88	1.9445	4537	3.6568	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	1	0	0	0	0
H	78	1.8921	5140	3.7110	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0
I	111	2.0453	3200	3.5052	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	0	1
J	87	1.9395	3708	3.5691	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0
K	87	1.9395	3455	3.5384	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0
L	60	1.7782	5000	3.6990	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1	0	0	0	0	0
M	95	1.9777	4018	3.6040	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
N	77	1.8865	4250	3.6284	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	1	0	0
O	107	2.0294	3948	3.5964	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0
P	71	1.8513	6815	3.8335	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
Q	103	2.0128	4008	3.6029	0	1	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0
R	72	1.8573	4500	3.6532	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0
S	98	1.9912	4108	3.6136	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	1	0	0	1	0	1	0	1	0	0
T	71	1.8513	4500	3.6532	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	1	0	0	1	0	1	0	0	1	0	0
U	75	1.8751	5997	3.7779	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0	0
V	88	1.9445	3926	3.5940	0	0	1	1	0	0	1	0	0	1	1	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0
W	91	1.9590	3691	3.5671	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0
X	93	1.9685	3872	3.5879	0	1	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	1	0	0	0	0	0	0
Y	87	1.9395	3940	3.5955	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1
Z	102	2.0086	3600	3.5563	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0

NOTE: the connectivity matrix recording errors appearing in Table A8.1 (p. 229) of Cliff and Ord (1981) have been corrected.

APPENDIX 2-B.

ILLUSTRATIVE PUERTO RICAN DATA:
PRODUCTION DENSITY FOR THE MAYAGUEZ AGRICULTURAL ADMINISTRATIVE REGION

municipio	# farms	farm land	milk	sugar-cane	coffee	tobacco	bananas/ plantains	# families
Aguada	21.15	41.66	0.06	26.05	10286.9	0.00	2.24	1.26
Aguadilla	5.29	28.14	5.65	14.32	199.6	0.84	0.15	2.33
Anasco	9.29	37.04	0.00	14.67	11811.4	0.00	2.14	2.10
Cabo Rojo	5.54	48.28	22.50	14.79	7.0	0.00	0.17	3.62
Guanica	2.41	39.02	4.40	9.94	147.9	0.00	0.04	1.01
Hormigueros	9.77	44.50	0.07	27.66	4130.0	0.00	1.25	1.97
Isabela	11.00	32.31	39.64	5.79	683.6	8.29	0.48	2.02
Lajas	5.31	62.53	6.07	13.13	31.5	0.13	0.18	2.04
Las Marias	18.24	40.31	0.00	0.36	34270.9	3.20	9.33	2.47
Maricao	14.00	42.02	0.00	0.00	48140.1	0.00	14.26	4.75
Mayaguez	9.22	26.35	3.80	7.62	9763.0	0.33	2.92	1.62
Moca	14.30	37.87	2.09	15.29	6482.3	0.07	1.21	2.58
Rincon	8.42	21.60	0.00	7.54	2391.8	0.00	0.35	0.85
Sabana Grande	4.91	24.33	3.53	4.87	4898.7	0.00	0.05	1.65
San German	11.86	49.41	0.50	14.06	10993.9	5.79	1.71	4.30
San Sebastian	14.69	45.94	6.86	17.43	11689.8	1.79	1.83	2.74

	Connectivity Matrix																coordinates		C	R
																	x	y	I	U
Aguada	0	1	1	0	0	0	0	0	0	0	0	1	1	0	0	0	54.64	20.03	0	0
Aguadilla	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	61.11	23.19	0	1
Anasco	1	0	0	0	0	0	0	0	1	0	1	1	1	0	0	1	50.18	23.86	0	0
Cabo Rojo	0	0	0	0	0	1	0	1	0	0	1	0	0	0	1	0	34.42	22.63	0	0
Guanica	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	30.89	36.65	0	1
Hormigueros	0	0	0	1	0	0	0	0	0	0	1	0	0	0	1	0	40.42	24.72	1	1
Isabela	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	60.81	30.10	0	0
Lajas	0	0	0	1	1	0	0	0	0	0	0	0	0	1	1	0	32.81	29.49	0	0
Las Marias	0	0	1	0	0	0	0	0	0	1	1	0	0	0	0	1	47.25	32.10	1	0
Maricao	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1	0	43.14	34.99	1	0
Mayaguez	0	0	1	1	0	1	0	0	1	1	0	0	0	0	1	0	45.09	24.84	0	1
Moca	1	1	1	0	0	0	1	0	0	0	0	0	0	0	0	1	56.03	25.87	1	0
Rincon	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	52.88	17.08	0	0
Sabana Grande	0	0	0	0	1	0	0	1	0	1	0	0	0	0	1	0	37.47	35.07	1	0
San German	0	0	0	1	0	1	0	1	0	1	1	0	0	1	0	0	39.04	29.24	1	0
San Sebastian	0	0	1	0	0	0	1	0	1	0	0	1	0	0	0	0	53.13	32.61	1	0

NOTE: the x- and y- coordinates are municipio centroids, obtained from a map digitization process; C/I (coastal lowlands/interior highlands) and R/U (rural/urban) are binary classification variables.

CHAPTER 3.
STATISTICAL TECHNIQUES THAT ARE EXECUTABLE AS AN OLS REGRESSION

A number of statistical techniques can be rewritten as regression problems. In fact, the general linear model developments of the late 1960s and early 1970s did much to synthesize statistical theory, as well as solve certain problems that had been intractable or even unsolvable until then. Moreover, the advent of high-speed electronic computers has propagated a flurry of revived interest in the multiple regression approach to statistical problems, especially since tedious and horrendous calculations or computer programming requirements are no longer an obstacle to this approach. This chapter will review several different statistical techniques, in terms of their multiple regression versions, and in each case apply the test for non-zero residual spatial autocorrelation presented in Chapter 2. Those techniques to be addressed here include test of a single mean, analysis of variance (ANOVA), two-groups discriminant function analysis, correlation, and trend surface modelling.

3.1. Inference about the population mean

One can posit an hypothesis about a single regional mean. But the presence of non-zero spatial autocorrelation complicates the affiliated hypothesis testing. While the expected value of a sample mean is still its parent population mean, since this parameter estimate remains unbiased in the presence of non-zero spatial autocorrelation, complications arises from a biased estimate of the variance of the corresponding sampling distribution of sample means. In other words, usually non-zero spatial autocorrelation primarily impacts upon variance estimates, altering associated Type I and Type II error probabilities. This complication causes many sample results to be far more unstable, although still correct on average.

A simple bivariate regression approach can be taken to the estimation of a sample mean. In this formulation, the variable whose mean is sought becomes the dependent variable, **Y**, and this variate is regressed on a vector of ones, denoted by **1**, which becomes the matrix of "independent variables." Accordingly, $(\mathbf{X}^t\mathbf{X})^{-1}$ becomes $(\mathbf{1}^t\mathbf{1})^{-1}$, with $\mathbf{1}^t\mathbf{1}$ being the sum of n products of

1*1; this term yields 1/n. Meanwhile, $\mathbf{X}^t\mathbf{Y}$ becomes $\mathbf{1}^t\mathbf{Y}$, which is the sum of the observed values. Thus, $(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$ becomes $(\mathbf{1}^t\mathbf{1})^{-1}\mathbf{1}^t\mathbf{Y}$, which equals $\bar{y}$. Consequently, regressing vector **Y** on vector **1** yields an intercept value of zero, a regression coefficient value of $\bar{y}$, and a standard error for the regression coefficient of $s_{\bar{y}}$. Because the intercept term must equal zero, this type of regression is carried out with a no-intercept model. And, since a regression analysis is being conducted, the test for spatial autocorrelation outlined in Chapter 2 is applicable here.

The MINITAB code appearing in file a:CLASSIC.REG has been modified in order to exploit the reduced form of the regression analysis needed for handling this problem, and in order to estimate a no-intercept regression model. This revised version of regression analysis with a test for spatial autocorrelation is housed in file a:REG.MU; its set of code appears as follows:

a:REG.MU

```
trans m1 m9
sub m1 m9 m3
print m3
NOTE c1 is Y; c49 is a vector of 1s that will be the predictor
set c49
k1(1)
end
mean c1
regress c1 1 c49;
noconstant;
resids c50.
mult m1 c49 c47
mult m1 c50 c46
NOTE Moran Coefficient (MC) calculated for residuals; printed as k4
let k4 = (k1/sum(c47))*sum(c50*c46)/sum(c50**2)
NOTE Expected value of MC for residuals
let k5 = -1/(k1-1)
add m1 m9 m10
mult m10 c49 c45
let k6 = 50 + k1
copy m10 c51-ck6
let k7 = 51
exec 'a:square.mcv' k1
rsum c51-ck6 c42
let k11 = 1/((sum(c47)**2)*(k1-1)*(k1+1))
NOTE Variance of MC for residuals
let k8=(k1**2)*sum(c42)/2-k1*sum(c45**2)+2*(sum(c47)**2)*(k1-2)/(k1-1)
NOTE t-score calculated for MC; printed as k9, df printed as k10
let k9 = (k4 - k5)/sqrt(k11*k8)
let k10 = k1 - 1
print k4,k9,k10
NOTE Shapiro-Wilk test for normality performed on residuals
nscores c50 c45
corr c50 c45
end
```

This macro expects MINITAB column C1 to contain variable Y. It differs fundamentally from file a:CLASSIC.REG only in that it contains the reduced and simplified forms of many general regression terms. Also, the MINITAB command

```
MEAN C1
```

is included so that one can see that the regression solution indeed yields the mean of variable Y.

As a benchmark, data from Griffith (1987, p. 57) were analyzed with this program; only the results for Figure 5.4(c) [p. 37] will be summarized here. The first CRT screen display was

```
9 ROWS READ
9 ROWS READ
```

This first response indicates that 9 rows of data have been read from the connectivity matrix file, while this second response indicates that 9 rows of data have been read from the variate file [neither data set has been included on the diskette, since they are available in Griffith (1987)].

Next, the regression analysis was completed by executing the MINITAB command

```
EXEC 'a:REG.MU'  ↵
```

which results in the following CRT screen display for assessing the symmetry of the connectivity matrix:

```
MATRIX M3
 0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0
```

Since every entry in matrix M3 is zero, then matrix **C** and its transpose must have identical cell entries.

Then the OLS regression results were displayed on the CRT screen, together with the variable mean and a brief note.

```
c1 is Y; c49 is a vector of 1s that will be the predictor

MEAN    =      3.0000

The regression equation is
```

```
C1 = 3.00 C49

Predictor       Coef       Stdev    t-ratio       p
Noconstant
C49           3.0000      0.4082       7.35   0.000

s = 1.225

Analysis of Variance
SOURCE       DF         SS          MS         F        p
Regression    1     81.000      81.000     54.00    0.000
Error         8     12.000       1.500
Total         9     93.000

* NOTE  * ALL VALUES IN COLUMN ARE IDENTICAL
```

First the mean is reported; being 3, and next the regression results are reported, with no constant in the model (the intercept term) and a regression coefficient of 3; these two statistics are identical. The standard error of the mean is given by the standard error of the regression coefficient, and is 0.4082 here $(s_y/\sqrt{n})$. Finally, the analyst is notified that MINITAB column C49 is a constant, which is what it is to be. These standard OLS calculations were supplemented with the spatial autocorrelation and normality test results, producing the CRT screen display

```
Moran Coefficient (MC) calculated for residuals; printed as k4
Expected value of MC for residuals
Variance of MC for residuals
t-score calculated for MC; printed as k9, df printed as k10
K4       -0.875000
K9       -3.25396
K10      8.00000
Shapiro-Wilk test for normality performed on residuals
Correlation of C50 and C45 = 1.000
```

The Moran Coefficient and t-statistic outcomes agree with those findings reported in Griffith (1987). The Shapiro-Wilk statistic implies that invoking the normality assumption is reasonable to do, and demonstrates the difference between a raw value of this statistic, which is found in Griffith (1987), and a modified value of this statistic, which appears in the MINITAB Reference Manual.

Density of coffee production in the Mayaguez Agricultural Administrative Region, from the Puerto Rican data, is used here for illustrative purposes. These data are accessed with the file a:DEMOMAY.MU; this set of code appears as follows:

```
a:DEMOMAY.MU

noecho
let k1 = 16
let k2 = 0
let k3 = k2 + 1
read 'a:mayaguez.con' k1 k1 m1
read 'a:mayaguez.dat' c50-c54 c1
end
```

This file defines n as being equal to 16, reads in the connectivity matrix, discards the first five variables in the data file, and fills the MINITAB column C1 with the density of coffee production data. This file is executed with the MINITAB command

```
EXEC 'a:DEMOMAY.MU'  ↵
```

which results in the now familiar CRT screen display

```
16 ROWS READ
16 ROWS READ
```

As mentioned in Chapter 2, this first response indicates that 16 rows of data have been read from the file a:MAYAGUEZ.CON, while this second response indicates that 16 rows of data have been read from the file a:MAYAGUEZ.DAT.

Next, the regression analysis is completed by executing the MINITAB command

```
EXEC 'a:REG.MU'  ↵
```

which results in the following CRT screen display for assessing the symmetry of the connectivity matrix:

```
MATRIX M3
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
 0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
```

Again the symmetric of matrix **C** can be verified.

Succeeding this output, the OLS regression results are displayed on the

CRT screen, together with the variable mean and a brief note.

```
cl is Y; c49 is a vector of 1s that will be the predictor

MEAN    =      9745.5

The regression equation is
C1 = 9746 C49

Predictor       Coef       Stdev    t-ratio        p
Noconstant
C49             9746        3329       2.93    0.010

s = 13316

Analysis of Variance
SOURCE        DF          SS          MS          F        p
Regression     1  1519604096  1519604096       8.57    0.010
Error         15  2659847680   177323184
Total         16  4179451904

Unusual Observations
Obs.     C49        C1       Fit Stdev.Fit  Residual    St.Resid
 10     1.00     48140      9746      3329     38395       2.98R

R denotes an obs. with a large st. resid.

* NOTE  * ALL VALUES IN COLUMN ARE IDENTICAL
```

First the mean of the density of coffee production is reported, being 9745.5, and next the regression results are reported, with a zero intercept term (by construction) and a regression coefficient of 9746; these two statistics are identical, except for rounding error. The standard error of this mean is given as 3329 (the standard error of the regression coefficient). These standard OLS calculations are supplemented with the spatial autocorrelation and normality test results, producing the CRT screen display

```
Moran Coefficient (MC) calculated for residuals; printed as k4
Expected value of MC for residuals
Variance of MC for residuals
t-score calculated for MC; printed as k9, df printed as k10
K4       0.243819
K9       2.11573
K10      15.0000
Shapiro-Wilk test for normality performed on residuals
Correlation of C50 and C45 = 0.836
```

These findings suggest that there is near-significant spatial autocorrelation present in the density of coffee production variable, since the relevant critical value of the t-statistic is 2.13. Furthermore, this sample variate does not seem to come from a population whose frequency distribution conforms to a normal distribution, given the relevant Shapiro-Wilk statistic critical value of 0.9383; this specific outcome raises a question concerning the appropriateness of the normality assumption for testing MC in this situation.

3.2. One-way Analysis of Variance (ANOVA)

Analysis of variance involving a single factor design can be rewritten as a regression problem by converting the classification variable into a set of binary indicator variables, and then using these new arrays as the independent variables of a regression analysis. Because these indicator variables sum to the vector **1** (they are mutually exclusive and collectively exhaustive), a variable whose regression coefficient is the intercept term (see Section 3.1), jointly they constitute a case of perfect multicollinearity in regression analysis. This perfect multicollinearity can be reduced either by employing a regression model with a zero intercept term (see Section 3.1), or by removing any one of the indicator variables (this selection is arbitrary) from the analysis. The removal of a single indicator variable is possible because if an observation is not in any of the groups represented by those indicator variables retained in an analysis, then this observation must be a member of the group represented by that indicator variable that has been dropped from the analysis; in other words, the removed indicator variable represents redundant information, which is exactly what multicollinearity refers to. Now, since there often is a desire to compare group means, the best way to remove one of the indicator variables from an analysis is to subtract it from all other indicator variables. This differencing results in regression coefficients that are of the form $(\mu_j - \mu_k)$, or pairwise difference of means tests. Clearly, when one of these regression coefficients is not significantly different from zero, then the two population group means in question are expected to be equal. This is the standard formulation of the analysis of variance regression model; it yields identical results to those from standard analysis of variance calculations.

MINITAB allows indicator variables to be generated from a classification variable by using the command INDICATOR; this MINITAB command constructs one indicator variable for each group identified in the MINITAB column housing the classification variable. For the purposes of this workbook, this conversion procedure is achieved with the following set of MINITAB command:

```
LET K6 = 11 + K2 - 1
INDICATOR C2 C11-CK6
```

In this context, K2 defines the number of groups existing in the classification variable, which is read into the statistical package as MINITAB column C2. The set of indicator variables is housed in MINITAB columns C11-CK6. The differenced indicator variables are constructed with the set of MINITAB commands

```
LET K4 = K2 - 1
LET K5 = 11
EXEC 'a:SUB.IND' K4
LET K5 = K5 - 1
```

where the macro a:SUB.IND appears as

```
a:SUB.IND

let ck5 = ck5 - ck3
let k5 = k5 + 1
end
```

This procedure subtracts the last indicator variable from all others.

The regression analysis seeks to produce an analysis of variance; so, the MINITAB macro devised for this task includes the standard MINITAB command to achieve this end, namely

```
ONEWAY C1 C2
```

ANOVA results obtained with this MINITAB command then can be compared with those obtained from the regression analysis.

Again density of coffee production in the Mayaguez Agricultural Administrative Region, from the Puerto Rican data, is used here for illustrative purposes. These data are accessed with the file a:DEMOMAY.AOV; this set of code appears as follows:

```
a:DEMOMAY.AOV

noecho
let k1 = 16
let k2 = 2
let k3 = k2
read 'a:mayaguez.con' k1 k1 m1
read 'a:mayaguez.dat' c50-c54 c1 c55-c59 c2
end
```

This file defines n as being equal to 16, reads in the connectivity matrix, discards the first five variables in the data file, sets the sixth variable equal to Y (MINITAB column C1), discards the next 5 variables, and sets the twelfth variable equal to the classification variable (MINITAB column C2). The classification variable captured here is the coastal lowland/interior highland dichotomy, which means K2 = 2. This file is executed with the MINITAB command

```
EXEC 'a:DEMOMAY.AOV'  ↵
```

which results in the now familiar CRT screen display

```
16 ROWS READ
16 ROWS READ
```

As mentioned above, this first response indicates that 16 rows of data have been read from the file a:MAYAGUEZ.CON, while this second response indicates that 16 rows of data have been read from the file a:MAYAGUEZ.DAT.

The MINITAB code appearing in file a:REG.AOV is a modified version of a:CLASSIC.REG in which indicator variables are created from the MINITAB column C2 variable, and a one-way ANOVA is produced. The set of code for this revised version of regression analysis with a test for spatial autocorrelation appears as follows:

a:REG.AOV

```
trans m1 m9
sub m1 m9 m3
print m3
oneway c1 c2
let k6 = 11 + k2 - 1
indicator c2 c11-ck6
let k4 = k2 - 1
let k5 = 11
exec 'a:sub.ind' k4
let k5 = k5 - 1
copy c11-ck5 c2-ck3
NOTE c1 is Y; difference indicator variables c2-ck3 are predictors
regress c1 k4 c2-ck3;
resids c50.
set c49
k1(1)
end
copy c49 c11-ck5 m2
trans m2 m3
mult m3 m2 m4
invert m4 m5
mult m1 m2 m6
mult m3 m6 m7
mult m5 m7 m8
diag m8 c48
mult m1 c49 c47
mult m1 c50 c46
NOTE Moran Coefficient (MC) calculated for residuals; printed as k4
let k4 = (k1/sum(c47))*sum(c50*c46)/sum(c50**2)
NOTE Expected value of MC for residuals
let k5 = (-k1/(k1-k2))*sum(c48)/sum(c47)
add m1 m9 m10
mult m10 m10 m11
mult m3 m11 m12
mult m12 m2 m13
mult m5 m13 m14
diag m14 c44
mult m8 m8 m15
diag m15 c43
let k6 = 50 + k1
copy m10 c51-ck6
let k7 = 51
exec 'a:square.mcv' k1
rsum c51-ck6 c42
```

```
let k8 = (k1**2)/((sum(c47)**2)*(k1-k2)*(k1-k2+2))
NOTE Variance of MC for residuals
let k8 = k8*(sum(c42)/2+2*sum(c43)-sum(c44)-2*(sum(c48)**2)/(k1-k2))
NOTE t-score calculated for MC; printed as k9, df printed as k10
let k9 = (k4 - k5)/sqrt(k8)
let k10 = k1 - k2
print k4,k9,k10
NOTE Shapiro-Wilk test for normality performed on residuals
nscores c50 c45
corr c50 c45
end
```

The regression analysis is completed by executing this program with the MINITAB command

```
EXEC 'a:REG.AOV' ↵
```

which results in the usual CRT screen display for assessing the symmetry of the connectivity matrix. This matrix M3 display is followed by the output from the MINITAB command ONEWAY, standard analysis of variance results, which appears as

```
ANALYSIS OF VARIANCE ON C1
SOURCE     DF        SS        MS        F       p
C2          1 696990592 696990592     4.97   0.043
ERROR      14 1.963E+09 140204080
TOTAL      15 2.660E+09
                                  INDIVIDUAL 95 PCT CI'S FOR MEAN
                                  BASED ON POOLED STDEV
LEVEL       N      MEAN     STDEV ----+---------+---------+---------+-
    0       9      3925      5102 (--------*--------)
    1       7     17229     17101              (---------*---------)
                                  ----+---------+---------+---------+-
POOLED STDEV =    11841               0     10000     20000     30000
```

Next the OLS regression results are displayed on the CRT screen, together with a brief note.

```
c1 is Y; difference indicator variables c11-ck5 are predictors

The regression equation is
C1 = 10577 - 6652 C11

Predictor       Coef       Stdev    t-ratio        p
Constant       10577        2984       3.55    0.003
C11            -6652        2984      -2.23    0.043

s = 11841       R-sq = 26.2%     R-sq(adj) = 20.9%

Analysis of Variance
SOURCE       DF          SS          MS         F        p
Regression    1   696990592   696990592      4.97    0.043
Error        14  1962857216   140204080
```

```
Total         15  2659847680

Unusual Observations
Obs.     C11       C1       Fit Stdev.Fit  Residual   St.Resid
 10    -1.00    48140     17229      4475     30911      2.82R

R denotes an obs. with a large st. resid.
```

From the regression equation, when the differenced indicator variable is 1, then the mean is 10577 - 6652*(1) = 3925, which is the "Level 0" mean found in the preceding ONEWAY results. Similarly, when the differenced indicator variable is -1, then the mean is 10577 - 6652*(-1) = 17229, which is the "Level 1" mean found in the preceding ONEWAY results. From the ONEWAY results, the F-ratio is 4.97, and its probability is 0.043. From the regression results, the ANOVA table reveals these same two values, while the regression coefficient for the differenced indicator variables MINITAB column C11 has a t-statistic equaling $\sqrt{4.97} = -2.229$, with exactly the same probability. These standard OLS calculations are supplemented with the spatial autocorrelation and normality test results, producing the CRT screen display

```
Moran Coefficient (MC) calculated for residuals; printed as k4
Expected value of MC for residuals
Variance of MC for residuals
t-score calculated for MC; printed as k9, df printed as k10
K4        0.137169
K9        1.39289
K10       14.0000
Shapiro-Wilk test for normality performed on residuals
Correlation of C50 and C45 = 0.920
```

These findings suggest that there is insignificant spatial autocorrelation present in the density of coffee production variable difference of means residuals, based upon the two groups of coastal lowlands/interior highlands, since the relevant critical value of the t-statistic is 2.14. Furthermore, these residuals do not seem to come from a population whose frequency distribution conforms to a normal distribution, given the relevant Shapiro-Wilk statistic critical value of 0.9326 (it is close, though); this specific outcome raises some question concerning the appropriateness of the normality assumption for testing MC.

3.3. Two-groups discriminant function analysis

One popular multivariate technique is discriminant function analysis, which may be used to find linear combinations that maximize differences among group means, or to classify observations into groups, given some set of original independent variables **X**. When only two groups are present in a classification scheme, the computations for determining the single canonical discriminant function simplify to those of estimating a regression equation; moreover, the single canonical discriminant function can be obtained without solving an eigenfunction problem. In this context, the variable Y is a binary indicator variable, taking on the values of zero and unity. Overlooking the intercept term, the regression coefficients obtained by regressing this binary vector **Y** on matrix **X** are proportional to the actual canonical discriminant function coefficients. Except for rounding error, the actual canonical discriminant function coefficients are obtained by taking any one of the regression coefficients, dividing all other regression coefficients (excluding the intercept term) by this selected one, and then multiplying all of these quotients by the canonical discriminant function coefficient for that variable whose regression coefficient served as a divisor. In other words, in a two-group case the canonical discriminant weights are proportional to the weights for a multiple regression equation where a dichotomous group-membership variable regressed has been regressed on p regressors. Here each correlation between Y and X_j is a point-biserial coefficient.

Classificatory discriminant function analysis constructs a function for each group, with these functions being of the form, for group k,

$$L_k = \mathbf{X}^t\mathbf{W}^{-1}\bar{\mathbf{X}}_k - \bar{\mathbf{X}}_k^t\mathbf{W}^{-1}\bar{\mathbf{X}}_k/2 ,$$

where matrix **W** is the pooled within-groups covariance matrix, and $\bar{\mathbf{X}}_k$ is the vector of variable means for group k. The individual variable coefficients of a classificatory discriminant function result from the term $\mathbf{W}^{-1}\bar{\mathbf{X}}_k$, whereas the intercept for this function is provided by the term $\bar{\mathbf{X}}_k^t\mathbf{W}^{-1}\bar{\mathbf{X}}_k/2$. For the two-group problem,

$$L_1 - L_2 = \mathbf{X}^t\mathbf{W}^{1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2) + \text{intercept},$$

with the coefficients $\mathbf{W}^{1}(\bar{\mathbf{X}}_1 - \bar{\mathbf{X}}_2)$ again being proportional to the canonical discriminant function coefficients. This classificatory discriminant function problem is discussed here because this is the solution that MINITAB yields.

The binary variable Y used in the example for this technique is the coastal lowland/interior highlands dichotomy. Discrimination between these two groups will be attempted on the basis of density of farms, density of farm land, and density of milk production, for the Mayaguez Agricultural Administrative Region of Puerto Rico. The file a:DEMOMAY.DFA establishes the parameters for utilizing regression to achieve a two-group discriminant function analysis. This macro also renders the MINITAB output for a classificatory discriminant function analysis (this procedure allows a maximum of twenty groups). The contents of this file are as follows:

```
a:DEMOMAY.DFA

noecho
let k1 = 16
let k2 = 3
read 'a:mayaguez.con' k1 k1 m1
let k3 = k2 + 1
read 'a:mayaguez.dat' c50 c2-ck3 c51-c57 c1
disc c1 c2-ck3;
ldf c40 c41.
let c42 = c40 - c41
```

```
      print c42
      end
```

This setup defines the MINITAB column C1 as the classificatory variable, and MINITAB columns C2-CK3 as the predictor variables. The MINITAB command DISC executes a classificatory discriminant function analysis, and the MINITAB subcommand LDF stores the individual group discriminant function coefficients into MINITAB columns C40 and C41. MINITAB column C42 is the difference between these two sets of coefficients; these numerical values will be proportional to the regression coefficients obtained by regressing C1 on C2-CK3. The contents of MINITAB column C42 are printed. This file is executed with the MINITAB command

```
      EXEC 'a:DEMOMAY.DFA'  ↵
```

which results in the CRT screen display

```
     16 ROWS READ
     16 ROWS READ

Linear Discriminant Analysis for C1

Group        0       1
Count        9       7

Summary of Classification

Put into     ....True Group....
Group             0        1
0                 7        1
1                 2        6
Total N           9        7
N Correct         7        6
Proport.     0.778    0.857

N =   16     N Correct =    13      Prop. Correct = 0.812

Squared Distance Between Groups
                0        1
0         0.00000  1.13288
1         1.13288  0.00000

       Linear Discriminant Function for Group
                0        1
Constant  -7.3906  -9.5732
C2         0.3143   0.4610
C3         0.3030   0.3284
C4         0.0798   0.0140

Summary of Misclassified Observations

Observation     True     Pred    Group Sqrd Distnc Probability
                Group    Group
     1 **         0        1        0       6.987       0.122
```

```
                                        1      3.037       0.878
     3 **       0        1              0      0.7982      0.470
                                        1      0.5554      0.530
    14 **       1        0              0      2.179       0.746
                                        1      4.332       0.254
C42
  2.18263  -0.14672  -0.02540   0.06580
```

The first two lines of output should be quite familiar by now, and indicate that 16 rows of the connectivity and 16 rows of the data matrices have been read. The first entry in MINITAB column C42 is the intercept term, and is arbitrary. The second, third and fourth entries in this column are the coefficients, respectively, of the three variables (a) density of farms, (b) density of farm land, and (c) density of milk production. These three coefficients are proportional to the canonical discriminant function coefficients for this two-group problem.

Since the only change in canonical discriminant function analysis as a regression problem is the definition of the dependent variable, numerical results for this procedure may be secured by executing the original OLS regression macro a:CLASSIC.REG, or

```
EXEC 'a:CLASSIC.REG'  ↵
```

Of course this macro produces the CRT screen display for verifying symmetry of the connectivity matrix. This output is followed by the CRT screen display

```
c1 is Y; c2-ck3 are predictors

The regression equation is
C1 = - 0.013 + 0.0313 C2 + 0.0054 C3 - 0.0140 C4

Predictor       Coef       Stdev    t-ratio        p
Constant     -0.0129      0.5388      -0.02    0.981
C2           0.03129     0.02556       1.22    0.244
C3           0.00542     0.01224       0.44    0.666
C4          -0.01404     0.01242      -1.13    0.280

s = 0.4988      R-sq = 24.2%     R-sq(adj) = 5.2%

Analysis of Variance
SOURCE       DF          SS          MS         F        p
Regression    3      0.9514      0.3171      1.27    0.327
Error        12      2.9861      0.2488
Total        15      3.9375

SOURCE       DF      SEQ SS
C2            1      0.5897
C3            1      0.0438
C4            1      0.3180

Unusual Observations
Obs.      C2          C1        Fit Stdev.Fit  Residual    St.Resid
```

```
  1    21.1    0.000    0.874    0.297   -0.874    -2.18R
  7    11.0    0.000   -0.050    0.449    0.050     0.23 X

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

Again, the intercept term is not of interest. Now the three regression slopes are (carrying them out several decimal places beyond what was printed by MINITAB)

0.0312943   0.0054185   -0.0140351 ,

dividing each of these three coefficients by the last one yields

-2.2297169   -0.3860678   1 ,

and multiplying these by 0.06580, which is the classificatory discriminant function coefficient difference from above (see C42) for the third variable, renders

-0.1467154   -0.0254033   0.06580 ,

which round off to the differenced individual functions coefficients obtained with the MINITAB command DISC.

These standard OLS calculations are supplemented with the spatial autocorrelation and normality test results, producing the CRT screen display

```
Moran Coefficient (MC) calculated for residuals; printed as k4
Expected value of MC for residuals
Variance of MC for residuals
t-score calculated for MC; printed as k9, df printed as k10
K4       0.0879981
K9       1.32995
K10      12.0000
Shapiro-Wilk test for normality performed on residuals
Correlation of C50 and C45 = 0.979
```

These findings suggest that there is no significant spatial autocorrelation present in the coastal lowlands/interior highlands discriminant function residuals, since the relevant critical value of the t-statistic is 2.18. Furthermore, these residuals seem to come from a population whose frequency distribution conforms to a normal distribution, given the relevant Shapiro-Wilk statistic critical value of 0.9261.

3.4. Bivariate correlation

OLS regression can be used to calculate a bivariate correlation coefficient merely by converting both the Y and the X variables to their corresponding z-score counterparts. By construction, again, the regression of

Y on X yields a zero intercept; but multicollinearity restrictions are not encountered here, so that the no-intercept model is not required. For this problem the MINITAB code appearing in file a:CLASSIC.REG has been modified in order to convert both X and Y to z-score variates. This conversion is achieved with the following set of MINITAB commands:

```
CENTER C1-C2 C3-C4;
LOCATION;
SCALE.
```

The CENTER command instructs MINITAB to operate on columns C1 and C2, storing the results of these operations in columns C3 (for C1) and C4 (for C2). The operation performed when MINITAB command LOCATION is executed is subtraction of variable means from their respective columns; the operation performed when MINITAB command SCALE is executed is division of columns by their respective unbiased variable standard deviations. Thus, the mean of each variate is zero, and the standard deviation of each variate is one; the MINITAB command DESCRIBE is executed in the ensuing modified version of a:CLASSIC.REG so that these two features of the data may be checked.

The macro of interest here is constrained to treatment of two variables, and incorporates this above set of z-score transformations code. In addition, it includes the MINITAB command CORR in order to show that the regression coefficient obtained indeed equals the correlation coefficient in question. This revised version of regression analysis with a test for spatial autocorrelation is housed in file a:REG.COR; its set of code appears as follows:

```
a:REG.COR

trans m1 m9
sub m1 m9 m3
print m3
NOTE c1 is Y; c2 is the predictor
center c1-c2 c3-c4;
location;
scale.
let c1 = c3
let c2 = c4
describe c1-c2
corr c1 c2
let k3 = k2 + 1
regress c1 k2 c2;
resids c50.
set c49
k1(1)
end
copy c49 c2 m2
trans m2 m3
mult m3 m2 m4
invert m4 m5
mult m1 m2 m6
mult m3 m6 m7
mult m5 m7 m8
diag m8 c48
mult m1 c49 c47
mult m1 c50 c46
NOTE Moran Coefficient (MC) calculated for residuals; printed as k4
let k4 = (k1/sum(c47))*sum(c50*c46)/sum(c50**2)
NOTE Expected value of MC for residuals
let k5 = (-k1/(k1-k3))*sum(c48)/sum(c47)
add m1 m9 m10
mult m10 m10 m11
mult m3 m11 m12
mult m12 m2 m13
mult m5 m13 m14
diag m14 c44
mult m8 m8 m15
diag m15 c43
let k6 = 50 + k1
copy m10 c51-ck6
let k7 = 51
exec 'a:square.mcv' k1
rsum c51-ck6 c42
let k8 = (k1**2)/((sum(c47)**2)*(k1-k3)*(k1-k3+2))
NOTE Variance of MC for residuals
let k8 = k8*(sum(c42)/2+2*sum(c43)-sum(c44)-2*(sum(c48)**2)/(k1-k3))
NOTE t-score calculated for MC; printed as k9, df printed as k10
let k9 = (k4 - k5)/sqrt(k8)
let k10 = k1 - k3
print k4,k9,k10
NOTE Shapiro-Wilk test for normality performed on residuals
nscores c50 c45
corr c50 c45
end
```

This macro expects MINITAB column C1 to contain Y and column C2 to contain X.

Density of sugarcane production and density of farm families in the Mayaguez Agricultural Administrative Region, from the Puerto Rican data, are used here to exemplify correlation as a regression problem. These data are accessed with the file a:DEMOMAY.COR; this set of code appears as follows:

```
a:DEMOMAY.COR

noecho
let k1 = 16
let k2 = 1
read 'a:mayaguez.con' k1 k1 m1
let k3 = k2 + 1
read 'a:mayaguez.dat' c50-c54 c1 c55-c56 c2
end
```

As usual, this macro defines n as being equal to 16, defines the number of regressors as 1, reads in the connectivity matrix, discards the first five variables in the data file, fills the MINITAB column C1 with the density of sugarcane production, discards the next two variables, and fills the MINITAB column C2 with the density of farm families. This file is executed with the MINITAB command

```
EXEC 'a:DEMOMAY.COR' ↵
```

which results in the now familiar CRT screen display indicating that both the connectivity and data files have been read.

Next, the regression analysis is completed by executing the MINITAB command

```
EXEC 'a:REG.COR' ↵
```

which results in the recurrent MINITAB matrix M3 display for a connectivity matrix symmetry check. Then OLS regression results are displayed on the CRT screen, together with a description of the mean and standard deviation of each variable, the correlation coefficient for the two variables, and a brief note.

```
c1 is Y; c2 is the predictor


               N     MEAN   MEDIAN   TRMEAN    STDEV   SEMEAN
C1            16   -0.000   -0.305   -0.154    1.000    0.250
C2            16   -0.000   -0.239   -0.061    1.000    0.250
             MIN      MAX       Q1       Q3
C1        -0.731    2.883   -0.708    0.133
C2        -1.351    2.204   -0.642    0.336


Correlation of C1 and C2 = 0.546


The regression equation is
C1 = 0.000 + 0.546 C2


Predictor       Coef       Stdev    t-ratio       p
Constant      0.0000      0.2168       0.00   1.000
C2            0.5460      0.2239       2.44   0.029


s = 0.8672      R-sq = 29.8%     R-sq(adj) = 24.8%


Analysis of Variance
SOURCE       DF          SS          MS         F        p
Regression    1      4.4712      4.4712      5.95    0.029
Error        14     10.5288      0.7521
Total        15     15.0000


Unusual Observations
Obs.      C2        C1        Fit Stdev.Fit  Residual   St.Resid
  9     0.13     1.842      0.069     0.219     1.773      2.11R
 10     2.20     2.883      1.203     0.539     1.680      2.47RX


R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

This output demonstrates that the mean and the variance of each z-score variate is, respectively, zero and unity. It also shows that the intercept of the z-score regression is zero, and that the correlation coefficient for the two variables equals the slope of the z-score regression line. These standard OLS calculations are supplemented with the spatial autocorrelation and normality test results, producing the CRT screen display

```
Moran Coefficient (MC) calculated for residuals; printed as k4
Expected value of MC for residuals
Variance of MC for residuals
t-score calculated for MC; printed as k9, df printed as k10
K4      0.279060
K9      2.46053
K10     14.0000
Shapiro-Wilk test for normality performed on residuals
Correlation of C50 and C45 = 0.962
```

These findings suggest that significant spatial autocorrelation is present in the residuals for this correlation analysis, since the relevant critical value of the t-statistic is 2.14. Furthermore, these residuals seem to come from a parent population whose frequency distribution conforms to a normal distribution, given the relevant Shapiro-Wilk statistic critical value of 0.9342.

3.5. Trend surface models: linear, quadratic, and cubic forms

Trend surface models merit attention here because of their close relationship with spatial autocorrelation analysis, and because they rely so heavily on the statistical notion known as the "extra sum of squares" principle. In this first instance, spatial autocorrelation converges upon a trend surface as the average juxtaposed dependency approaches an absolute value correlation of one. In other words, if a trend surface component is necessary, but is not included in an analysis, the regression residuals will exhibit spatial autocorrelation; once these trend surface terms are introduced into the model specification, this feature of the residuals may well disappear. This is an example of the missing variables problem extensively discussed in econometrics. What a trend surface means is that the spatial mean is not stationary from location to location. Rather, it becomes a function of absolute location. Thus, a non-homogeneous process is operating over the planar surface.

The extra sum of squares principle pertains to the idea that the contribution a subset of variables makes to the regression sum of squares, appearing in the regression ANOVA table, may be investigated by determining the difference between the regression sum of squares with these regressors in the model, and this sum of squares with these regressors out of the model. Because the marginal contribution to the regression sum of squares is being calculated, these are conditional quantities. This increase in the regression sum of squares that results from adding r regressors is distributed as an F-ratio, with r and (n - p - 1) degrees of freedom. This F-ratio is calculated by dividing this sum of squares quantity by r, and then dividing this ratio by the mean square error quantity obtained when all p regressors are in the model.

Macros have been developed here for evaluating a linear, a quadratic, and a cubic trend surface model. The linear model has two orthogonal coordinate variables, say U and V, the quadratic model adds three additional terms of second degree (two squared coordinate terms, U^2 and V^2, and a cross-product term, UV), and the cubic model adds four final terms of third degree (two cubed terms, U^3 and V^3, and two cross-product terms, U^2V and UV^2). These terms can become highly collinear by construction; centering (i. e., subtracting the respective means) the coordinate variables U and V dramatically helps to avoid this situation, at least until considerably high order models are dealt with. This end is achieved because centering moves the absolute co-ordinate system to a trough/peak of the parabola, for example, in a quadratic trend surface model; positions away from such exaggerated curvatures may result in nonlinear portions of the curve being closely approximated by a straight line. So, again, the regression macro includes a MINITAB LOCATION command.

Besides centering the data, the only two other tasks to be completed are (a) to identify the coordinates of each areal unit, and (b) to construct the various polynomial terms. Once these three data management issues have been resolved, then the original OLS regression macro can be employed. Once more density of milk production is used for illustrative purposes. The relevant data are secured by executing the file a:DEMOMAY.TSM, whose code appears as

```
      a:DEMOMAY.TSM

      noecho
      let k1 = 16
      read 'a:mayaguez.con' k1 k1 m1
      read 'a:mayaguez.dat' c50-c52 c1 c53-c57 c2 c3
      end
```

This file is executed with the MINITAB command

```
      EXEC 'a:DEMOMAY.TSM'  ↵
```

It causes the connectivity matrix to be read, and it retrieves the third variable from the data set as Y, and the the tenth and eleventh variables from this data set as U and V. Of course it generates the CRT screen display indicating that both the connectivity matrix and the data file have been read.

The linear trend surface model is obtained by executing the macro a:LINEAR.TSM, whose code appears as

```
      a:LINEAR.TSM

      center c2-c3 c4-c5;
      location;
      scale.
```

```
let c2=c4
let c3=c5
let k2 = 2
let k3 = k2 + 1
exec 'a:classic.reg'
end
```

This macro centers the coordinate system, and executes the original regression macro. It is executed with the MINITAB command

```
EXEC 'a:LINEAR.TSM'  ↵
```

Of course it generates the familiar M3 matrix. And, it produces the following CRT screen display of OLS results:

```
c1 is Y; c2-ck3 are predictors


The regression equation is
C1 = 5.95 + 2.58 C2 + 1.68 C3


Predictor       Coef        Stdev    t-ratio        p
Constant       5.948        2.764       2.15    0.051
C2             2.578        3.121       0.83    0.424
C3             1.684        3.121       0.54    0.599

s = 11.05       R-sq = 5.3%       R-sq(adj) = 0.0%

Analysis of Variance
SOURCE        DF          SS          MS         F        p
Regression     2        89.6        44.8      0.37    0.700
Error         13      1588.6       122.2
Total         15      1678.2

SOURCE        DF      SEQ SS
C2             1        54.0
C3             1        35.6

Unusual Observations
Obs.      C2        C1        Fit Stdev.Fit  Residual    St.Resid
  4    -1.20     22.50       1.37      6.12     21.13       2.30R
  7     1.49     39.64      10.51      6.01     29.13       3.14R

R denotes an obs. with a large st. resid.
```

Clearly neither the F-ratio for the regression ANOVA, nor the individual regression coefficient t-statistics suggest that a linear trend surface is appropriate for describing the spatial variation of milk production density. These standard OLS calculations are supplemented with the spatial autocorrelation and normality test results, producing the CRT screen display

```
Moran Coefficient (MC) calculated for residuals; printed as k4
Expected value of MC for residuals
Variance of MC for residuals
```

```
t-score calculated for MC; printed as k9, df printed as k10
K4      -0.00785980
K9      1.20199
K10     13.0000
Shapiro-Wilk test for normality performed on residuals
Correlation of C50 and C45 = 0.804
```

These findings suggest that there is little spatial autocorrelation in the regression residuals; but, the frequency distribution of these residuals deviates markedly from a normal one, raising a concern about the appropriateness of the normality assumption for testing MC in this situation.

Once the linear trend surface has been appraised, a quadratic trend surface regression model can be estimated; these two models must be explored sequentially, with the linear model estimated first. The quadratic trend surface model is obtained by executing the macro a:QUADRATI.TSM, whose code appears as

```
a:QUADRATI.TSM

let c4 = c2**2
let c5 = c2*c3
let c6 = c3**2
let k2 = 5
let k3 = k2 + 1
exec 'a:classic.reg'
end
```

This macro builds upon the linear trend surface model constructed with macro a:LINEAR.TSM, and also executes the original regression macro a:CLASSIC.REG. It is executed with the MINITAB command

```
EXEC 'a:QUADRATI.TSM'  ↵
```

Of course it, too, generates the familiar M3 matrix. And, it produces the following CRT screen display of OLS results:

```
c1 is Y; c2-ck3 are predictors


The regression equation is
C1 = - 3.43 + 1.83 C2 + 1.54 C3 + 10.2 C4 + 11.5 C5 + 4.45 C6


Predictor       Coef       Stdev    t-ratio        p
Constant      -3.426       2.569      -1.33    0.212
C2             1.830       1.487       1.23    0.246
C3             1.535       1.507       1.02    0.332
C4            10.182       1.713       5.94    0.000
C5            11.451       2.213       5.17    0.000
C6             4.446       1.938       2.29    0.045

s = 5.220       R-sq = 83.8%     R-sq(adj) = 75.6%


Analysis of Variance
```

```
SOURCE        DF        SS         MS         F        p
Regression     5   1405.70     281.14     10.32    0.001
Error         10    272.51      27.25
Total         15   1678.21

SOURCE        DF    SEQ SS
C2             1     54.01
C3             1     35.57
C4             1    500.71
C5             1    672.00
C6             1    143.41

Unusual Observations
Obs.      C2         C1         Fit Stdev.Fit  Residual   St.Resid
  7     1.49      39.64       30.70      4.06      8.94      2.73R

R denotes an obs. with a large st. resid.
```

Using the extra sum of squares principle to compare this quadratic model with the preceding linear model yields the following F-ratio for the marginal contribution to the regression sum of squares:

$$F \;=\; [(1405.70 - 89.6)/3]/27.25 \;=\; 16.099 \;>\; F_{0.05,3,10} \;=\; 3.71 \;.$$

This result suggests that the quadratic terms add a significant amount of statistical explanation to the trend surface model. This suggestion is further reinforced by the dramatic increase in the R^2 value, as well as the sizeable decrease in the residual mean square value. These standard OLS calculations are supplemented with the spatial autocorrelation and normality test results, producing the CRT screen display

```
Moran Coefficient (MC) calculated for residuals; printed as k4
Expected value of MC for residuals
Variance of MC for residuals
t-score calculated for MC; printed as k9, df printed as k10
K4       -0.317211
K9       -0.683824
K10      10.0000
Shapiro-Wilk test for normality performed on residuals
Correlation of C50 and C45 = 0.988
```

These findings also suggest that there is only spurious spatial autocorrelation in the regression residuals. One should note, though, by adding the quadratic terms to the trend surface model, the t-statistic for MC has moved closer to zero in absolute value, demonstrating to some degree the close relationship between trend surfaces and spatial autocorrelation. In addition, the Shapiro-Wilk statistic implies that the frequency distribution for the regression residuals, in the population, conforms to a normal distribution, given its affiliated critical value of 0.9180; this particular outcome is a welcomed improvement over its counterpart for the linear trend surface model.

Finally, after the linear and quadratic trend surface models have been evaluated, a cubic trend surface regression model can be estimated; these three models must be explored sequentially, with the linear model estimated first, and the quadratic model estimated second. The cubic trend surface model is obtained by executing the macro a:CUBIC.TSM, whose code appears as

```
a:CUBIC.TSM

let c7 = c2**3
let c8 = c4*c3
let c9 = c2*c6
let c10 = c3**3
let k2 = 9
let k3 = k2 + 1
exec 'a:classic.reg'
end
```

This macro builds upon the linear and quadratic trend surface models constructed above, and also executes the original regression macro a:CLASSIC.REG. It is executed with the MINITAB command

```
EXEC 'a:CUBIC.TSM' ↵
```

Of course it, too, generates the familiar M3 matrix. And, it produces the following CRT screen display of OLS results:

```
c1 is Y; c2-ck3 are predictors


The regression equation is

C1 = - 2.27 - 3.03 C2 - 2.23 C3 + 8.49 C4 + 12.0 C5 + 4.60 C6 + 3.99 C7 + 4.34 C8 - 1.30 C9 - 0.46 C10


Predictor       Coef      Stdev   t-ratio       p
Constant      -2.268      2.976     -0.76   0.475
C2            -3.030      4.227     -0.72   0.500
C3            -2.231      4.671     -0.48   0.650
C4             8.487      1.876      4.52   0.004
C5            12.032      2.396      5.02   0.002
C6             4.603      2.760      1.67   0.146
C7             3.994      2.103      1.90   0.106
C8             4.336      2.651      1.64   0.153
C9            -1.302      4.692     -0.28   0.791
C10           -0.465      3.261     -0.14   0.891

s = 4.671       R-sq = 92.2%     R-sq(adj) = 80.5%

Analysis of Variance

SOURCE       DF          SS          MS         F        p
Regression    9     1547.29      171.92      7.88    0.010
Error         6      130.92       21.82
Total        15     1678.21

SOURCE       DF      SEQ SS
C2            1       54.01
```

```
C3          1      35.57
C4          1     500.71
C5          1     672.00
C6          1     143.41
C7          1      17.22
C8          1     121.53
C9          1       2.40
C10         1       0.44
```

Using the extra sum of squares principle to compare this cubic model with the preceding quadratic model yields the following F-ratio for the marginal contribution to the regression sum of squares:

$$F = [(1547.29 - 1405.70)/4]/21.82 = 1.622 < F_{0.05,4,6} = 4.53 .$$

This result suggests that the cubic terms fail to add any significant amount of statistical explanation to the trend surface model. Thus, the quadratic model should be accepted as a reasonable description of the non-stationary spatial mean landscape. Returning to the quadratic trend surface output, one should recognize that its linear trend surface terms do not have significant regression coefficients, and so perhaps they need to be removed from the model. The acknowledgement of strictly quadratic terms in this example is consistent with the nature of agricultural production in Puerto Rico, where restrictions and constraints placed on agricultural production by the interior highlands could lead to such a surface. These standard OLS calculations are supplemented with the spatial autocorrelation and normality test results, too, producing the CRT screen display

```
Moran Coefficient (MC) calculated for residuals; printed as k4
Expected value of MC for residuals
Variance of MC for residuals
t-score calculated for MC; printed as k9, df printed as k10
K4       -0.241015
K9       0.572380
K10      6.00000
Shapiro-Wilk test for normality performed on residuals
Correlation of C50 and C45 = 0.970 next
```

Not surprising, these findings are consistent with those for the quadratic trend surface model; after all, no noticeable increase in statistical explanation has been achieved by adding the cubic terms. Noteworthy here, though, is the additional shrinkage of the distance separating the MC t-statistic and zero; again, the close relationship between trend surface models and spatial autocorrelation is alluded to.

CHAPTER 4.
ESTIMATING AN SAR ERROR MODEL

As was mentioned in Chapter 1, there are several autoregressive/moving average models that can be fitted to or estimated with geographic data exhibiting a non-zero level of spatial autocorrelation. The model to be focused on in this Discussion Paper is the simultaneous autoregressive model (SAR), for autocorrelated errors. The present plan is for these models to be topics of future Discussion Papers. This chapter will be devoted to issues of estimation, and the description of a MINITAB macro for performing this estimation; this algorithm is based upon Griffith (1988a). The macro is interactive and iterative.

A word of caution is in order here. Spatial autoregressive models are nonlinear in nature. As is found with their time series counterparts, the nonlinear nature of the mathematical structure means there always is a risk that convergence of the iterative estimation procedure will not occur within the feasible parameter space [for the SAR model, with matrix **C** converted to matrix **W**, this region is (-1,1)]. The convergence criterion is a weighted mean square error. The optimization problem is one of minimizing this weighted average sum of squares. For most cases convergence will be achieved. But when it is not, then the analyst must evaluate two conditions of the prevailing model specification. First, the simple SAR model assumes a stationary spatial mean; failure of convergence could mean that the spatial mean is non-stationary, and hence trend surface terms need to be introduced into the regression equation specification. The relationship between spatial autocorrelation and trend surfaces was briefly sketched in Chapter 3. As Upton and Fingleton (1985) demonstrate, because the likelihood function tends to be well behaved within the feasible parameter space, changing the starting values of the iterative procedure should prove futile and fruitless in a quest for convergence. Second, the model specification under study could have too many parameters (overparameterization), and hence some parameters need to be deleted. In other words, the SAR model should be replace with a CAR, an MA, or an autoregressive response model. Inspection of the MC t-statistic should prove illuminating here. If this statistic is not significant, then fitting any autoregressive model to the regression residuals would tend to be questionable; such an estimation exercise may well involve an overparameterization of the model. If this statistic is significant, then experience indicates that convergence should occur.

4.1. The estimation algorithm

The iterative algorithm developed for estimating a spatial SAR model is comprised of two macros, and three subroutines. This algorithm accepts input from any of the macros presented in Chapter 3; these sorts of output constitute its starting point. As was mentioned in Chapter 1, one major problem with spatial statistical analysis is that the Jacobian term is present and complex. A first step in estimating the spatial autocorrelation parameter is to construct the necessary components of this Jacobian term. The SAR model usually converts the connectivity matrix **C** to its stochastic counterpart, say **W** (recall that each entry of matrix **C** is divided by its respective row sum, resulting in each of the rows of matrix **W** summing to unity). This conversion is achieved with the MINITAB commands

```
LET K19 = 50
COPY M1 C51-CK19
RSUM C51-CK19 C50
LET K18 = 51
EXEC 'a:WFROM.C' K1
COPY C51-CK19 M2
```

where the MINITAB command RSUM sums the rows of a set of MINITAB columns. The subroutine a:WFROM.C has the following code:

```
a:WFROM.C

let ck18 = ck18/c50
let k18 = k18 + 1
end
```

Griffith (1988a), after Ord (1975), shows how to extract the eigenvalues of this non-symmetric matrix while satisfying the symmetry constraint imposed by MINITAB. This task is achieved with the MINITAB commands

```
LET C51 = SQRT(1/C50)
DIAG C51 M3
MULT M3 M1 M4
MULT M4 M3 M5
EIGEN M5 C48
```

The eigenvalues of matrix **W** are basic ingredients for evaluating the Jacobian term. These eigenvalues must sum to zero, and the principal eigenvalue must equal unity (except for rounding error).

The feasible parameter space is (-1,1), which is determined by the principal eigenvalue of matrix **W** (values close to ± 1 can be analyzed, such as 0.999, but use of the values ± 1 themselves will result in an error statement and subsequent erroneous calculations). This range of values must be searched for the minimum mean square error for the nonlinear regression model. This search procedure is initiated with the MINITAB macro a:START.SAR, whose code appears as

```
a:START.SAR

let k20 = -0.9
let k21 = 0.1
let k19 = 50 + k1
copy m1 c51-ck19
rsum c51-ck19 c50
let k18 = 51
exec ’a:wfrom.c’ k1
copy c51-ck19 m2
let c51 = sqrt(1/c50)
diag c51 m3
mult m3 m1 m4
mult m4 m3 m5
eigen m5 c48
let k18 = 1
let k19 = 51
exec ’a:slag.x’ k3
mult m2 c49 c50
exec ’a:classic.sar’ 19
end
```

This macro generates 19 regressions, each involving a set of spatially lagged variables; these variables are constructed with the MINITAB commands

```
LET K18 = 1
LET K19 = 51
EXEC "a:SLAG.X’ K3
```

The code for subroutine a:SLAG.X appears as

```
a:SLAG.X

mult m2 ck18 ck19
let k18 = k18 + 1
let k19 = k19 + 1
end
```

The analyst must monitor each of the 19 resulting regression mean square errors, searching for the minimum (this is the interactive feature of this method). These regressions are executed with the MINITAB macro a:CLASSIC.SAR, whose code appears as

```
a:CLASSIC.SAR

let k22 = exp(mean(loge((1-k20*c48)**2))/2)
let k17 = 1
let k18 = 51
let k19 = 15
exec ’a:trans.x’ k3
let c47 = (c49 - k20*c50)/k22
let k19 = k19 - 1
```

```
regress c15 k3 c47 c16-ck19;
noconstant;
mse k15.
print k20, k15
let k20 = k20 + k21
end
```

This set of MINITAB commands constructs synthetic variables that are products of the spatial linear operator ($\mathbf{I} - \rho\mathbf{W}$) and the original X_j and Y variates. Because the likelihood function is a summation of squared quantities multiplied by the Jacobian term, each synthetic variate can be multiplied by the square root of this Jacobian term, which is equivalent to distributing the Jacobian term over the individual summation terms of the likelihood function. Since this Jacobian term is raised to a negative exponential power, this multiplication operation can be written as a division. And, since its logarithmic form is equivalent to a sum divided by n, it also can be rewritten as the function of an arithmetic mean. Thus, the modified Jacobian term is calculated with the MINITAB command

```
LET K22 = EXP(MEAN(LOGE((1-K20*C48)**2))/2)
```

This quantity becomes the denominator of a fraction in which the numerator is the spatial linear operator times a given variable. Hence, the synthetic variables are constructed by the MINITAB macro a:TRANS.X, whose code appears as

```
a:TRANS.X

let ck19 = (ck17 - k20*ck18)/k22
let k17 = k17 + 1
let k18 = k18 + 1
let k19 = k19 + 1
end
```

Finally, an OLS regression is executed using these synthetic variables, which actually have some level of spatial autocorrelation filtered out of them. This OLS yields a mean square error value, printed here as K15, which is the quantity to be minimized; the corresponding value of the spatial autocorrelation parameter is printed as K20.

Once the algorithm has searched over the range (-0.9,0.9) by increments of 0.1, then the analyst must identify that value of $\hat{\rho}$ from the 19 choices being displayed that has the minimum mean square error, and then initiate search over the range ($\hat{\rho}$ - 0.09, $\hat{\rho}$ + 0.09) by increments of 0.01. Again, once the improved estimate of $\hat{\rho}$ is uncovered by identifying the new and more exact minimum mean square error from a second set of 19 choices, the search must be extended to the range ($\hat{\rho}$ - 0.009, $\hat{\rho}$ + 0.009) using increments of 0.001. This third search completes the process. This second step of the search process is completed by executing the MINITAB commands

```
LET K20 = ρ̂
LET K21 = .01
EXEC 'a:CLASSIC.SAR' 19
```

This third algorithm step is completed by executing the MINITAB commands

```
LET K20 = ρ̂
LET K21 = .001
EXEC 'a:CLASSIC.SAR' 19
```

Upon completion of these three steps, the analyst can identify the minimum mean square error (MSE) value, and hence the maximum likelihood estimate of $\hat{\rho}$.

4.2. Illustrative estimations for problems from Chapter 3

The first problem investigated in Chapter 3 has to do with inference about the population mean. The t-statistic for MC calculated with the regression residuals for this example is significant at the 10% (critical value of 1.753), but not at the 5% level. Given this context, one may want to explore a SAR specification for this model. The iterations generated by the algorithm presented in Section 4.1, for this problem, are summarized in the following tabulations:

LET K20 = -.9		LET K20 = .51		LET K20 = .571	
LET K21 = .1		LET K21 = .01		LET K21 = .001	
$\hat{\rho}$	MSE	$\hat{\rho}$	MSE	$\hat{\rho}$	MSE
-.9	343379264	.51	148794128	.571	148262064
-.8	312978400	.52	148651984	.572	148260368
-.7	286835840	.53	148530544	.573	148258928
-.6	264183360	.54	148430384	.574	148257712
-.5	244446880	.55	148352160	.575	148256736
-.4	227189872	.56	148296432	.576	148256016
-.3	212077152	.57	148264016	.577	148255504
-.2	198850832	.58	148255552*	.578	148255280*
-.1	187314400	.59	148271952	.579	148255312
0.0	177323200	.60	148314048	.580	148255568
.1	168779168	.61	148382816	.581	148256080
.2	161631056	.62	148479264	.582	148256832
.3	155880448	.63	148604512	.583	148257824
.4	151597504	.64	148759824	.584	148259088
.5	148956512	.65	148946480	.585	148260608
.6	148314032*	.66	149165888	.586	148262352
.7	150405312	.67	149419664	.587	148264384
.8	156947232	.68	149709568	.588	148266656
.9	173425392	.69	150037408	.589	148269168

Consequently, the maximum likelihood estimate of the spatial autocorrelation parameter is $\hat{\rho} = 0.578$.

The second problem investigated in Chapter 3 has to do with analysis of variance. Here the t-statistic for MC is not significant, even at the 10% level. Hence, fitting of a SAR model to these data may falter, in that the MSE may fail to converge upon a minimum within the interval (-1,1) due to overparameterization of the model. As an illustrative exercise, consider the estimation of a SAR model here. The iterations generated by the algorithm presented in Section 4.1, for this problem, are summarized in the following tabulations:

LET K20 = -.9		LET K20 = .31		LET K20 = .381	
LET K21 = .1		LET K21 = .01		LET K21 = .001	
$\hat{\rho}$	MSE	$\hat{\rho}$	MSE	$\hat{\rho}$	MSE
-.9	227614112	.31	132675208	.381	132305784
-.8	210258912	.32	132589032	.382	132304768
-.7	195644032	.33	132513736	.383	132303872
-.6	183269216	.34	132449544	.384	132303088
-.5	172758688	.35	132396504	.385	132302408
-.4	163825840	.36	132354960	.386	132301880
-.3	156250656	.37	132325016	.387	132301464
-.2	149865360	.38	132306952	.388	132301184
-.1	144545232	.39	132300944*	.389	132300992
0.0	140204096	.40	132307288	.390	132300936*
.1	136792832	.41	132326216	.391	132301008
.2	134302416	.42	132358024	.392	132301200
.3	132772112	.43	132403016	.393	132301552
.4	132307280*	.44	132461520	.394	132301976
.5	133116272	.45	132533816	.395	132302552
.6	135590112	.46	132620336	.396	132303256
.7	140500528	.47	132721408	.397	132304072
.8	149604832	.48	132837440	.398	132305016
.9	168429104	.49	132968920	.399	132306104

Consequently, the maximum likelihood estimate of the spatial autocorrelation parameter is $\hat{\rho}$ = 0.390. One should note that this estimate is lower than that for the preceding problem, as well as the t-statistic for MC accompanying this problem is closer to zero than the one for the preceding problem.

The third problem investigated in Chapter 3 has to do with two-group canonical discriminant function analysis. Again the t-statistic for MC is not significant, even at the 10% level. Hence, fitting of a SAR model to these data may falter, too, in that the MSE may fail to converge upon a minimum within the interval (-1,1). Once more, as an illustrative exercise, consider the estimation of a SAR model here. The iterations generated by the algorithm presented in Section 4.1, for this problem, are summarized in the following tabulations:

LET K20 = -.9		LET K20 = .31		LET K20 = .371	
LET K21 = .1		LET K21 = .01		LET K21 = .001	
$\hat{\rho}$	MSE	$\hat{\rho}$	MSE	$\hat{\rho}$	MSE
-.9	.379174	.31	.236464	.371	.235930
-.8	.354060	.32	.236325	.372	.235928
-.7	.332775	.33	.236207	.373	.235926
-.6	.314600	.34	.236107	.374	.235924
-.5	.299002	.35	.236028	.375	.235922
-.4	.285582	.36	.235970	.376	.235921
-.3	.274042	.37	.235933	.377	.235919
-.2	.264163	.38	.235917*	.378	.235918
-.1	.255792	.39	.235923	.379	.235917
0.0	.248839	.40	.235953	.380	.235917

.1	.243278	.41	.236005	.381	.235917
.2	.239157	.42	.236082	.382	.235916*
.3	.236621	.43	.236184	.383	.235916*
.4	.235953*	.44	.236310	.384	.235917
.5	.237653	.45	.236464	.385	.235917
.6	.242599	.46	.236644	.386	.235918
.7	.252430	.47	.236852	.387	.235919
.8	.270695	.48	.237089	.388	.235920
.9	.308114	.49	.237355	.389	.235922

Consequently, the maximum likelihood estimate of the spatial autocorrelation parameter is $\hat{\rho} = 0.3825$. One should note that this estimate is lower than that for the preceding two problems, and its t-statistic for MC also is closer to zero than are those for the preceding two problems.

The fourth problem investigated in Chapter 3 has to do with the correlation between two variables. This time the t-statistic for MC is significant at the 5% level. Hence, fitting of a SAR model to these data is advisable. The iterations generated by the algorithm presented in Section 4.1, for this problem, are summarized in the following tabulations:

LET K20 = −.9		LET K20 = .51		LET K20 = .591	
LET K21 = .1		LET K21 = .01		LET K21 = .001	
$\hat{\rho}$	MSE	$\hat{\rho}$	MSE	$\hat{\rho}$	MSE
−.9	1.515220	.51	.617594	.591	.613487
−.8	1.375280	.52	.616759	.592	.613477
−.7	1.255010	.53	.616011	.593	.613468
−.6	1.150860	.54	.615353	.594	.613460
−.5	1.060180	.55	.614787	.595	.613454
−.4	.980945	.56	.614316	.596	.613448
−.3	.911593	.57	.613942	.597	.613443
−.2	.850913	.58	.613668	.598	.613440
−.1	.797969	.59	.613499	.599	.613437
0.0	.752060	.60	.613436*	.600	.613436*
.1	.712687	.61	.613485	.601	.613436*
.2	.679556	.62	.613649	.602	.613437
.3	.652599	.63	.613932	.603	.613439
.4	.632035	.64	.614341	.604	.613442
.5	.618514	.65	.614879	.605	.613446
.6	.613436*	.66	.615554	.606	.613452
.7	.619747	.67	.616370	.607	.613458
.8	.644395	.68	.617336	.608	.613466
.9	.709681	.69	.618459	.609	.613475

Consequently, the maximum likelihood estimate of the spatial autocorrelation parameter is $\hat{\rho} = 0.6005$. One should note that this estimate is higher than that for each of the preceding three problems, and its t-statistic for MC also is farther away from zero than are those for the preceding three problems. Furthermore, one should note that this form of modified regression analysis, even though in its classical form it is independent of which variable is labelled X and which is labelled Y, may differ in the estimate $\hat{\rho}$ it renders if these two variables are switched; one should consult Griffith (1988b) and Mardia (1988) for a better understanding of this problem.

The fifth problem investigated in Chapter 3 has to do with the construction of trend surface regression models. As in previous problems discussed in this Chapter, the t-statistic for MC is insignificant, even at the 10% level. Hence, fitting of a SAR model to these data may well falter, in that the MSE may fail to converge upon a minimum within the feasible parameter space interval (-1,1). Once more, as an illustrative exercise, consider the estimation of a SAR model here. The iterations generated by the algorithm presented in Section 4.1, for this problem, are summarized in the following tabulations:

LET K20 = -.9		LET K20 = -.19		LET K20 = -.079	
LET K21 = .1		LET K21 = .01		LET K21 = .001	
$\hat{\rho}$	MSE	$\hat{\rho}$	MSE	$\hat{\rho}$	MSE
-.9	156.514	-.19	122.596	-.079	121.975
-.8	147.718	-.18	122.496	-.078	121.974
-.7	140.575	-.17	122.404	-.077	121.974
-.6	134.823	-.16	122.321	-.076	121.973
-.5	130.274	-.15	122.247	-.075	121.973
-.4	126.795	-.14	122.182	-.074	121.973
-.3	124.292	-.13	122.125	-.073	121.972*
-.2	122.706	-.12	122.077	-.072	121.972*
-.1	122.009*	-.11	122.039	-.071	121.972*
0.0	122.202	-.10	122.009	-.070	121.972*
.1	123.321	-.09	121.988	-.069	121.973
.2	125.441	-.08	121.976	-.068	121.973
.3	128.691	-.07	121.972*	-.067	121.973
.4	133.283	-.06	121.978	-.066	121.974
.5	139.565	-.05	121.993	-.065	121.974
.6	148.129	-.04	122.017	-.064	121.975
.7	160.088	-.03	122.049	-.063	121.975
.8	177.891	-.02	122.091	-.062	121.976
.9	209.024	-.01	122.142	-.061	121.977

Consequently, the maximum likelihood estimate of the spatial autocorrelation parameter is $\hat{\rho} = -0.0715$. One should note that this estimate is lower than that for each of the preceding four problems, and its t-statistic for MC also is closer to zero than are those for the preceding four problems. Again there is a consistency in the emerging relationship between the MC t-statistic for this problem and the estimate of the spatial autocorrelation parameter; this estimate is the lowest thus far, and its affiliated MC t-statistic is the closest to zero.

Before evaluating the last two numerical examples that are treated in this Discussion Paper, this relationship between the estimated spatial autocorrelation parameter value and its affiliated MC and MC t-statistic will be further explored; there does appear to be a direct relationship between this first variable and these latter two variates. Testing each for normality results in the following modified Shapiro-Wilk statistics:

estimate of the spatial autocorrelation parameter: 0.910
MC calculated for the regression residuals: 0.985
MC t-statistic calculated for the regression residuals: 0.934

In all three cases this statistic is well beyond the 5% level critical value of 0.8804, and hence suggests that the associated parent population frequency distributions conform to a normal distribution; these three values may well be correlated, though, since they are based upon the same sample. Meanwhile, the correlation between the estimated spatial autocorrelation parameter value and its affiliated MC value is 0.937, which is significant at the 5% level (t = 4.65). A plot of the relationship between the spatial autocorrelation parameter estimate and its corresponding MC t-statistic suggests that it is nonlinear, but direct. These data lead to the following estimated regression equation:

```
The regression equation is
C1 = 0.0546 + 2.17 C2

Predictor       Coef      Stdev    t-ratio       p
Constant     0.05464    0.08467       0.65   0.565
C2            2.1701     0.4676       4.64   0.019

s = 0.1090     R-sq = 87.8%     R-sq(adj) = 83.7%

Analysis of Variance
SOURCE       DF         SS          MS          F        p
Regression    1    0.25604     0.25604      21.54    0.019
Error         3    0.03566     0.01189
Total         4    0.29170
```

These results suggest that a one-to-one correspondence exists between the calculated MC value and $\hat{\rho}$, such that $\hat{\rho}$ can be computed as roughly twice this statistic's value (the slope is 2.17, and the intercept term in the population is expected to be zero, given the t-statistic of 0.65). If this finding holds in general (considerable subsequent meticulous research is needed in order to establish such an empirical rule), then a reasonable and quick estimate of the SAR spatial autocorrelation parameter is revealed by the value of MC (a considerably easier and less numerically intensive calculation).

The remaining two numerical examples have been isolated from this foregoing analysis because both are examples of non-convergence, presumably due to model overparameterization. These two problems are the initial regression results reported in Chapter 2, and the quadratic trend surface results reported in Chapter 3. The first stage iterations generated by the algorithm presented in Section 4.1, for these two problems, are summarized in the following tabulations:

LET K20 = -.9		LET K21 = .1	
initial regression model		quadratic trend surface model	
$\hat{\rho}$	MSE	$\hat{\rho}$	MSE
-.999	8.05973	-.999	14.3141
-.9	8.22823	-.9	15.0566
-.8	8.44161	-.8	15.9236
-.7	8.68890	-.7	16.9033
-.6	8.96259	-.6	17.9947

-.5	9.25657
-.4	9.56582
-.3	9.88638
-.2	10.21550
-.1	10.55190
0.0	10.89640
.1	11.25290
.2	11.62960
.3	12.04100
.4	12.51070
.5	13.07690
.6	13.80250
.7	14.80010
.8	16.30510
.9	19.01210

-.5	19.2003
-.4	20.5255
-.3	21.9789
-.2	23.5726
-.1	25.3228
0.0	27.2510
.1	29.3856
.2	31.7649
.3	34.4418
.4	37.4913
.5	41.0265
.6	45.2288
.7	50.4260
.8	57.3336
.9	68.2257

In both cases the $\hat{\rho}$ value of -0.999 has been evaluated, too, to demonstrate that in fact this mean square error function is not achieving a minimum between -1.0 and -0.9. In both cases the mean square error quantity continually increases; in other situations it could just as easily decrease, without achieving a minimum, as it moves toward 1.0. Inspection of the associated MC t-statistics for these two problems reveals that both are very close to zero (-0.65 and -0.68, respectively). In both cases one would expect that non-convergence is attributable to overparameterization of the regression model.

4.3. Benchmark output for the California plant species data

The results reported in Upton and Fingleton (1985) are duplicated here in order to verify that the MINITAB computer code macros indeed do yield correct calculations; these data come from Tables 5.2 (p. 273) and 5.6(b) (p. 292). First comparable results will be presented here, and second improvements offered by the algorithm of this Discussion Paper will be outlined.

Upton and Fingleton (1985, p. 292) present matrix **W**, rather than its underlying symmetric form. Hence, the eigenvalues of this asymmetric matrix cannot be extracted by MINITAB; rather, those eigenvalues reported in Table 5.10 (p. 299) have been used as part of the input to this analysis. Reading the eigenvalues rather than calculating them internally compels the making of a slight modification to the MINITAB code for analysis; MINITAB column C48 is read from the digital file, as is matrix **W**, and so the EIGEN command and the matrix **C** conversion macro are not part of this particular analysis. A connectivity matrix **C** has been constructed here from matrix **W**, so that MC and its accompanying t-statistic can be calculated. These modifications have necessitated the construction of an additional separate file for this example. Furthermore, as will be shown subsequently, Upton and Fingleton (1985, p. 289) do not calculate the spatial autocorrelation parameter estimate with the same degree of precision as is done with the present algorithm; hence, $\hat{\rho}$ is set to 0.75 in order to replicate their results, but also will be estimated to a third decimal place in this section. Therefore, their California plant species example results are obtained by executing the following sequence of

MINITAB commands:

```
EXEC 'a:UFEXAMP.TST'
EXEC 'a:CLASSIC.REG'
EXEC 'a:UFSTART.SAR'
```

Only selected relevant portions of the output from these commands will be presented here. Of course the various CRT screen displays, such as the symmetry check, appear, too.

The OLS regression results obtained are as follows:

```
cl is Y: c2-ck3 are predictors


The regression equation is
C1 = - 1668 + 0.164 C2 + 0.116 C3 + 52.5 C4


Predictor        Coef       Stdev    t-ratio        p
Constant      -1668.2       370.0      -4.51    0.000
C2            0.16418     0.05024       3.27    0.004
C3            0.11647     0.02813       4.14    0.000
C4              52.51       10.83       4.85    0.000

s = 165.9        R-sq = 87.6%      R-sq(adj) = 85.9%

Analysis of Variance
SOURCE         DF          SS          MS          F        p
Regression      3     4263957     1421319      51.62    0.000
Error          22      605744       27534
Total          25     4869700

SOURCE         DF      SEQ SS
C2              1     3290883
C3              1      325737
C4              1      647337

Unusual Observations
Obs.      C2         C1        Fit Stdev.Fit  Residual    St.Resid
 13        0       12.0      353.4      63.0    -341.4      -2.22R
 19     4260     1450.0     1525.1     131.0     -75.1      -0.74 X
 21      529     1060.0      723.3      62.6     336.7       2.19R

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

Within the limits of acceptable and slight rounding error, these results agree with those reported in Upton and Fingleton (1985, p. 275).

Results for the SAR model are as follows:

```
* NOTE *      C47 is highly correlated with other  predictor variables
* NOTE *      C18 is highly correlated with other  predictor variables
```

```
The regression equation is
C15 = - 861 C47 + 0.148 C16 + 0.102 C17 + 27.1 C18

Predictor        Coef       Stdev    t-ratio        p
Noconstant
C47            -860.9       415.6      -2.07    0.050
C16           0.14787     0.04115       3.59    0.002
C17           0.10173     0.02681       3.80    0.001
C18             27.06       12.48       2.17    0.041

s = 154.7

Analysis of Variance
SOURCE        DF          SS          MS         F        p
Regression     4     3042078      760520     31.77    0.000
Error         22      526570       23935
Total         26     3568648

SOURCE        DF      SEQ SS
C47            1      652612
C16            1     2007856
C17            1      268997
C18            1      112614

Unusual Observations
Obs.     C47       C15        Fit Stdev.Fit  Residual   St.Resid
  2     1.05     171.8      428.3      97.1    -256.5     -2.13R
 19     0.26    1022.6     1062.2     121.5     -39.7     -0.41 X
 21     0.26     721.5      319.6      43.7     401.8      2.71R

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.

K20      0.750000
K15      23935.0
```

Within the limits of acceptable and slight rounding error, these regression coefficient results also agree with those reported in Upton and Fingleton (1985, p. 293). But, the standard errors, the mean square error, and the t-statistics are not correct. The final MINITAB macro, which is discussed in Chapter 5, will present the correct values for these assorted statistics; the problem here primarily has to do with degrees of freedom.

These foregoing results now will be augmented, in order to emphasize some of the advantages of the MINITAB macros being presented in this Discussion Paper. First, MC results based upon the matrix **C** are provided by the macro a:CLASSIC.REG; for these data they are

```
Moran Coefficient (MC) calculated for residuals; printed as k4
Expected value of MC for residuals
Variance of MC for residuals
t-score calculated for MC; printed as k9, df printed as k10
K4       0.341540
```

```
K9       2.56557
K10      22.0000
Shapiro-Wilk test for normality performed on residuals
Correlation of C50 and C45 = 0.993
```

These results indicate that (a) there is significant spatial autocorrelation in the regression residuals obtained with a traditional OLS solution (the t-statistic equals 2.57, which is greater than the 5% critical value of 2.07), and (b) these residuals come from a population whose frequency distribution should conform to a normal distribution (the modified Shapiro-Wilk statistic equals 0.993, which is greater than the 5% critical value of approximately 0.992).

The three-stage estimation procedure promoted by the MINITAB algorithm yields slightly more precise results than are reported in Upton and Fingleton (1985, p. 293). The iterations generated by the algorithm presented in Section 4.1, for this problem, are summarized in the following tabulations:

LET K20 = -.9		LET K20 = .61		LET K20 = .741	
LET K21 = .1		LET K21 = .01		LET K21 = .001	
$\hat{\rho}$	MSE	$\hat{\rho}$	MSE	$\hat{\rho}$	MSE
-.9	32248.3	.61	24390.7	.741	23939.1
-.8	31172.8	.62	24339.0	.742	23938.4
-.7	30397.9	.63	24289.2	.743	23937.8
-.6	29802.6	.64	24241.5	.744	23937.2
-.5	29322.7	.65	24196.1	.745	23936.7
-.4	28917.7	.66	24153.2	.746	23936.2
-.3	28558.4	.67	24113.3	.747	23935.8
-.2	28221.5	.68	24076.4	.748	23935.5
-.1	27886.3	.69	24042.9	.749	23935.2
0.0	27533.8	.70	24013.3	.750	23935.0
.1	27145.9	.71	23987.7	.751	23934.8
.2	26706.7	.72	23966.7	.752	23934.8
.3	26204.7	.73	23950.6	.753	23934.7*
.4	25639.4	.74	23939.9	.754	23934.8
.5	25031.7	.75	23935.0*	.755	23934.9
.6	24444.0	.76	23936.5	.756	23935.1
.7	24013.3*	.77	23945.0	.757	23935.3
.8	24018.9	.78	23961.1	.758	23935.7
.9	25117.2	.79	23985.4	.759	23936.1

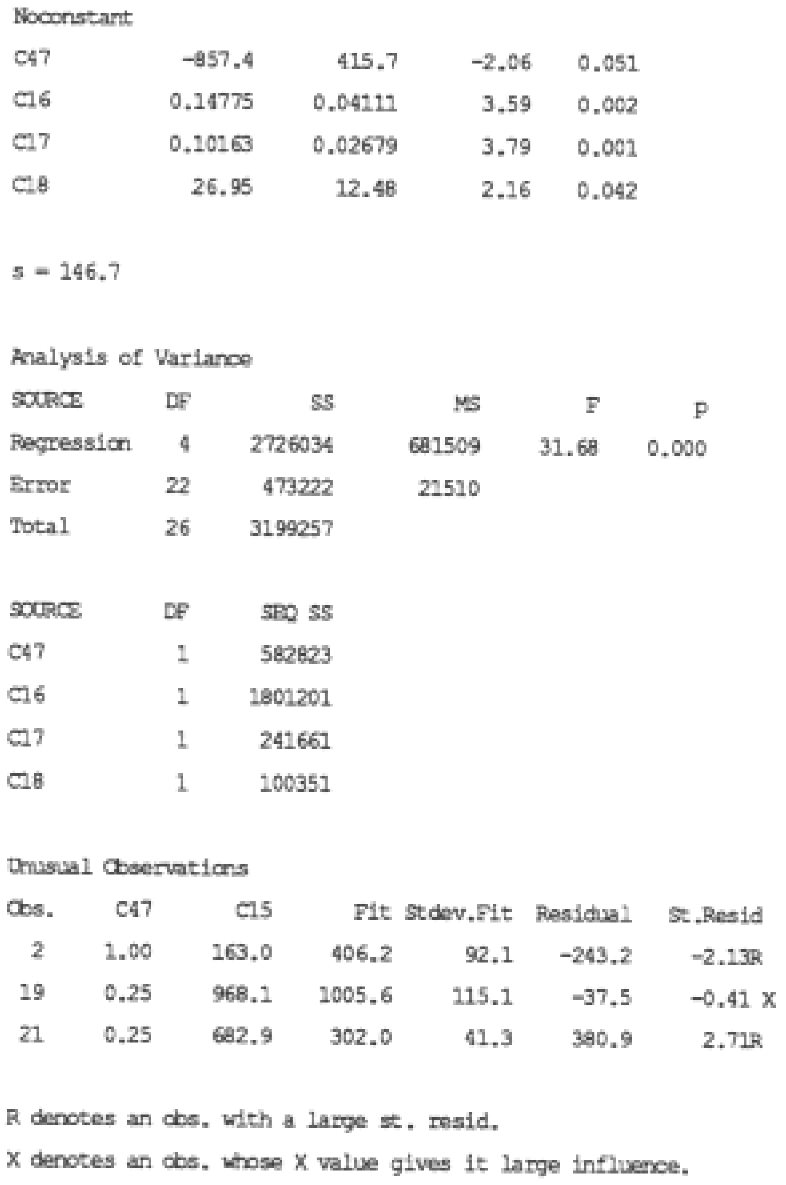
Consequently, the maximum likelihood estimate of the spatial autocorrelation parameter is $\hat{\rho} =$ 0.753, which is slightly larger than what Upton and Fingleton (1985, p. 293) report. The appropriate SAR results for this more precise estimate of the spatial autocorrelation parameter are

---

```
* NOTE *     C47 is highly correlated with other  predictor variables
* NOTE *     C18 is highly correlated with other  predictor variables

The regression equation is
C15 = - 857 C47 + 0.148 C16 + 0.102 C17 + 27.0 C18

Predictor        Coef        Stdev    t-ratio        p
```



```
Noconstant
C47          -857.4       415.7      -2.06    0.051
C16         0.14775     0.04111       3.59    0.002
C17         0.10163     0.02679       3.79    0.001
C18           26.95       12.48       2.16    0.042

s = 146.7

Analysis of Variance
SOURCE        DF          SS          MS          F         p
Regression     4     2726034      681509      31.68     0.000
Error         22      473222       21510
Total         26     3199257

SOURCE        DF      SEQ SS
C47            1      582823
C16            1     1801201
C17            1      241661
C18            1      100351

Unusual Observations
Obs.     C47       C15         Fit Stdev.Fit  Residual   St.Resid
  2     1.00     163.0       406.2      92.1    -243.2     -2.13R
 19     0.25     968.1      1005.6     115.1     -37.5     -0.41 X
 21     0.25     682.9       302.0      41.3     380.9      2.71R

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

As one can see, and as one should expect, these calculations differ only slightly from those reported in Upton and Fingleton (1985, p. 293).

APPENDIX 4-A.

CALIFORNIA DATA FROM UPTON AND FINGLETON

County	Number of Plant Species	Area	Maximum Elevation	Latitude	Eigenvalues of Island Isolation Weights Matrix
1	205	134	3950	28.2	1.00
2	163	98	4600	29.0	.90
3	420	96	2470	34.0	- .87
4	340	84	1560	34.0	.14
5	392	75	2125	33.3	- .01
6	235	56	1965	32.9	- .27
7	120	22	910	33.2	- .19
8	190	14	830	34.0	- .70
9	42	2.8	490	27.9	0.00
10	40	1.0	635	33.4	0.00
11	62	0.9	470	30.5	0.00
12	4	0.2	130	29.8	0.00
13	12	0.1	360	37.7	0.00
14	40	0.02	60	37.1	0.00
15	39	2.5	660	28.3	0.00
16	70	1.1	930	34.0	0.00
17	83	1.0	670	32.4	0.00
18	72	0.5	315	31.8	0.00
19	1450	4260	6535	33.0	0.00
20	1400	3324	5860	36.2	0.00
21	1060	529	2610	38.1	0.00
22	1200	1386	3810	37.3	0.00
23	640	320	3110	34.1	0.00
24	680	110	3985	34.4	0.00
25	640	45	930	37.8	0.00
26	370	5.9	750	37.9	0.00

Appendix 4-A, Continued

County	Connectivity Matrix 1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1
21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	1
22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	1
23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0
24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1
26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0

County	Stochastic Version of Lower Right-hand Partition of Connectivity Matrix 19	20	21	22	23	24	25	26
19	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000
20	0.0000	0.0000	0.1162	0.3285	0.0752	0.1708	0.1665	0.1427
21	0.0000	0.0082	0.0000	0.0493	0.0000	0.0000	0.3016	0.6409
22	0.0000	0.0797	0.1693	0.0000	0.0000	0.0000	0.4759	0.2751
23	0.2139	0.0933	0.0000	0.0000	0.0000	0.6928	0.0000	0.0000
24	0.0000	0.2342	0.0000	0.0000	0.7658	0.0000	0.0000	0.0000
25	0.0000	0.0080	0.2041	0.0938	0.0000	0.0000	0.0000	0.6941
26	0.0000	0.0057	0.3649	0.0456	0.0000	0.0000	0.5838	0.0000

CHAPTER 5.
COMPARISON OF OLS AND SAR RESULTS: EVALUATING THE SAR SOLUTION

Two questions remain to be answered here. First, if the t-statistics obtained with the modified OLS regression are incorrect, then what are the correct t-statistics? Similarly, what is the t-statistic for the spatial autocorrelation parameter estimate? Second, does this more complicated regression modelling approach really make an important difference in an analysis? Answers to these sorts of questions will be outlined in this chapter.

In calculating the t-statistics for SAR regression coefficients, one needs to determine their correct covariance matrix, the correct mean square error, and the correct number of degrees of freedom. The correct asymptotic covariance matrix is given by $[\mathbf{X}^t(\mathbf{I} - \hat{\rho}\mathbf{W})^t(\mathbf{I} - \hat{\rho}\mathbf{W})\mathbf{X}]^{-1}\hat{\sigma}^2$. Although this matrix is a function of the estimated spatial autocorrelation parameter value and the mean square error value, it does not covary with either the asymptotic variance of $\hat{\rho}$ or the asymptotic variance of $\hat{\sigma}^2$. The mean square error estimate, $\hat{\sigma}^2$, is computed from the sum of square error term, which can be written as $(\mathbf{Y} - \hat{\mathbf{Y}})^t(\mathbf{Y} - \hat{\mathbf{Y}})$; the value for this term that is produced in Chapter 4 is based upon the filtered vector $(\mathbf{I} - \hat{\rho}\mathbf{W})\mathbf{Y}$, which is why it is incorrect. For an SAR model, the vector of predicted Y values may be written as

$$\hat{\mathbf{Y}} = \hat{\rho}\mathbf{W}\mathbf{Y} + (\mathbf{I} - \hat{\rho}\mathbf{W})\mathbf{X}\mathbf{b} \quad ,$$

where the spatial autocorrelation parameter estimate is obtained from the iterative and interactive procedure outlined in Chapter 4. Again, one should note that the Jacobian term does not appear here, since it is used only in calculating the estimate $\hat{\rho}$, and is not part of the model specification. The predicted vector $\hat{\mathbf{Y}}$ also can be correlated with the observed vector $\mathbf{Y}$ in order to determine the multiple correlation coefficient goodness-of-fit measure for this final regression equation. And, since one degree of freedom should be subtracted for each parameter that is estimated, an additional degree of freedom is lost for the estimate of the spatial autocorrelation parameter; the MINITAB regression procedure does not realize that this estimate has been obtained, and thus fails to adjust degrees of freedom for it. Division by the number of degrees of freedom will render an unbiased estimate of the mean square error; Upton and Fingleton (1985, p. 285) use the maximum likelihood estimate of this parameter, which instead involves division by n. In addition, the standard error of $\hat{\rho}$ is obtained by calculating the square root of a fraction whose numerator is (n/2), and whose denominator is

$$(n/2)\{\mathrm{tr}[(\mathbf{I} - \hat{\rho}\mathbf{W}^{t})^{-1}\mathbf{W}^{t}\mathbf{W}(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}] - (-\sum_{i=1}^{i=n} \lambda_i^2/(1 - \rho\lambda_i)^2\} - \{\mathrm{tr}[\mathbf{W}(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}]\}^2 \quad ,$$

This fraction results from a matrix inversion. If the term $\mathrm{tr}[\mathbf{W}(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}]$ is zero, then this fraction reduces to

$$1/\{\mathrm{tr}[(\mathbf{I} - \hat{\rho}\mathbf{W}^{t})^{-1}\mathbf{W}^{t}\mathbf{W}(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}] - (-\sum_{i=1}^{i=n} \lambda_i^2/(1 - \rho\lambda_i)^2\} \quad .$$

This later simplification furnishes a useful approximation for when extremely large numbers are involved, and hence the matrix in question numerically is not able to be inverted.

Whether or not the SAR results are an improvement over conventional OLS results may be assessed by seeing whether or not inferences about parameters change, given the correct t-statistics, and whether or not the percent of variance accounted for in variate Y increases by a marked amount. At this point in an analysis, though, attempts to diagnose remanent spatial autocorrelation are ill-advised. Although the final regression residuals may be tested for normality, Cliff and Ord (1981, p. 240) note that

> ... the standard tests for autocorrelation cease to be valid when the model contains an autoregressive component. ... The greater complexity of the estimation procedures for spatial models [as opposed to time series models] is such that no satisfactory technique has been developed to handle the problem [of testing for residual autocorrelation once an autoregressive component has been added] ...

Because of this argument, the final SAR regression residuals determined with the algorithm presented in this Discussion Paper will not be tested for the presence of additional spatial autocorrelation.

5.1. The estimation algorithm

Individual vectors of the matrix $(\mathbf{I} - \hat{\rho}\mathbf{W})\mathbf{X}$ are produced by the macro MINITAB a:TRANS2.X, whose code appears as

```
a:TRANS2.X

let ck19 = ck17 - k20*ck18
let k17 = k17 + 1
let k18 = k18 + 1
let k19 = k19 + 1
end
```

This macro differs from the macro a:TRANS.X in that it does not divide each vector by the Jacobian term that is needed in order to properly estimate the spatial autocorrelation parameter; the Jacobian is not part of a model specification, just part of an estimation procedure. The matrix inverse portion of the asymptotic regression coefficients variance term can be obtained by using the MINITAB subcommand XPXINV with the command REGRESS, which will calculate $[\mathbf{X}^t(\mathbf{I} - \hat{\rho}\mathbf{W})^t(\mathbf{I} - \hat{\rho}\mathbf{W})\mathbf{X}]^{-1}$. Meanwhile, the REGRESS subcommand COEF C61 will store the sample regression coefficients $\mathbf{b}$ in MINITAB column C61. Including additional columns on the REGRESS command line allows predicted values to be captured; here these predictions are stored in MINITAB column C30. These predicted values are for the filtered Y variate, and are calculated as $(\mathbf{I} - \hat{\rho}\mathbf{W})\mathbf{Xb}$. The eigenvalue term

$$-\sum_{i=1}^{i=n} \lambda_i^2/(1 - \rho\lambda_i)^2$$

needed for the asymptotic variance estimate of the spatial autocorrelation parameter estimate is calculated with the MINITAB command

```
      LET K30 = -SUM((C48**2)/((1 - K20*C48)**2))
```

The matrix trace $\mathrm{tr}[\mathbf{W}(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}]$ needed for the asymptotic variance calculations is obtained with the MINITAB commands

```
      MULT K20 M2 M5
      DIAG C49 M4
      SUB M5 M4 M6
      INVERT M6 M7
      MULT M2 M7 M8
      DIAG M8 C45
      LET K31 = SUM(C45)
```

Once the matrix $[\mathbf{W}(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}$ is calculated (MINITAB matrix M8), the matrix trace $\mathrm{tr}[(\mathbf{I} - \hat{\rho}\mathbf{W}^t)^{-1}\mathbf{W}^t\mathbf{W}(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}]$ also needed for the asymptotic variance calculations is obtained with the MINITAB commands

```
      TRANS M8 M9
      MULT M9 M8 M10
      DIAG M10 C45
      LET K32 = SUM(C45)
```

Since the predicted vector $\hat{\mathbf{Y}}$ is the sum of the terms $\hat{\rho}\mathbf{WY}$ and $(\mathbf{I} - \hat{\rho}\mathbf{W})\mathbf{Xb}$, which is stored in MINITAB column C30, it is calculated with the MINITAB command

```
      LET C32 = K20*C51 + C30
```

recalling that K20 is the spatial autocorrelation parameter estimate, and column C51 is constructed by the subroutine a:TRANS2.X. The correct mean square error can be calculated now with the MINITAB command

```
      LET K15 = SUM((C1 - C32)**2)/(K1 - K3 - 1)
```

where K1 is the number of areal units, K3 is the number of regression parameters (including the intercept), and the additional degree of freedom is subtracted because of the estimate $\hat{\rho}$. The correct standard errors for the regression coefficients are calculated with the MINITAB command

```
LET C63 = SQRT(K15*C62)
```

while the corresponding t-statistics are calculated with the MINITAB command

```
LET C64 = C61/C63
```

The multiple correlation coefficient is computed with the MINITAB command

```
CORR C1 C32
```

Finally, the t-statistic for the estimate $\hat{\rho}$ is secured by first executing either the macro a:DEPSE.RHO or a:INDEPSE.RHO, which returns the appropriate number K31, and then using the MINITAB command

```
LET K35 = K31/K15
```

All of these commands are contained in the macro entitled a:FINAL.SAR; the code for this file appears as

```
a:FINAL.SAR

let k17 = 1
let k18 = 51
let k19 = 15
exec 'a:trans2.x' k3
mult m2 c49 c50
let c47 = c49 - k20*c50
let k19 = k19 - 1
NOTE regression model with filtered errors
regress c15 k3 c47 c16-ck19 c31 c30;
noconstant;
xpxinv m11;
coef c61.
let k30 = -sum((c48**2)/((1 - k20*c48)**2))
mult k20 m2 m5
diag c49 m4
sub m5 m4 m6
invert m6 m7
mult m2 m7 m8
diag m8 c45
let k31 = sum(c45)
trans m8 m9
mult m9 m8 m10
diag m10 c45
let k32 = sum(c45)
let c32 = k20*c51 + c30
let k15 = sum((c1 - c32)**2)/(k1 - k3 - 1)
NOTE correct mean square error
print k15
NOTE correct regression coefficients (c61), standard errors (c63),
t-statistics (c64)
diag m11 c62
```

```
let c63 = sqrt(k15*c62)
let c64 = c61/c63
print c61 c63 c64
NOTE correlation between the expected and observed values
corr c1 c32
NOTE Shapiro-Wilk statistic for filtered residuals
let c33 = c1 - c32
nscores c33 c34
corr c33 c34
NOTE if k35 is very close to zero, then execute a:INDEPSE.RHO; otherwise
execute a:DEPSE.RHO
let k35 = k31/k15
print k35
end
```

The value of constant K31 is returned from a subroutine. Before it is returned, though, instructions appear on the CRT screen notifying the user of the numerical value for tr$[\mathbf{W}(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}]$; if this value is near zero, then the matrix inversion will be problematic. Hence, a value close to zero means that the macro a:INDEPSE.RHO should be executed; its code appears as

```
a:INDEPSE.RHO

NOTE spatial autocorrelation parameter estimate (k20), t-statistic
(k36), df (k37)
let k35 = 1/sqrt(k32-k30)
let k36 = k20/k35
let k37 = k1 - k3 - 1
print k20,k36,k37
end
```

A value substantially different from zero means that the macro a:DEPSE.RHO should be executed; its code appears as

```
a:DEPSE.RHO

NOTE spatial autocorrelation parameter estimate (k20), t-statistic
(k36), df (k37)
let c38(1) = k1/(2*(k15**2))
let c38(2) = k31/k15
let c39(1) = c38(2)
let c39(2) = k32-k30
copy c38-c39 m11
invert m11 m12
copy m12 c38-c39
let k35 = sqrt(c39(2))
let k36 = k20/k35
let k37 = k1 - k3 - 1
print k20,k36,k37
end
```

These executions terminate the SAR algorithm.

5.2. Illustrative evaluations for selected problems from Chapter 4

Only two examples discussed in Chapters 2 and 3 exhibit a significant level of spatial autocorrelation. Specifying an SAR model for the remaining examples is questionable. Accordingly, only those two situations in which significant spatial autocorrelation is present will be evaluated in this section.

The first illustration in which significant non-zero spatial autocorrelation has been uncovered is the inference about the population mean problem. First the following MINITAB command must be executed:

```
LET K20 = .578
```

This is the value of the spatial autocorrelation parameter estimate identified in the three-stage iterative and interactive estimation procedure outlined in Chapter 4, which uses the MINITAB macro a:CLASSIC.SAR. The inference about the population mean problem requires the making of a slight adjustment to the MINITAB computer code presented in Section 5.1; the regression command appearing on the ninth line must be rewritten as

```
REGRESS C15 K3 C47 C31 C30;
```

This modified form of the final solution is housed in the macro a:FINALM.SAR; all other problems use the macro a:FINAL.SAR. Hence, executing the MINITAB macro a:FINALM.SAR for these data generates the CRT screen display

```
regression model with filtered errors


The regression equation is
C15 = 9387 C47


Predictor        Coef       Stdev    t-ratio        p
Noconstant
C47              9387        6839       1.37    0.190

s = 11544

Analysis of Variance
SOURCE       DF          SS          MS         F        p
Regression    1   251078336   251078336      1.88    0.190
Error        15  1998944384   133262960
Total        16  2250022656

Unusual Observations
Obs.     C47       C15       Fit Stdev.Fit  Residual   St.Resid
 10    0.422     39481      3961      2886     35519      3.18R

R denotes an obs. with a large st. resid.

* NOTE  * ALL VALUES IN COLUMN ARE IDENTICAL
```

Therefore the estimate of the population mean has changed from 9746 to 9387.

The correct inferential statistics for this problem are as follows:

```
correct mean square error
K15      142781728


correct regression coefficients (c61), standard errors (c63), t-statistics (c64)
ROW       C61       C63       C64
  1   9387.12   7078.87   1.32608


correlation between the expected and observed values
Correlation of C1 and C32 = 0.654
```

In other words, the correct mean square error is 142781728 rather than 177323184, the OLS population mean estimate appears to be slightly inflated, and nearly 43% of the variance exhibited by the Y variable can be statistically explained by its spatially lagged term. The standard error of the sample mean seems too low, as well, increasing from 3329 to 7079. This change in the standard error estimate has completely reversed the statistical inference for this problem. In the OLS analysis this mean is determined to be significantly different from zero; in this SAR analysis it is found not to be significantly different from zero. Consequently, overlooking the positive spatial autocorrelation latent in the geographic distribution of density of coffee production in the Mayaguez Agricultural Administrative Region of Puerto Rico leads one to erroneously infer that it is non-zero in the population.

The new residuals still suggest the absence of a normal distribution in the population.

```
Shapiro-Wilk statistic for filtered residuals
Correlation of C33 and C34 = 0.813
```

The appropriate critical Shapiro-Wilk statistic value here, conservatively speaking, is approximately 0.9343. This failure to satisfy the normality assumption should be viewed as quite troublesome.

Finally, the spatial autocorrelation parameter estimate is significantly different from zero.

```
if k35 is very close to zero, then execute a:INDEPSE.RHO; otherwise execute a:DEPSE.RHO
K35      0.000000025


spatial autocorrelation parameter estimate (k20), t-statistic (k36), df (k37) K20      0.578000
K36      2.84872
K37      14.0000
```

Because K35 is so close to zero, the MINITAB macro a:INDEPSE.RHO was selected for execution. The appropriate critical t-statistic value here is 2.14.

The second instance in which significant non-zero spatial autocorrelation has been uncovered is the z-score regression problem. This final step of the analysis is initiated with the following MINITAB command execution:

```
LET K20 = .6005
```

This is the value of the spatial autocorrelation parameter estimate identified in the three-stage iterative and interactive estimation procedure outlined in Chapter 4, which uses the MINITAB macro a:CLASSIC.SAR. Executing the MINITAB macro a:FINAL.SAR for this data generates the CRT screen display

```
regression model with filtered errors


The regression equation is
C15 = 0.032 C47 + 0.495 C16


Predictor        Coef        Stdev    t-ratio        p
Noconstant
C47            0.0316       0.4629       0.07    0.947
C16            0.4953       0.1950       2.54    0.024

s = 0.7387

Analysis of Variance
SOURCE       DF         SS          MS         F       p
Regression    2     3.5231      1.7616      3.23   0.070
Error        14     7.6392      0.5457
Total        16    11.1623

SOURCE       DF     SEQ SS
C47           1     0.0022
C16           1     3.5209

Unusual Observations
Obs.     C47       C15       Fit Stdev.Fit  Residual   St.Resid
 10    0.400     2.647     1.056     0.459     1.591     2.75RX

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.

* NOTE  * ALL VALUES IN COLUMN ARE IDENTICAL
```

Therefore the two regression coefficients embrace the following changes:

b_0 has gone from 0 in OLS (by construction) to 0.0316 in SAR
b_1 has gone from 0.546 in OLS to 0.495 in SAR

This second value no longer is equivalent to a traditional correlation coefficient, though.

The correct inferential statistics for this problem are as follows:

```
correct mean square error
K15       0.587631


correct regression coefficients (c61), standard errors (c63), t-statistics (c64)
ROW       C61        C63        C64
  1   0.031566   0.480356   0.06571
  2   0.495301   0.202347   2.44778


correlation between the expected and observed values
Correlation of C1 and C32 = 0.717
```

In other words, the correct mean square error is 0.5876 rather than 0.8672, the OLS sample correlation between the two variables in question appears to be slightly inflated, and conditionally nearly 22% of the variance exhibited by the Y variable can be statistically explained by its spatially lagged term. The standard error of the correlation coefficient seems too large, as well, decreasing from 0.2239 to 0.2023.

The new residuals still suggest the presences of a normal distribution in the population.

```
Shapiro-Wilk statistic for filtered residuals
Correlation of C33 and C34 = 0.946
```

The appropriate critical Shapiro-Wilk statistic value here, conservatively speaking, is approximately 0.9302.

Finally, the spatial autocorrelation parameter estimate is significantly different from zero.

```
if k35 is very close to zero, then execute a:INDEPSE.RHO; otherwise execute a:DEPSE.RHO
 K35      6.57024


spatial autocorrelation parameter estimate (k20), t-statistic (k36), df (k37) K20      0.600500
K36      2.97607
K37      13.0000
```

The value of K35 indicates without a doubt that the MINITAB macro a:DEPSE.RHO should be selected for execution. The appropriate critical t-statistic value here is 2.16.

5.3. Benchmark output for the California plant species data

Again the results reported in Upton and Fingleton (1985) are duplicated here in order to verify that the MINITAB computer code macros generate correct results.

Once the three-stage iterative and interactive estimation of the spatial autocorrelation parameter is completed, using the macro a:CLASSIC.SAR, then its value must be set equal to the constant K20. Because Upton and Fingleton (1985) have used the value of 0.75, then the MINITAB command

```
LET K20 = .75
```

needs to be executed. Next the final regression results are obtained by executing the MINITAB command

```
EXEC 'a:FINAL.SAR' ↵
```

which results in the following CRT screen displays:

```
regression model with filtered errors
* NOTE *      C47 is highly correlated with other  predictor variables
* NOTE *      C18 is highly correlated with other  predictor variables

The regression equation is
C15 = - 861 C47 + 0.148 C16 + 0.102 C17 + 27.1 C18

Predictor        Coef       Stdev    t-ratio       p
Noconstant
C47            -860.9       415.6      -2.07   0.050
C16           0.14787     0.04115       3.59   0.002
C17           0.10173     0.02681       3.80   0.001
C18             27.06       12.48       2.17   0.041

s = 146.8

Analysis of Variance
SOURCE       DF          SS          MS         F        p
Regression    4     2737430      684358     31.77    0.000
Error        22      473836       21538
Total        26     3211266

SOURCE       DF      SEQ SS
C47           1      587256
C16           1     1806779
C17           1      242059
C18           1      101336

Unusual Observations
Obs.     C47      C15        Fit Stdev.Fit  Residual   St.Resid
  2     1.00    163.0      406.3      92.1    -243.3     -2.13R
 19     0.25    970.0     1007.7     115.2     -37.7     -0.41 X
 21     0.25    684.4      303.2      41.4     381.2      2.71R

R denotes an obs. with a large st. resid.
X denotes an obs. whose X value gives it large influence.
```

These results are followed by the display of the correct mean square error.

```
correct mean square error
 K15       22563.6
```

This value agrees with that reported by Upton and Fingleton (1985, p. 293), within the limits of acceptable and slight rounding error, because it is the unbiased counterpart to the biased estimate, which is computed as (21/26)*K15 = (21/26)*22563.6 = 18224.5 [Upton and Fingleton (1985, p. 275) divide this value by 10^6].

The correct standard errors and corresponding t-statistics for the regression coefficients, and the correct multiple correlation coefficient are displayed next.

```
correct regression coefficients (c61), standard errors (c63), t-statistics
(c64)
ROW       C61        C63        C64
  1  -860.931    425.393   -2.02385
  2     0.148      0.042    3.51083
  3     0.102      0.027    3.70790
  4    27.061     12.769    2.11923

correlation between the expected and observed values
Correlation of C1 and C32 = 0.951
```

Once more these standard errors and t-statistics agree with those presented in Upton and Fingleton (1985, p. 293), within the limits of acceptable and slight rounding error, when multiplied by $\sqrt{21/26}$. In other words,

$$\begin{aligned}
425.393*0.898717 &= 382.3 \\
0.042*0.898717 &= 0.038;\ t = 3.91 \\
0.027*0.898717 &= 0.025;\ t = 4.13 \\
12.769*0.898717 &= 11.476;\ t = 2.36
\end{aligned}$$

Meanwhile, $R^2 = (0.951)^2 = 0.904$.

Succeeding this CRT screen display is one for the normality test results.

```
Shapiro-Wilk statistic for filtered residuals
Correlation of C33 and C34 = 0.984
```

Clearly this model assumption is satisfied for these California plant species data; Upton and Fingleton (1985) do not report this statistic.

Finally, using the unbiased mean square error estimate rather than the biased estimate that is employed by Upton and Fingleton (1985), such that the sample size in the denominator is replaced with the number of degrees of freedom, the test of the spatial autocorrelation parameter estimate, $\hat{\rho}$, may be undertaken. First the value of $\mathrm{tr}[\mathbf{W}(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}]/\hat{\sigma}^2$ is displayed on the CRT screen.

```
if k35 is very close to zero, then execute a:INDEPSE.RHO; otherwise execute a:DEPSE.RHO
K35      0.000244072
```

Since this value is not approximately zero, although it is quite small, the macro a:DEPSE.RHO is selected for execution; this value equals 0.000302184 if the biased mean square error estimate is used, which agrees with the corresponding calculation alluded to by Upton and Fingleton (1985, p. 299). Consequently, the final CRT screen display is

```
spatial autocorrelation parameter estimate (k20), t-statistic (k36), df (k37) K20      0.750000
K36      5.57784
K37      21.0000
```

This t-statistic value agrees with that reported by Upton and Fingleton (1985, p. 293), within the limits of acceptable and slight rounding error. It will be the same regardless of which mean square error estimate is used; as is established in the introduction to this chapter, the asymptotic variance of $\hat{\rho}$ is not directly dependent upon $\hat{\sigma}^2$, although it is indirectly dependent upon the quantity tr$[\mathbf{W}(\mathbf{I} - \hat{\rho}\mathbf{W})^{-1}]/\hat{\sigma}^2$ through the matrix inversion operation. If either 0.0002 or 0.0003 are considered very close to zero, then this t-statistic changes to 5.66 (the difference between these two t-statistic values is attributable to specification error).

The interested reader should consult Upton and Fingleton (1985) for a discussion of the differences between these SAR results and their OLS counterparts, as well as an evaluation of this SAR solution.

CHAPTER 6.
SUMMARY

The execution sequences of macros for each of the problems treated in this Discussion Paper have the following concatenations:

multiple regression

```
EXEC 'a:START.___'
EXEC 'a:CLASSIC.REG'
EXEC 'a:START.SAR'
LET K20 = ?
LET K21 = .01
EXEC 'a:CLASSIC.SAR' 19
LET K20 = ?
LET K21 = .001
EXEC 'a:CLASSIC.SAR' 19
LET K20 = ?
EXEC 'a:FINAL.SAR'
EXEC 'a:INDEPSE.RHO'
            or
      'a:DEPSE.RHO'
```

inference about a population mean

```
EXEC 'a:START.___'
EXEC 'a:REG.MU'
EXEC 'a:START.SAR'
LET K20 = ?
LET K21 = .01
EXEC 'a:CLASSIC.SAR' 19
LET K20 = ?
LET K21 = .001
EXEC 'a:CLASSIC.SAR' 19
LET K20 = ?
EXEC 'a:FINALM.SAR'
EXEC 'a:INDEPSE.RHO'
            or
      'a:DEPSE.RHO'
```

analysis of variance

```
EXEC 'a:START.___'
EXEC 'a:REG.AOV'
EXEC 'a:START.SAR'
LET K20 = ?
LET K21 = .01
EXEC 'a:CLASSIC.SAR' 19
LET K20 = ?
LET K21 = .001
EXEC 'a:CLASSIC.SAR' 19
LET K20 = ?
```

two-groups discriminant function analysis

```
EXEC 'a:START.___'
EXEC 'a:CLASSIC.REG'
EXEC 'a:START.SAR'
LET K20 = ?
LET K21 = .01
EXEC 'a:CLASSIC.SAR' 19
LET K20 = ?
LET K21 = .001
EXEC 'a:CLASSIC.SAR' 19
LET K20 = ?
```

```
EXEC 'a:FINAL.SAR'
EXEC 'a:INDEPSE.RHO'
             or
      'a:DEPSE.RHO'
```

```
EXEC 'a:FINALM.SAR'
EXEC 'a:INDEPSE.RHO'
             or
      'a:DEPSE.RHO'
```

bivariate
correlation

```
EXEC 'a:START.___'
EXEC 'a:REG.COR'
EXEC 'a:START.SAR'
LET K20 = ?
LET K21 = .01
EXEC 'a:CLASSIC.SAR' 19
LET K20 = ?
LET K21 = .001
EXEC 'a:CLASSIC.SAR' 19
LET K20 = ?
EXEC 'a:FINAL.SAR'
EXEC 'a:INDEPSE.RHO'
             or
      'a:DEPSE.RHO'
```

trend surface models

```
EXEC 'a:START.___'
EXEC 'a:LINEAR.TSM'
EXEC 'a:QUADRATI.TSM'
EXEC 'a:CUBIC.TSM'
EXEC 'a:START.SAR'
LET K20 = ?
LET K21 = .01
EXEC 'a:CLASSIC.SAR' 19
LET K20 = ?
LET K21 = .001
EXEC 'a:CLASSIC.SAR' 19
LET K20 = ?
EXEC 'a:FINAL.SAR'
EXEC 'a:INDEPSE.RHO'
             or
      'a:DEPSE.RHO'
```

Meanwhile, the examples from Cliff and Ord (1981) and from Upton and Fingleton (1985) have the execution sequences

Cliff and Ord

```
EXEC 'a:START.TST'
EXEC 'a:CLASSIC.REG'
LET C1 = LOGTEN(C1)
LET C2 = LOGTEN(C2)
EXEC 'a:CLASSIC.REG'
```

Upton and Fingleton

```
EXEC 'a:UFEXAMP.TST'
EXEC 'a:CLASSIC.REG'
EXEC 'a:UFSTART.SAR'
EXEC 'a:CLASSIC.SAR' 19
LET K20 = .61
LET K21 = .01
EXEC 'a:CLASSIC.SAR' 19
LET K20 = .741
LET K21 = .001
EXEC 'a:CLASSIC.SAR' 19
LET K20 = .75
EXEC 'a:FINAL.SAR'
EXEC 'a:DEPSE.RHO'
```

Finally, the two complete examples from the Puerto Rican data set have the execution sequences

inference about a
population mean for
density of coffee production

```
EXEC 'a:DEMOMAY.MU'
EXEC 'a:REG.MU'
EXEC 'a:START.SAR'
LET K20 = .51
LET K21 = .01
EXEC 'a:CLASSIC.SAR' 19
LET K20 = .571
LET K21 = .001
EXEC 'a:CLASSIC.SAR' 19
LET K20 = .578
EXEC 'a:FINALM.SAR'
EXEC 'a:INDEPSE.RHO'
```

correlation between density
of sugarcane production and
density of number of farm families

```
EXEC 'a:DEMOMAY.COR'
EXEC 'a:REG.COR'
EXEC 'a:START.SAR'
LET K20 = .51
LET K21 = .01
EXEC 'a:CLASSIC.SAR' 19
LET K20 = .591
LET K21 = .001
EXEC 'a:CLASSIC.SAR' 19
LET K20 = .6005
EXEC 'a:FINAL.SAR'
EXEC 'a:DEPSE.RHO'
```

All files in this chapter are identified as being located on Disk Drive A because that is where they are housed, and from where they are retrieved, when the digital companion for this workbook is accessed.

CHAPTER 7.
REFERENCES

Anselin, L., and D. Griffith. "Do spatial effects really matter in regression analysis?" Papers of the Regional Science Association, 1988 in press.

Cliff, A., and J. Ord. Spatial Processes: Models and Applications. London: Pion, 1981.

Griffith, D. Spatial Autocorrelation: A Primer. Washington, D. C.: Association of American Geographers, Resource Publication 1985/4, 1987.

Griffith, D. "Estimating spatial autoregressive model parameters with commercial statistical packages," Geographical Analysis. Vol. 20 (1988a), 176-186.

Griffith, D. Advanced Spatial Statistics. Boston: Kluwer, 1988b.

Griffith, D. "Interpretation of selected standard regression diagnostics in the presence of spatial dependence," paper presented to the 35th North American meeting of the Regional Science Association, Toronto, November 11-13, 1988c.

Mardia, K. "Multi-dimensional multivariate Gaussian Markov random fields with application to image processing," unpublished paper, Department of Statistics, University of Leeds, 1988.

Mardia, K., and R. Marshall. "Maximum likelihood estimation of models for residual covariance in spatial regression," Biometrika. Vol. 71 (1984), 135-146.

Minitab, Inc. MINITAB Reference Manual, Release 6.1. State College, Pa.: Minitab, Inc., 1988.

Ord, K. "Estimation methods for models of spatial interaction," Journal of the American Statistical Association. Vol. 70 (1975), 120-126.

Upton, G., and B. Fingleton. Spatial Data Analysis by Example, Vol. 1. New York: Wiley, 1985.

APPENDIX A.
MSE PLOTS VERSUS
SPATIAL AUTOCORRELATION ESTIMATES

```
       MSE
3.50E+08+   *
        -
        -
        -      *
        -
2.80E+08+         *
        -            *
        -
        -               *
        -                  *
2.10E+08+                    *
        -                       *
        -                          *  *
        -                                *  *                   *
        -                                      * *  *4++342  *
1.40E+08+
        -
          --------+---------+---------+---------+---------+-------- rho
              -0.70     -0.35      0.00      0.35      0.70
```

mean square error (MSE) for the inference about a mean problem

```
       MSE
        -   *
        -
        -
2.10E+08+      *
        -
        -         *
        -
        -            *
1.80E+08+
        -               *
        -                                                       *
        -                  *
        -                    *
1.50E+08+                       *                            *
        -                          *
        -                             *  *              *  *
        -                                   * 43+343
          --------+---------+---------+---------+---------+-------- rho
              -0.70     -0.35      0.00      0.35      0.70
```

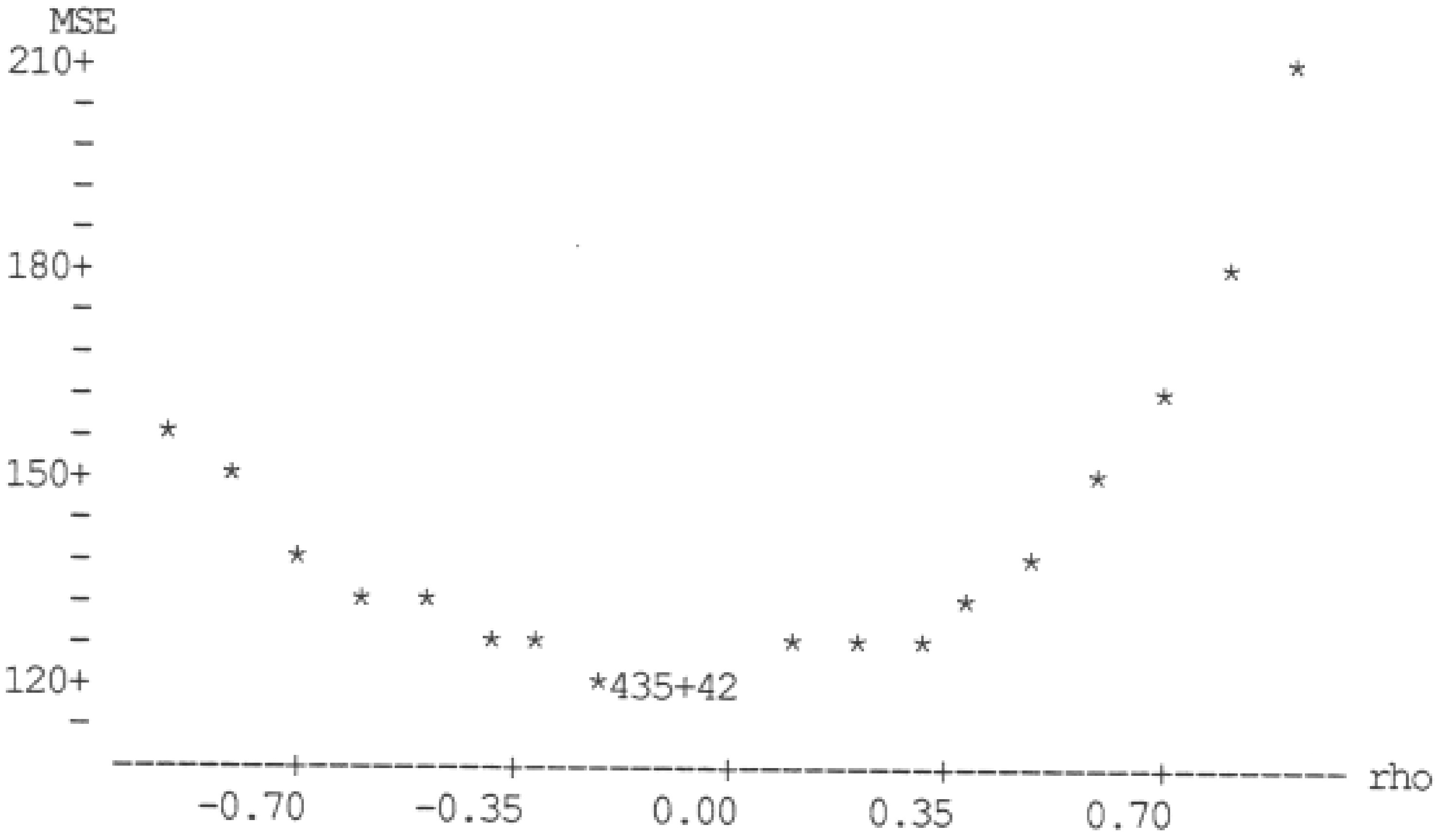
mean square error (MSE) for the analysis of variance problem

```
    MSE
     -   *
     -
     -
0.350+      *
     -
     -         *
     -
     -            *                                          *
0.300+               *
     -                  *
     -
     -                    *                               *
     -                       *  *
0.250+                             *                   *
     -                                *  *  43+343  *
     -
     -
       --------+---------+---------+---------+---------+-------- rho
           -0.70     -0.35      0.00      0.35      0.70
```

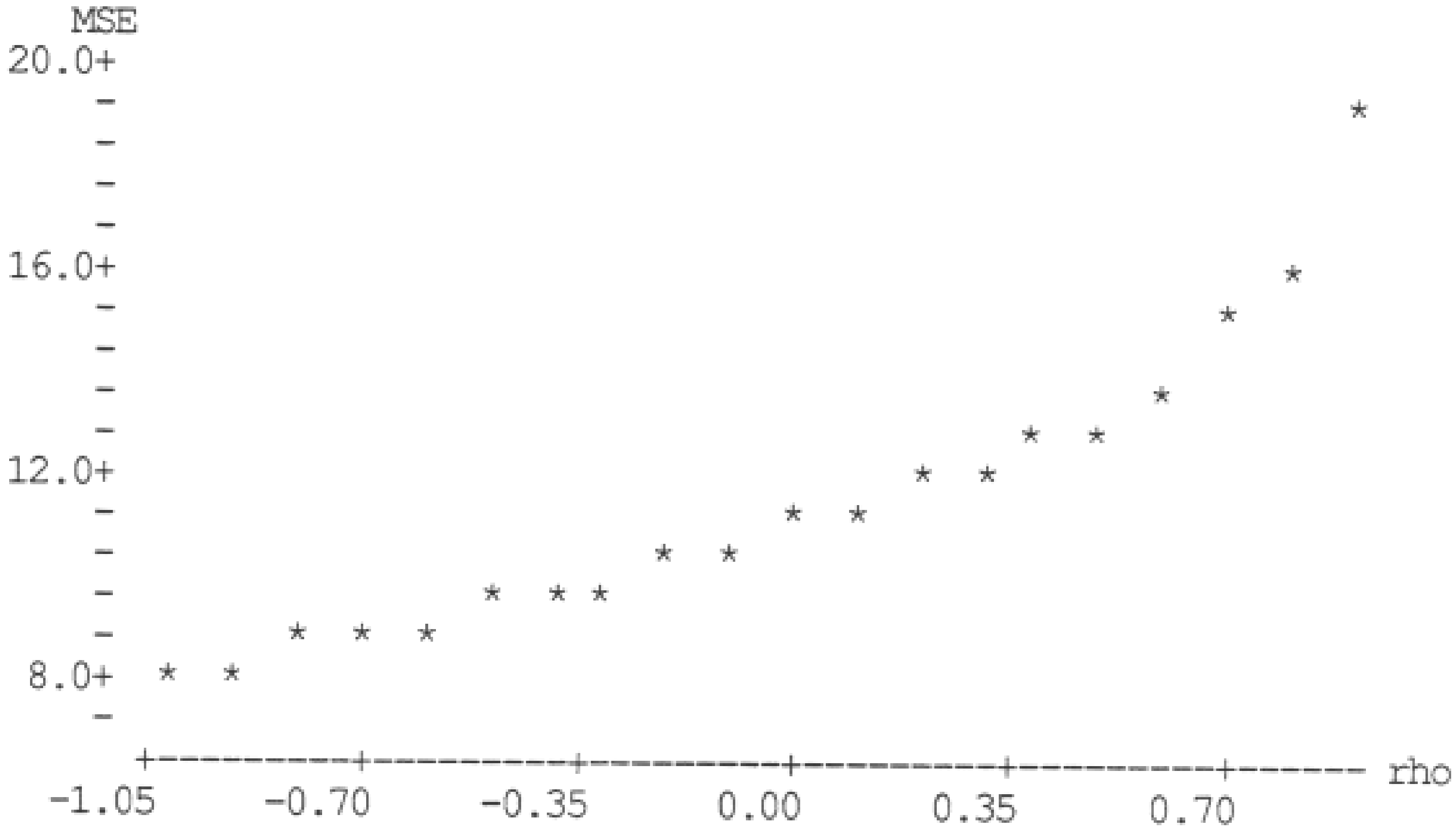
mean square error (MSE) for the discriminant function analysis problem

```
    MSE
 1.50+   *
     -
     -      *
     -
     -         *
 1.20+
     -            *
     -               *
     -
     -                  *
 0.90+                    *
     -                       *
     -                          *  *
     -                                *                      *
     -                                   *  *  *           *
 0.60+                                          *43+342
       --------+---------+---------+---------+---------+-------- rho
           -0.70     -0.35      0.00      0.35      0.70
```

mean square error (MSE) for the correlation problem



mean square error (MSE) for the linear trend surface model problem



mean square error (MSE) for the classical regression problem

```
  MSE
    -                                                         *
    -
  60+
    -                                                      *
    -                                                   *
    -
    -                                                *
  40+                                             *
    -                                        * *
    -                                     *
    -                               *  *
    -                         *  *
  20+                 *  * *
    -  *  *  *  *  *
    -
    -
      +---------+---------+---------+---------+---------+---------+------ rho
 -1.05     -0.70     -0.35      0.00      0.35      0.70
```

mean square error (MSE) for the quadratic trend surface model

```
   MSE
     -   *
     -
     -      *
     -         *
30000+            *
     -               *
     -                  * *
     -                       *
     -                          *  *
27000+                                *  *
     -                                      *
     -                                        *
     -                                           *           *
     -                                              2*
24000+                                               244++2
     -
     -
       --------+---------+---------+---------+---------+-------- rho
           -0.70     -0.35      0.00      0.35      0.70
```

mean square error (MSE) for the Upton & Fingleton data problem

ORDER FORM

Digital forms of the software, entitled "GRIFFITH'S MINITAB MACROS FOR SPATIAL STATISTICS," housed in ASCII files located on high-density, dual sided, double density 5.25" diskettes, may be purchased by contacting the following person:

Ms. Ann Hammersla
Director of Technology Transfer
Office of Sponsored Programs
Skytop Office Building
Syracuse University
Syracuse, New York 13244-5300

This digital form of the MINITAB macros is designed for PC use, but is uploadable to mainframes with few modifications, and is priced as follows:

$ 100 for commercial buyers
$ 40 for academic institutions/persons
$ 20 for students (verification of student status must accompany the order)

The reduced prices are offered exclusively to recognized educational institutions, students, and employees thereof. Syracuse University, at its sole discretion, reserves the right to refuse any order. All macros are provided "as-is," without warranty of any kind, either expressed or implied. Please allow up to four weeks for delivery.

Payment must be in U. S. dollars, drawn on a U. S. bank, and payable to Syracuse University. Prices and availability subject to change without notice. Orders must be prepaid, or accompanied by a purchase requisition.

NAME: ______________________________________

ADDRESS: ______________________________________

____

____

____

STATUS: [ ] commercial, [ ] academic, [ ] student (attach verification)

INSTITUTE OF MATHEMATICAL GEOGRAPHY (IMaGe)
2790 BRIARCLIFF
ANN ARBOR, MI 48105; U.S.A.

"Imagination is more important than knowledge"
Albert Einstein

IMaGe MONOGRAPH SERIES–1988 PRICE LIST

Exclusive of shipping; prices listed and payable in U.S. funds.

1. *Mathematical Geography and Global Art: the Mathematics of David Barr's "Four Corners Project,"* Sandra L. Arlinghaus, Director of IMaGe, and John D. Nystuen, Professor of Geography and Urban Planning, College of Architecture and Urban Planning, The University of Michigan, Ann Arbor, MI 48109. IMaGe@umichum; Nystuen@umichum. 1986. $9.95.

This monograph contains Nystuen's calculations, actually used by sculptor David Barr to position his abstract tetrahedral sculpture within the earth, as well as a Preface by Barr. Placement of the sculpture vertices in Easter Island, South Africa, Greenland, and Indonesia was chronicled in film by The Archives of American Art for The Smithsonian Institution. In addition to the archival material, this monograph also contains Arlinghaus's solutions to broader theoretical questions–was Barr's choice of a tetrahedron unique within his initial geographic constraints, and, within the set of Platonic solids?

2. *Down the Mail Tubes: the Pressured Postal Era, 1853-1984*, Sandra L. Arlinghaus, Director of IMaGe. IMaGe@umichum. 1986. $9.95.

The history of the pneumatic post, in Europe and in the United States, is examined for the lessons it might offer to the technological scenes of the late twentieth century. As Sylvia L. Thrupp, Alice Freeman Palmer Professor Emeritus of History, The University of Michigan, commented in her review of this work "Such brief comment does far less than justice to the intelligence and the stimulating quality of the author's writing, or to the breadth of her reading. The detail of her accounts of the interest of American private enterprise, in New York and other large cities on this continent, in pushing for construction of large tubes in systems to be leased to the government, brings out contrast between American and European views of how the new technology should be managed. This and many other sections of the monograph will set readers on new tracks of thought."

3. *Essays on Mathematical Geography*, Sandra L. Arlinghaus, Director of IMaGe. 1986. $15.95.

A collection of essays intended to show the range of power in applying pure mathematics to human systems. There are two types of essay: those which employ traditional mathematical proof, and those which do not. As mathematical proof may itself be regarded as art, the former style of essay might represent "traditional" art, and the latter, "surrealist" art. Essay titles are: "The well-tempered map projection," "Antipodal graphs," "Analogue clocks," "Steiner transformations," "Concavity and urban settlement patterns," "Measuring the vertical city," "Fad and permanence in human systems," "Topological exploration in geography," "A space for thought," and "Chaos in human systems–the Heine-Borel Theorem."

4. *A Historical Gazetteer of Southeast Asia*, Robert F. Austin, Director of Automated Mapping and Facility Management Systems, Baystar Service Corporation, 311 Park Place Blvd. Suite 650, Clearwater, FL 34619. 1986. $12.95.

Dr. Austin's Gazetteer draws geographic coordinates of Southeast Asian place-names together with references to these place-names as they have appeared in historical and literary documents. This book is of obvious use to historians and to historical geographers specializing in Southeast Asia. At a deeper level, it might serve as a valuable source in establishing place-name linkages which have remained previously unnoticed, in documents describing trade or other communications connections, because of variation in place-name nomenclature.

5. *Essays on Mathematical Geography–II*, Sandra L. Arlinghaus, Director of IMaGe. IMaGe@umichum. 1987. $12.95.

Written in the same format as IMaGe Monograph #3, that seeks to use "pure" mathematics in real-world settings, this volume contains the following material: "Frontispiece–the Atlantic Drainage Tree," "Getting a Handel on Water-Graphs," "Terror in Transit: A Graph Theoretic Approach to the Passive Defense of Urban Networks," "Terrae Antipodum," "Urban Inversion," "Fractals: Constructions, Speculations, and Concepts," "Solar Woks," "A Pneumatic Postal Plan: The Chambered Interchange and ZIPPR Code," "Endpiece."

6. *Theoretical Market Areas Under Euclidean Distance*, Pierre Hanjoul, Hubert Beguin, and Jean-Claude Thill: respectively, Electrical Engineer and Ph.D. candidate in Sciences, University of Louvain-la-Neuve; Professor of Economic and Quantitative Geography, University of Louvain-la-Neuve; National Fund for Scientific Research (Belgium). Address: Université Catholique de Louvain, Batîment Mercator, Place Pasteur 3, B-1348, Louvain-la-Neuve, Belgium. Beguin@bucllnl1. 1988. (English language text; abstracts written in French and in English.) $15.95.

Though already initiated by Rau in 1841, the economic theory of the shape of two-dimensional market areas has long remained concerned with a representation of transportation costs as linear in distance. In the general gravity model, to which the theory also applies, this corresponds to a decreasing exponential function of distance deterrence. Other transportation cost and distance deterrence functions also appear in the literature, however. They have not always been considered from the viewpoint of the shape of the market areas they generate, and their disparity asks the question whether other types of functions would not be worth being investigated. There is thus a need for a general theory of market areas: the present work aims at filling this gap, in the case of a duopoly competing inside the Euclidean plane endowed with Euclidean distance.

(Bien qu'ébauchée par Rau dès 1841, la théorie économique de la forme des aires de marché planaires s'est longtemps contentée de l'hypothèse de coûts de transport proportionnels à la distance. Dans le modèle gravitaire généralisé, auquel on peut étendre cette théorie, ceci correspond au choix d'une exponentielle décroissante comme fonction de dissuasion de la distance. D'autres fonctions de coût de transport ou de dissuasion de la distance apparaissent cependant dans la littérature. La forme des aires de marché qu'elles engendrent n'a pas toujours été étudiée ; par ailleurs, leur variété amène à se demander si d'autres fonctions encore ne mériteraient pas d'être examinées. Il paraît donc utile de disposer d'une théorie générale des aires de marché : ce à quoi s'attache ce travail en cas de duopole, dans le cadre du plan euclidien muni d'une distance euclidienne.)

7. *Nystuen—Dacey Nodal Analysis*, Keith J. Tinkler Editor, Professor, Department of Geography, Brock University, St. Catharine's, Ontario, Canada L2S 3A1. 1988. $15.95.

Professor Tinkler's volume displays the use of this graph theoretical tool in geography, from the original Nystuen—Dacey article, to a bibliography of uses, to original uses by Tinkler. Some reprinted material is included, but by far the larger part is of previously unpublished material. (Unless otherwise noted, all items listed below are previously unpublished.) Contents: " 'Foreward' " by Nystuen, 1988; "Preface" by Tinkler, 1988; "Statistics for Nystuen—Dacey Nodal Analysis," by Tinkler, 1979; Review of Nodal Analysis literature by Tinkler (pre–1979, reprinted with permission; post—1979, new as of 1988); FORTRAN program listing for Nodal Analysis by Tinkler; "A graph theory interpretation of nodal regions" by John D. Nystuen and Michael F. Dacey, reprinted with permission, 1961; Nystuen—Dacey data concerning telephone flows in Washington and Missouri, 1958, 1959 with comment by Nystuen, 1988; "The expected distribution of nodality in random (p, q) graphs and multigraphs," by Tinkler, 1976.

8. *The Urban Rank-size Hierarchy: A Mathematical Interpretation* by James W. Fonseca, Associate Professor of Geography and Acting Dean of the Graduate School, George Mason University, Fairfax, Virginia 22030. Jfonseca@gmuvax.bitnet. 1989. $15.95.

The urban rank-size hierarchy can be characterized as an equiangular spiral of the form $r = ae^{\theta \cot \alpha}$. An equiangular spiral can also be constructed from a Fibonacci sequence. The urban rank-size hierarchy is thus shown to mirror the properties derived from Fibonacci characteristics such as rank-additive properties. A new method of structuring the urban rank-size hierarchy is explored which essentially parallels that of the traditional rank-size hierarchy below rank 11. Above rank 11 this method may help explain the frequently noted concavity of the rank-size distribution at the upper levels. The research suggests that the simple rank-size rule with the exponent equal to 1 is not merely a special case, but rather a theoretically justified norm against which deviant cases may be measured. The spiral distribution model allows conceptualization of a new view of the urban rank-size hierarchy in which the three largest cities share functions in a Fibonacci hierarchy.

9. *An Atlas of Steiner Networks*, Sandra L. Arlinghaus, Director of IMaGe. IMaGe@umichum. 1989. $15.95.

A Steiner network is a tree of minimum total length joining a prescribed, finite, number of locations; often new locations are introduced into the prescribed set to determine the minimum tree. This Atlas explains the mathematical detail behind the Steiner construction for prescribed sets of n locations and displays the steps, visually, in a series of Figures. The proof of the Steiner construction is by mathematical induction, and enough steps in the early part of the induction are displayed completely that the reader who is well-trained in Euclidean geometry, and familiar with the concepts of graph theory and elementary number theory, should be able to replicate the constructions for full as well as for degenerate Steiner trees.

10. *Simulating $K = 3$ Christaller Central Place Structures: An Algorithm Using A Constant Elasticity of Substitution Consumption Function*, Daniel A. Griffith, Professor of Geography, Syracuse University, 343 H.B. Crouse Hall, Syracuse, NY 13244-1160. Griffith@sunrise. 1989. $15.95.

An algorithm is presented that uses BASICA or GWBASIC on IBM compatible machines. This algorithm simulates Christaller $K = 3$ central place structures, for a four–level hierarchy. It is based upon earlier published work by the author. A description of the spatial theory, mathematics, and sample output runs appears in the monograph. A digital version is available from the author, free of charge, upon request; this request must be accompanied by a 5.5–inch formatted diskette. This algorithm has been developed for use in Social Science classroom laboratory situations, and is designed to (a) cultivate a deeper understanding of central place theory, (b) allow parameters of a central place system to be altered and then graphic and tabular results attributable to these changes viewed, without experiencing the tedium of massive calculations, and (c) help promote a better comprehension of the complex role distance plays in the space–economy. The algorithm also should facilitate intensive numerical research on central place structures; it is expected that even the sample simulation results will reveal interesting insights into abstract central place theory.

The background spatial theory concerns demand and competition in the space–economy; both linear and non–linear spatial demand functions are discussed. The mathematics is concerned with (a) integration of non–linear spatial demand cones on a continuous demand surface, using a constant elasticity of substitution consumption function, (b) solving for roots of polynomials, (c) numerical approximations to integration and root extraction, and (d) multinomial discriminant function classification of commodities into central place hierarchy levels. Sample output is presented for contrived data sets, constructed from artificial and empirical information, with the wide range of all possible central place structures being generated. These examples should facilitate implementation testing. Students are able to vary single or multiple parameters of the problem, permitting a study of how certain changes manifest themselves within the context of a theoretical central place structure. Hierarchical classification criteria may be changed, demand elasticities may or may not vary and can take on a wide range of non–negative values, the uniform transport cost may be set at any positive level, assorted fixed costs and variable costs may be introduced, again within a rich range of non–negative possibilities, and the number of commodities can be altered. Directions for algorithm execution are summarized. An ASCII version of the algorithm, written directly from GWBASIC, is included in an appendix; hence, it is free of typing errors.