Linear Regression Under Multiple Changepoints

by

QiQi Lu

(Under the direction of Robert Lund and Lynne Seymour)

Abstract

This dissertation studies the least squares estimator of a trend parameter in a simple linear regression model with multiple changepoints when the changepoint times are known. The error component in the model is allowed to be autocorrelated. The least squares estimator of the trend and the variance of the trend estimator are derived. Consistency and asymptotic normality of the trend estimator are established under wide generality. The Lund *et al.* (2001) temperature trend study of the contiguous 48 United States is updated as an application.

Index words: Ordinary Least Squares, Trend Estimate, Autocorrelation, Consistency, Asymptotic Normality, Temperature Trends, Periodic Time Series, Head-Banging Algorithm

Linear Regression Under Multiple Changepoints

by

QiQi Lu

B.S., Peking University, China, 1995

M.S., China Seismological Bureau, China, 1998

M.S., The University of Georgia, 2002

A Dissertation Submitted to the Graduate Faculty
of The University of Georgia in Partial Fulfillment
of the
Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2004

Linear Regression Under Multiple Changepoints

by

QiQi Lu

Approved:

Major Professor:	Robert Lund and Lynne Seymour
Committee:	Daniel Hall Jaxk Reeves Xiangrong Yin

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
August 2004

To my parents

Acknowledgments

I am deeply indebted to Dr. Robert Lund and Dr. Lynne Seymour, my major professors, whose generous help, encouragement, and expert guidance were responsible for the completion of this dissertation. Dr. Seymour's kindness and endless patience provided me with a pleasurable studying environment. Dr. Lund's assistance extended beyond the usual requirements of the dissertation and contributed significantly to my professional development. Your brilliance never ceased to amaze me.

I am truly grateful to Dr. Daniel Hall, Dr. Jaxk Reeves, and Dr. Xiangrong Yin, my committee members, for their timely assistance and valuable suggestions.

A very special thanks goes to Dr. Robert Taylor for being an excellent instructor in the course he taught me and encouraging me to pursue my achievements in academics.

Sincere appreciation is extended to the Faculty, Staff, and students of the Department of Statistics, who helped me in many ways and made my stay at UGA both enjoyable and memorable. I also extend my special thanks to Dr. Roy Welch and Staff at the Center for Remote Sensing and Mapping Science for their financial assistance during first two years of my study and their enthusiastic inspiration.

Furthermore, I would like to thank my friends in China and the United States for their care and support. I am particularly thankful to Ms. Yuqiu Zhu for being such a good friend for the past fifteen years. You will have my friendship for life.

Above all, I express my deepest gratitude to my husband, Yangrong Ling, for his constant support, understanding and love. I also wish to express my love to my son, ”TongTong”, who has brought me a lot of happiness.

Finally, I would like to thank my parents and sisters for their love, solicitude, encouragement, and good wishes. Especially, I am highly grateful to my parents for taking good care

of my son during the last two years of my study. Without them, this dissertation would not have been possible.

Table of Contents

Page

Chapter 1

Introduction

There has always been a keen interest in statistical changepoint problems as they are encountered in diverse disciplines such as climatology, economics, finance, medicine, psychology, geology, etc. Changepoints are times where the series under study first undertakes a structural change. Such change is usually in mean structure, but can also refer to variance changes, or changes in the marginal distributions. Multiple changepoints may be present.

Changepoints are present in many climatic time series, our ultimate object of study here. For example, changepoints in temperature series are plausible whenever the recording station physically moves, the thermometer is changed, the thermometer shelter is altered, an observer changes, etc. The time of a changepoint may be explicitly noted in the station history logs or may not be documented at all. Fortunately, in many climate settings the time of the changepoint is known. Undocumented changepoint time(s) greatly complicate the analysis.

Changepoints can substantially alter conclusions made from climate series. Consider the New Bedford, MA, series plotted in Figure 1.1. This series has experienced 4 changepoints since 1812; in particular, the station was physically moved in 1906 and there were instrumentation changes in 1888, 1951, and 1985. The solid and broken lines depict least squares fitted lines to the series. A linear annual trend estimate and standard error for this series is 1.3835 ±0.3415°C per century when changepoints are ignored and -0.2817 ± 0.4979°C per century when changepoints are taken into account. Note that there is a large discrepancy in the estimates — they differ by a factor of about 5 and the directions (signs) are even different. Note also that the standard error has increased from 0.3415°C per century (ignoring

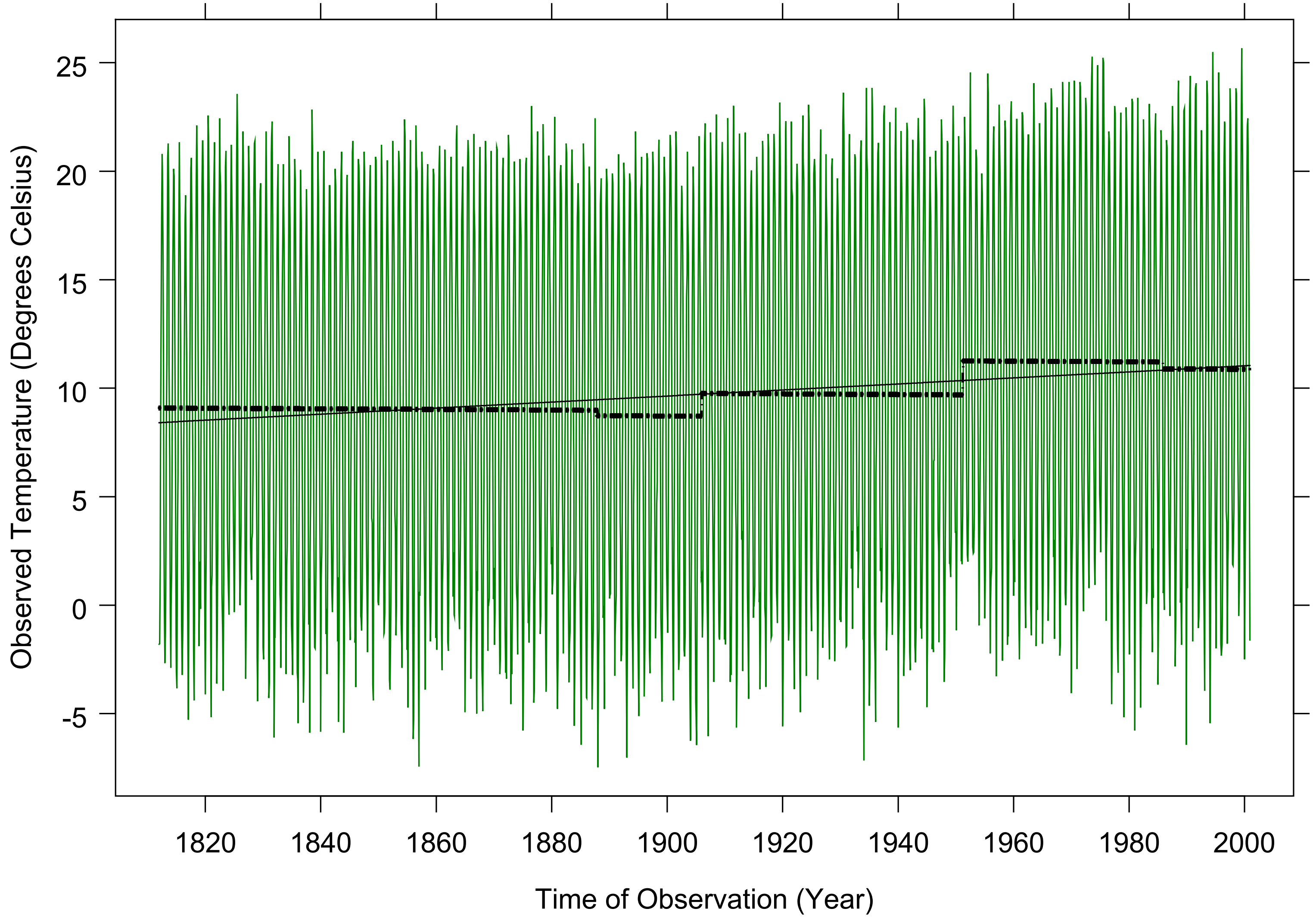

Figure 1.1: Monthly Temperatures at New Bedford, MA, with Least Squares Trends

changepoints) to 0.4979°C per century (taking changepoints into account). The estimated regression responses from the changepoint and non-changepoint models are compared graphically against the raw series in Figure 1.1. The New Bedford series is not pathological; in fact, Lund, Seymour, and Kafadar (2001) show that changepoints are the single most crucial factor in developing an accurate linear temperature trend estimator at a single station in the United States. Whereas one believes that a spatial law of large numbers holds, specifically that changepoints for series aggregated over large geographical areas average to zero, the overall importance of changepoints at an individual station is clear.

In this dissertation, the focus is on known changepoint times in the mean response of the series in the setting where many changepoints are present. Changepoints in variability, or more general structural changes in the marginal distributions of the series, are not considered.

The rest of the dissertation is organized as follows. Chapter 2 reviews historical and recent literature on changepoints. Chapter 3 studies a simple linear regression model with multiple changepoints under two different time series error structures. The large sample properties of the trend estimator are examined and its asymptotic normality is established under wide generality. Chapter 4 updates the Lund *et al.* (2001) temperature trend study of the contiguous 48 United States for data observed during the last four years and proposes some methodological improvements over past work.

1.1 References

[1] Lund, R.B., Seymour, L., and Kafadar, K. (2001). Temperature trends in the United States, *Envirometrics*, **12**, 673-679.

Chapter 2

Literature Review

2.1 Changepoint problems in statistics

Changepoint problems comprise a rapidly growing area of statistics, both in theory and applications. In an intuitive sense, a changepoint is a time point such that the observations follow one distribution up to that point and a distinct distribution thereafter. There are two fundamental changepoint problems: one is to detect if there are any changepoints in the sequence; the other is to estimate the number of changes and their corresponding locations when a changepoint is present. Frequently used methods for changepoint inference in the literature include maximum likelihood, Bayesian methods, and nonparametric tests. Different types of changepoints exist (mean, variance, etc.); however, the most often investigated changepoint problem is that of the change in the mean of a sequence of random variables.

2.1.1 Single changepoints

Many previous authors consider detecting changes in the mean of a sequence of normal random variables experiencing a single changepoint at an unknown time. This problem was first examined by Page (1955). Page employed a Shewhart control chart approach (Shewhart 1931) and constructed a sequential test based on a cumulative sum (CUSUM) scheme. Chernoff and Zacks (1964) proposed a Bayesian approach to detect a possible shift in a parameter of a distribution occurring at unknown time(s). Their results were later generalized by Kander and Zacks (1966) to the case of a one-parameter exponential family, and by Gardner (1969) to the case of an unknown positive mean shift. Sen and Srivastava (1975) obtained the precise null hypothesis distribution of Gardner's statistic and compared Bayesian tests with the maximum likelihood statistic for the single changepoint problem. The next important advance was made by Hinkley (1970, 1971a). Hinkley investigated a maximum likelihood approach for general changepoints in the mean response when the parameters of the time series are allowed to vary at the changepoint time. Hawkins (1977) and Worsley (1979) derived the null hypothesis distribution for the case of known and unknown variances of a single change in mean.

The above parametric methods apply to changepoint detection in a sequence of independent random variables. This assumption, however, falls far short of being practical for many applications. Since the 1980s, methods of changepoint detection for dependent random processes have attracted substantial attention. The work of Box and Tiao (1965) was among the first in this direction. They assumed a non-stationary integrated moving average model of observations. For detection of a shift in the mean value of such sequences, a Student's t-statistic was used. Picard (1985) estimated a shift in a Gaussian autoregressive process of a known autoregressive order. Tang and MacNeill (1993) proposed adjustments to test statistics to account for the effect of serial correlation. An influential area of changepoint detection for random sequences takes a non-parametric approach. Non-parametric methods have been proposed by Bhattacharya and Johnson (1968), Pettit (1979), and Bhattacharya and Friesson (1981).

2.1.2 Multiple changepoints

Most early changepoint works concentrate on the case of a single changepoint in the mean of a random sequence. The problem of multiple changepoints has not been as heavily trodden upon. To detect multiple changepoints, Vostrikova (1981) proposed a method, known as a binary segmentation procedure, to simultaneously estimate the number of changepoints and their locations. Vostrikova's advance was primarily an algorithm that saved computation time. Vostrikova's first step tests the null hypothesis of no changepoint versus the alternative of one changepoint. If the null hypothesis is rejected, the two subsequences before and after the changepoint found are analyzed in steps. The process is repeated until no further subsequences are found to have an additional changepoint.

Srivastava and Worsley (1986) proposed a likelihood ratio test for detecting a change in the mean of a sequence of independent normal random vectors. In Yao (1988), a version of Schwarz's criterion was used for changepoint detection. James, James and Siegmund (1992) derived an asymptotic approximation for the likelihood ratio test and confidence regions for the change in multivariate normal means.

In recent years, new techniques have been applied to multiple changepoint problems. Barry and Hartigan (1992) consider this problem in the framework of the Markov production model. Stephens (1994) discusses the use of the Gibbs sampler in multiple changepoint problems and demonstrate how it can be used to considerably reduce the computational analysis load. In Wang (1995), the problem of detection of multiple changepoints in a Gaussian random sequence with the use of wavelet transformations was considered. Chib (1998) introduced a Bayesian approach for models with multiple changepoints.

2.1.3 Changepoints in regression models

If the series indeed has a changepoint, then a one-regime regression model obviously leaves the data poorly explained. Changepoint problems in regression models were first considered by Quandt (1958, 1960), who derived a likelihood ratio test based on testing and estimating a linear regression model obeying two distinct regimes. Hinkley (1969, 1971b) studied estimation and inference in a two-phase regression model, examining the null hypothesis distribution of an F-type test statistic. Later, Lund and Reeves (2002) pointed out that Hinkley's claim of an F distribution is incorrect and leads to an overestimation of the number of undocumented changepoints. Ferreira (1975) studied a switching regression model from a Bayesian point of view with the assumption of a known number of changepoints. Brown, Durbin, and Evans (1975) introduce a method involving recursively computed residuals to test for changepoints in multiple regression models. Hawkins (1989) used a union and intersection approach to test for changes in a linear regression model. In Kim and Siegmund (1989), asymptotic properties of a maximum likelihood statistic were investigated for testing the null hypothesis of statistical homogeneity in a linear regression model against a changepoint alternative for the case when the regression function is not continuous at the time of the changepoint. Gombay and Horvath (1994) consider tests based on the maxima of weighted cumulative sums based processes to detect possible changepoints in a multiple linear regression. Andrews, Lee and Ploberger (1996) derive a class of finite-sample optimal tests for one or more changepoints at unknown times in a multiple linear regression model.

In this dissertation, the process which is studied is a simple linear regression model with multiple changepoints, where the changepoint times are known. This model allows for a mean shift at each changepoint time and simultaneous estimation of any trend and changepoint effects. The error component in the regression model is allowed to be autocorrelated, which is one of our main technical advances. In particular, we compute the least squares estimate of the trend and derive its variance in this setting. The large sample properties of the trend estimate are examined and its asymptotic normality is established under wide generality.

2.2 Changepoint problems in climatology

Changepoints (inhomogeneities) are present in many climatic time series. Climatologists have studied changepoint problems extensively (cf. Thompson 1984, Alexandersson 1986, Jones *et al.* 1986; Karl and Williams 1987, Rhoades and Salinger 1993, Easterling and Peterson 1995, Vincent 1998). The importance of using homogeneous climate series in climate research has received much attention. Analyzing raw inhomogenous climate data can seriously change the assessment of climate trends and variability.

2.2.1 Homogeneity Methods

A homogeneous time series in climatology is defined as one where variations are caused only by variations in weather and climate (Conrad and Pollak 1962). It has long been known that most of the long-term station air temperature records are not homogeneous, since they contain changes that result from nonclimatic effects such as new instruments, station relocation, and changes in averaging methods for time-averaged quantities, etc.

Inhomogeneities in climatic time series can occur either as gradual trends or as a discontinuity (sharp change). Gradual trends may occur due to urban warming or other effects that accumulate over time. However, discontinuities can be linked to many physical reasons: from station moves and instrument changes to changes in methods for calculating time-averaged values (Easterling and Peterson 1995).

In practice, it is not possible to discriminate between natural (climatic) and artificial (nonclimatic) variations in climatic time series. In other words, it is generally impossible to decide whether or not a series of observations is homogeneous. In this sense, we view homogeneity in terms of what Conrad and Pollak (1962) have defined as *relative homogeneity*: "*A climatological series is relatively homogeneous with respect to a synchronous series at another place if the temperature differences (precipitation ratios) of pairs of homogeneous averages constitute a series of random numbers that satisfies the law of errors.*" Statistical looseness aside, this definition indicates that the variations in weather have similar tendencies over rather large regions. So neighboring stations should paint the same picture of temperature change. For example, a cold winter in Athens, GA, is usually accompanied by a cold winter in Atlanta, GA.

2.2.2 Adjusting for inhomogeneities in climatological time series

To make homogeneity adjustments to a climatic series with possible changepoints, two basic approaches have been developed. One is based on the use of metadata (station history) when available, and the other is based on statistical methods to detect undocumented inhomogeneities when station history information is not available.

a. Adjustments for inhomogeneity by metadata

Our data source of monthly mean temperatures is the U.S. Historical Climatology Network (USHCN), which has reasonably good station history records (Karl *et al.* 1990). Station history information provides times of station moves, changes in instrumentation, height of instruments above the ground, etc., at each observing site. Such metadata permits adjustments for inhomogeneity.

Karl *et al.* (1986) derive a correction for the time-of-observation bias to convert daily temperature to a midnight-to-midnight scale and verify its validity from hourly data available at many U.S. stations. To apply such a bias correction, it is necessary to have reliable metadata defining all changepoint times in the station record. Karl and Williams (1987) have developed a detailed procedure for adjusting for site changes when the changepoint times are known a priori. The adjusted data retain its original scale and did not, for the most part, contain anomalies. Karl *et al.* (1988) have concluded that urban effects on temperature are detectable even in small towns (say with a population under 10,000). Also, systematic discontinuities were introduced by the change from liquid-in-glass thermometers to the maximum-minimum temperature system (MMTS) commonly used in the U.S. Cooperative Network (Quayle *et al.* 1991) today. Rhoades and Salinger (1993) propose a method for estimating the effect of known site changes on temperature and rainfall measurements. For temperature data, a site-change effect can be estimated by computing a difference between the target station and a weighted mean of neighboring stations.

b. Detection of and adjustment for inhomogeneities

Reliable monthly mean temperature records are usually very important in making useful decisions in many climatological applications. One way of checking the reliability of a climatic series is to compare a candidate station (a station which may require adjustments due to non-climate changepoints) with nearby, closely related (specifically, highly correlated) reference stations. This is the idea behind all tests of relative homogeneity. Neighboring stations can reveal discontinuities at the candidate station more readily when the correlation between the candidate and its neighbors is high and the year-to-year variances of the anomalies are small (Karl and Williams 1987).

A method for detecting changepoints without reference stations was proposed by Thompson (1984). This might accidentally remove the long-term trend from the climate records in the course of adjusting for identified changepoints, since there is no way of distinguishing changes of meteorological origin from site-change effects (Rhoades and Salinger 1993).

Jones *et al.* (1986) develop a visual technique to identify any major inhomogeneities. Alexandersson (1986) used a single reference series created from a number of neighboring stations. This helps minimize the effects of a discontinuity in one of the neighboring stations' time series. Easterling and Peterson (1995) present a method for detecting undocumented changepoints, which is based on regression models with the difference between the candidate and reference series as the response variable and time as the explanatory variable. A two-phase regression model was used to identify the position of the changepoint and significance of the changepoint was tested with an F-based statistic. Lund and Reeves (2002) showed that this F test overestimates the number of undocumented changepoints and they propose an $F_{\max}$ test, which has performed reasonably well. The method in Lund and Reeves (2002) allows for detecting both step- (discontinuity) and trend- (gradual trend) type changepoints. Vincent (1998) describes a technique to identify nonclimatic steps and trends in Canadian temperature series. This technique is based on the application of four linear regression models; a residual analysis is used to assess the fit of the model.

Easterling *et al.* (1996) note that using climatic time series with homogeneity adjustments can produce somewhat different results than using unadjusted data. Therefore, care must be taken when using either the adjusted or unadjusted time series at an individual station.

In this dissertation, under known changepoint times given by the station history records without data adjustment for inhomogeneities, our methods can simultaneously estimate both trend and changepoint effects. The methods allow for both mean shifts (step-type changepoint) at each changepoint time and overall trends effects.

2.3 Asymptotic normality of the OLS estimator

Consider the statistical regression model

$$y_t = f(x_t, \beta) + \epsilon_t, \qquad t = 1, 2, \ldots, n, \tag{2.3.1}$$

where the series $\{\epsilon_t\}$ are zero mean errors. The ordinary least squares (OLS) estimator, denoted by $\hat{\beta}$, is defined as an argument of β that minimizes the sum of squared residuals

$$S_n(\beta) = \sum_{t=1}^{n} \{y_t - f(x_t, \beta)\}^2. \tag{2.3.2}$$

For a nonlinear regression function $f(x_t, \beta)$, general conditions which ensure the asymptotic normality of the OLS estimator were first given by Jennrich (1969) and Malinvaud (1970) for the model in (2.3.1) with independent and identically distributed (IID) $\{\epsilon_t\}$. Hannan (1971) extended Jennrich's results to time series data by allowing for stationary, but ergodic $\{\epsilon_t\}$. Robinson (1972) generalized Hannan's results to systems of equations. The essential details of the asymptotic theory are summarized briefly by Amemiya (1983), who considers both IID and autocorrelated errors. White (1984) provided more general conditions ensuring the asymptotic normality of the OLS estimator when $\{\epsilon_t\}$ is serially correlated and/or heteroscedastic.

When $f(x_t, \beta)$ is linear in β, Eicker (1963) develops general conditions for the asymptotic normality of the OLS estimators in (2.3.1) under IID $\{\epsilon_t\}$. There is also a body of literature on the central limit theorem for stationary process. Conditions guaranteeing that least squares regression estimators are asymptotically normal can be extracted from Grenander and Rosenblatt (1957). Hannan (1961) proved a central limit theorem for least squares regression parameter estimators in a multiple linear regression model when the errors comprise a causal linear stationary time series. A very general discussion of this situation is given in

Eicker (1967). Hannan (1970) used Eicker's ideas to modify the results in Hannan (1961). Anderson (1971) extended Hannan's (1961) results to settings where the errors in the causal linear representation satisfy a Lindeberg-type condition. Unfortunately, these works are all phrased in frequency domain terminology and are difficult to interpret in the time domain, the setting most frequently used in modern analyses. Indeed, Fuller's (1996) main contribution was to translate various classical theorems into the time domain.

In the Chapter 3, consistency and asymptotic normality of the OLS estimator in a simple linear regression model under multiple changepoints with short-memory stationary autocorrelated errors is established and its limiting properties are quantified.

2.4 References

[1] Alexandersson, H. (1986). A homogeneity test applied to precipitation data, *Journal of Climatology*, **6**, 661-675.

[2] Amemiya, T. (1983). Non-linear regression models. In Z. Griliches and M.D. Intriligator (Eds.), *Handbook of Econometrics*, Vol. **I**, 333–389, North-Holland: Amsterdam.

[3] Anderson, T.W. (1971). *The Statistical Analysis of Time Series*, John Wiley, New York.

[4] Andrews, D., Lee, I., and Ploberger, W. (1996). Optimal changepoint tests for normal linear regression, *Journal of Econometrics*, **70**, 9–38.

[5] Barry, D. and Hartigan, J.A. (1992). Product partition models for change-point problems, *Annals of Statistics*, **20**, 260– 279.

[6] Bhattacharya, G.K. and Friesson, D. (1981). A nonparametric control chart for detecting small disorders, *Annals of Statistics*, **9**, 544-554.

[7] Bhattacharya, G.K. and Johnson, R.A. (1968). Non-parametric tests for shift at an unknown time point, *Annals of Mathematical Statistics*, **39**, 1731-1743.

[8] Box G.E.P. and Tiao G.C. (1965). A change in level of a nonstationary time series, *Biometrika*, **52**, 181-192.

[9] Brodsky, B.E. and Darkhovsky, B.S. (1993). *Nonparametric Methods in Change-Point Problems*, Kluwer, Dordrecht.

[10] Brodsky, B.E. and Darkhovsky, B.S. (2000). *Non-Parametric Statistical Diagnosis: Problems and Methods*, Kluwer, Dordrecht.

[11] Brown, R.L., Durbin, J., and Evans, J.M. (1975). Techniques for testing the constancy of regression relationships over time (with discussion), *Journal of the Royal Statistical Society, Series B,* **37**, 149-192.

[12] Chen, J. and Gupta, A.K. (2000). *Parametric Statistical Change Point Analysis*, Birkhauser.

[13] Chernoff, H. and Zacks, S. (1964). Estimating the current mean of a normal distribution which is subjected to changes in time, *Annals of Mathematical Statistics*, **35**, 999-1018.

[14] Chib, S. (1998). Estimation and comparison of multiple change-point models, *Journal of Econometrics*, **86**, 221-241.

[15] Conrad, V. and Pollak, C. (1950). *Methods in Climatology*, Harvard University Press, Cambridge, MA, 223pp and 226pp.

[16] Csorgo, M. and Horvath, L. (1997). *Limit Theorems in Change-Point Analysis*, Wiley, New York.

[17] Easterling, D.R. and Peterson, T. (1995). A new method for detecting undocumented discontinuities in climatological time series, *International Journal of Climatology*, **15**, 369-377.

[18] Easterling, D.R., Peterson, T., and Karl, T.R. (1996). On the development and use of homogenized climate datasets, *Journal of Climate*, **9**, 1429-1434.

[19] Eicker, F. (1963). Asymptotic normality and consistency of the least squares estimators for families of linear regressions, *Annals of Mathematical Statistics*, **34**, 447–456.

[20] Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors, In: *Proceeding of The Fifth Berkeley Symposium: Mathematical Statistics and Probability (Berkeley California 1965/1966)*, **1** Statistics, 59–82, University of California Press, Berkeley.

[21] Ferreira, P.E. (1975). A Bayesian analysis of a switching regression model: Known number of regimes, *Journal of the American Statistical Association*, **70**, 370-374.

[22] Fuller, W. (1996). *Introduction to Statistical Time Series*, Second Edition, John Wiley, New York.

[23] Gardner, L.A. (1969). On detecting change in the mean of normal variates, *Annals of Mathematical Statistics*, **40**, 116-126.

[24] Gombay, E. and Horvath, L. (1994). Limit theorems for change in linear regression, *Journal of Multivariate Analysis*, **48**, 43-69.

[25] Grenander, U. and Rosenblatt M. (1957). *Statistical Analysis of Stationary Time Series*, John Wiley, New York.

[26] Hannan, E.J. (1961). A central limit theorem for systems of regressions, *Proceedings of the Cambridge Philosophical Society*, **57**, 583–588.

[27] Hannan, E.J. (1970). *Multiple Time Series*, John Wiley, New York.

[28] Hannan, E.J. (1971). Non-linear time series regression, *Journal of Applied Probability*, **8**, 767–780.

[29] Hawkins, D.M. (1977). Testing a sequence of observations for a shift in location, *Journal of the American Statistical Association*, **72**, 180-186.

[30] Hawkins, D.M. (1989). A U-I approach to retrospective testing for shift parameters in a linear model, *Communications in Statistics: Theory and Methods*, **18**, 3117-3134.

[31] Hinkley, D.V. (1969). Inference about the intersection in two-phase regression, *Biometrika*, **56**, 495-504.

[32] Hinkley, D.V. (1970). Inference about the change-point in a sequence of random variables, *Biometrika*, **57**, 1-17.

[33] Hinkley, D.V. (1971a). Inference about the change-point from cumulative sum tests, *Biometrika*, **58**, 509-523.

[34] Hinkley, D.V. (1971b). Inference in two-phase regression, *Journal of the American Statistical Association*, **66**, 736-743.

[35] James, B.J., James, K.L., and Siegmund, D. (1992). Asymptotic approximations for likelihood ratio tests and confidence regions for a change-point in the mean of a multivariate normal distribution, *Statistica Sinica*, **2**, 69-90.

[36] Jennrich, R.I. (1969). Asymptotic properties of non-linear least squares estimation, *Annals of Mathematical Statistics*, **40**, 633-643.

[37] Jones, P.D., Paper, S.C.B., Bradley, R.S., Diaz, H.F., Kelly, P.M., and Wigley, T.M.L. (1986). Northern Hemisphere Surface Air Temperature Variations: 1851-1984, *Journal of Climate and Applied Meteorology*, **25**, 161-179.

[38] Kander, Z. and Zacks, S. (1966). Test procedures for possible changes in parameters of statistical distributions occurring at unknown time point, *Annals of Mathematical Statistics*, **37**, 1196-1210.

[39] Kim, H.J. and Siegmund, D. (1989). The likelihood ratio test for a change-point in simple linear regression, *Biometrika*, **76**, 409-423.

[40] Karl, T.R. and Williams, C.N., Jr. (1987). An approach to adjusting climatological time series for discontinuous inhomogeneities, *Journal of Climate and Applied Meteorology*, **26**, 1744-1763.

[41] Karl, T.R., Williams, C.N., Young, P.J., and Wendland, W.M. (1986). A model to estimate the time of observation bias associated with monthly mean maximum, minimum and mean temperatures for the United States, *Journal of Climate and Applied Meteorology*, **25**, 145-160.

[42] Karl, T.R., Diaz, H.F., and Kukla, G. (1988). Urbanization: Its detection and effect in the United States climate record, *Journal of Climatology*, **1**, 1099-1123.

[43] Karl, T.R., Williams, C.N., Quinlan, F.T., and Boden, T.A. (1990). In United States Historical Climatology Network (USHCN) Serial Temperature and Prcipitation Data, Environmental Sciences Division, Publication No. 3404, Carbon Dioxide Information and Analysis Center, Oak Ridge National Laboratory, Oak Ridge, TN, 389pp.

[44] Lund, R.B. and Reeves, J. (2002). Detection of undocumented changepoints: A revision of the two-phase regression model, *Journal of Climate*, **15**, 2547-2554.

[45] Malinvaud, E. (1970). The consistency of nonlinear regressions, *Annals of Mathematical Statistics*, **41**, 956-969.

[46] Page, E.S. (1955). A test for change in a parameter occurring at an unknown point, *Biometrika*, **42**, 523-526.

[47] Picard, D. (1985). Testing and estimating change-points in time series, *Journal of Applied Probability*, **14**, 411-415.

[48] Pettit, A.N. (1979). A nonparametric approach to the change point problem, *Applied Statistics*, **28**, 126-135.

[49] Quandt, R.E. (1958). The estimation of the parameters of a linear regression system obeys two separate regimes, *Journal of the American Statistical Association*, **53**, 873-880.

[50] Quandt, R.E. (1960). Test of the hypothesis that a linear regression system obeys two separate regimes, *Journal of the American Statistical Association*, **55**, 324-330.

[51] Quayle, R.G., Easterling, D.R., Karl, T.R., and Hughes, P.Y. (1991). Effects of recent thermometer changes in the cooperative station network, *Bulletin of the American Meteorological Society*, **72**, 1718-1724.

[52] Rhoades, D.A. and Salinger, M.J. (1993). Adjustment of temperature and rainfall records for site changes, *International Journal of Climatology*, **13**, 899-913.

[53] Robinson, P.M. (1972). Non-linear regression for multiple time-series, *Journal of Applied Probability*, **9**, 758-768.

[54] Sen, A.K. and Srivastava, M.S. (1975). On tests for detecting change in mean, *Annals of Statistics*, **3**, 980-108.

[55] Shewhart, W.A. (1931). *Economic Control of Quality of Manufactured Products*, D. Van Nostrand Company, Inc., New York.

[56] Sinha, B., Rukhin, A., and Ahsanullah, M. (1995). *Applied Change Point Problems in Statistics*, NOVA Seience Publishers, Inc., New York.

[57] Srivastava, M.S. and Worsley, K.J. (1986). Likelihood ratio tests for a change in the multivariate normal mean, *Journal of the American Statistical Association*, **81**, 199-204.

[58] Stephens, D.A. (1994). Bayesian retrospective multiple-changepoint identification, *Applied Statistics*, **43**, 159-178.

[59] Tang, S.M. and MacNeill, I.B. (1993). The effect of serial correlation on tests for parameter change at unknown time, *Annals of Statistics*, **21**, 552-557.

[60] Thompson, C.S. (1984). Homogeneity analysis of rainfall series: An application of the use of a realistic rainfall model, *Journal of Climatology*, **4**, 609-619.

[61] Vincent, L.A. (1998). A technique for the identification of inhomogeneities in Canadian temperature series, *Journal of Climate*, **11**, 1094-1104.

[62] Vostrikova, L.Ju. (1981). Detecting ”disorder” in multidimensional random processes, *Soviet Mathematics Doklady*, **24**, 55-59.

[63] Wang, Y. (1995). Jump and sharp cusp detection by wavelets, *Biometrika*, **82**, 385-397.

[64] White, H. and Domowitz, I. (1984). Nonlinear regression with dependent observations, *Econometrika*, **52**, 143-161.

[65] Worsley, K.J. (1979). On the likelihood ration test for a shift in location of normal populations, *Journal of the American Statistical Association*, **74**, 365-367.

[66] Yao, Y.C. (1988). Estimating the number of change-points via Schwarz's criterion, *Statistics and Probability Letters*, **6**, 181-189.

Chapter 3

Simple Linear Regression with Multiple Changepoints[1]

[1] Lu, Q. and Lund, R. Submitted to *Biometrika,* 7/15/2004.

Abstract

This paper studies the ordinary least squares trend estimator in a simple linear regression model with multiple changepoints when the changepoint times are known. The error component in the model is allowed to be a general short-memory stationary autocorrelated series. Consistency and asymptotic normality of the estimator is established and its limiting properties are quantified.

Key Words: Linear Trend; Time Series; Consistency; Asymptotic Normality.

3.1 Introduction

Consider a simple linear regression model with multiple changepoints:

$$X_t = \mu + \alpha t + \delta_t + \epsilon_t, \qquad t = 1, 2, \ldots, n, \tag{3.1.1}$$

when the changepoint times are known. The errors $\{\epsilon_t\}$ are zero mean and stationary in time t with autocovariance $\gamma(h) = \text{cov}(\epsilon_t, \epsilon_{t+h})$ at lag h. The regression location parameter is μ, α is the linear trend-slope parameter, which is our focus, and $\{\delta_t\}$ is a changepoint mean shift factor. In particular, the changepoints have the structure

$$\delta_t = \begin{cases} \Delta_1, & 1 \leq t < \tau_1 \\ \Delta_2, & \tau_1 \leq t < \tau_2 \\ \vdots & \vdots \\ \Delta_k, & \tau_{k-1} \leq t \leq n \end{cases}, \tag{3.1.2}$$

where $\tau_1 < \tau_2 < \ldots < \tau_{k-1}$ are the ordered known changepoint times. Here, n is the total number of observations recorded and $k = k(n)$ is the total number of regimes in the series up to time n. For parameter identifiability, we take $\Delta_1 = 0$ (else the $\Delta_i's$ and μ become confounded). Hence, the regression model in (3.1.1) contains the $k+1$ unknown parameters μ, α, $\Delta_2, \ldots, \Delta_k$.

There is much literature on regression models with changepoints similar to (3.1.1) (cf. Quandt 1958 and 1960; Hinkley 1969 and 1971b; Solow 1987; Easterling and Peterson 1995; Vincent 1998; Lund and Reeves 2002). Many of these models can be viewed as classic simple linear regressions that allow for k phases with *unknown* changepoint times. Developing a good changepoint detection method is the main goal in these works. We also note that the model in (3.1.1) can be viewed as an ANCOVA model (Analysis of Covariates) in the presence of autocorrelated errors and a linear trend covariate.

It is the purpose of this paper to derive the ordinary least squares (OLS) estimator of the trend parameter, discuss its consistency, and establish its asymptotic normality under wide generality. The asymptotic normality is not immediately clear from prior literature as we allow the number of changepoints to tend to infinity as the sample size tends to infinity. The nuances of this aspect will become clearer as we progress. This chapter concludes with an extension of the results to the periodic (multivariate) setting with multiple changepoints in the presence of periodically stationary time series errors.

3.2 OLS Trend Estimates and Consistency

Suppose that the series experienced k different regimes during the data record. Let n_i denote the number of observations recorded during regime i, $1 \leq i \leq k$. Here, $n = n_1 + n_2 + \ldots + n_k$. The set of all time indices for regime i, denoted by Y_i, is

$$Y_i = \left\{ \sum_{j=1}^{i-1} n_j + 1, \sum_{j=1}^{i-1} n_j + 2, \ldots, \sum_{j=1}^{i} n_j \right\}.$$

For the model in (3.1.1), the OLS estimator of α has the explicit form

$$\hat{\alpha}_{\mathrm{OLS}} = \frac{\sum_{i=1}^{k} \sum_{t \in Y_i} (t - \bar{t}_i)(X_t - \bar{X}_i)}{\sum_{i=1}^{k} \sum_{t \in Y_i} (t - \bar{t}_i)^2}, \tag{3.2.1}$$

where $\bar{t}_i = n_i^{-1} \sum_{t \in Y_i} t$ and $\bar{X}_i = n_i^{-1} \sum_{t \in Y_i} X_t$ are the time and observation averages during regime i, respectively. Of course, $\hat{\alpha}_{\mathrm{OLS}}$ can be viewed as an argument of α that minimizes the sum of squares

$$S_{\mathrm{OLS}}(\mu, \alpha, \Delta_2, \ldots, \Delta_k) = \sum_{i=1}^{k} \sum_{t \in Y_i} (X_t - \mu - \alpha t - \Delta_i)^2$$

over μ, α, $\Delta_2, \ldots, \Delta_k$. The denominator of (3.2.1) can be explicitly evaluated as

$$\sum_{i=1}^{k} \sum_{t \in Y_i} (t - \bar{t}_i)^2 = \frac{\sum_{i=1}^{k} {n_i}^3 - n}{12}. \tag{3.2.2}$$

The variance of the ordinary least squares trend estimator can be easily obtained from (3.2.1). As $\hat{\alpha}_{\mathrm{OLS}}$ is a linear combination of $X_1, X_2, \ldots, X_n$, $\mathrm{Var}(\hat{\alpha}_{\mathrm{OLS}})$ is seen to be

$$\mathrm{Var}(\hat{\alpha}_{\mathrm{OLS}}) = \frac{\gamma(0) + 2 \sum_{h=1}^{n-1} w_{n,h} \gamma(h)}{\sum_{i=1}^{k} \sum_{t \in Y_i} (t - \bar{t}_i)^2}, \tag{3.2.3}$$

where the weights $\{w_{n,h}\}$ are

$$w_{n,h} = \frac{\sum_{t=1}^{n-h} \eta_t \eta_{t+h}}{\sum_{t=1}^{n} \eta_t^2}, \qquad 0 \leq h \leq n-1, \tag{3.2.4}$$

and $\{\eta_t\}$ is

$$\eta_t = \begin{cases} t - \bar{t}_1, & 1 \leq t < \tau_1 \\ t - \bar{t}_i, & \tau_{i-1} \leq t < \tau_i \\ t - \bar{t}_k, & \tau_{k-1} \leq t \leq n \end{cases}. \tag{3.2.5}$$

Note that $\hat{\alpha}_{\mathrm{OLS}}$ is unbiased for any mean zero $\{\epsilon_t\}$. Theorem 3.2.1 below establishes that

$\text{Var}(\hat{\alpha}_{\text{OLS}}) \to 0$ as $n \to \infty$ for almost all short-memory time series $\{\epsilon_t\}$ and regime lengths $\{n_i\}_{i=1}^{\infty}$. Thus, we can conclude that $\hat{\alpha}_{\text{OLS}}$ is consistent in considerable generality.

Theorem 3.2.1 Suppose that $\{\epsilon_t\}$ is stationary and has short-memory in the sense that

$$\sum_{h=0}^{\infty} |\gamma(h)| < \infty.$$

Let $m_1(n)$ be the number of regimes in the series up to time n that are of length 1. If

$$\liminf_{k\to\infty} \frac{m_1(n)}{k} < 1, \tag{3.2.6}$$

then $\text{Var}(\hat{\alpha}_{\text{OLS}}) \to 0$ as $n \to \infty$.

The short-memory condition $\sum_{h=0}^{\infty} |\gamma(h)| < \infty$ is satisfied by all causal autoregressive moving-average (ARMA) processes (cf. Brockwell and Davis 1991). The regime length condition in (3.2.6) is very general and requires that not all segments in the series have unit length. In other words, there must be an infinite number of regimes that are of length two or more to obtain consistency. Intuitively, regimes of length two or more provide some information for α, and an infinite number of regimes with some information induces consistency. For comparison's sake against a practical setting, climate time series saw six changepoints on the average during the 88-year period 1873–1950 (Mitchell 1953). In almost every practical modelling situation, the conditions in Theorem 3.2.1. hold; indeed, the result is quite general.

The proof of Theorem 3.2.1 is given in the appendix at the end of this chapter. Here we present a lemma useful in proof. In later work, we will regard $\{n_i\}_{i=1}^{\infty}$ as random; hence all limits should be interpreted almost surely.

Lemma 1 If

$$\liminf_{k\to\infty} \frac{m_1(n)}{k} < 1,$$

then

$$\begin{aligned} \lim_{n\to\infty} \sum_{i=1}^{k} \sum_{t\in Y_i} (t - \bar{t}_i)^2 &= \lim_{n\to\infty} \frac{\sum_{i=1}^{k} {n_i}^3 - n}{12} \\ &= \infty. \end{aligned} \tag{3.2.7}$$

From Lemma 1 and (3.2.2) and (3.2.3), we see that consistency of $\hat{\alpha}_{\mathrm{OLS}}$ under short-memory errors take place whenever the 'changepoint design' $\{n_i\}_{i=1}^{\infty}$ satisfies

$$\lim_{n\to\infty} \sum_{i=1}^{k} \sum_{t\in Y_i} (t-\bar{t}_i)^2 = \infty.$$

As $n \to \infty$, either the number of regimes $k = k(n) \to \infty$ or $n_i = \infty$ for some fixed i. Of course, if $n_i = \infty$ for some i, one can construct an estimate of α from data during this regime only and consistency is implicit. It is important to note that an infinite number of changepoint mean shift parameters are allowed as $n \to \infty$. In most practical situations, having $n_i = \infty$ for some fixed i is unlikely. Indeed, we are primarily interested in the case where $k \to \infty$ as $n \to \infty$.

3.3 Asymptotic Normality

There is substantial previous literature on the central limit theorem for stationary process. We mention Grenander and Rosenblatt (1957), Hannan (1961, 1970), Eicker (1967), Anderson (1971) and Fuller (1996) for classic results in systems of regressions. Two types of regressions with stationary errors have been classically considered. The first is the causal linear process

$$\epsilon_t = \sum_{j=0}^{\infty} \psi_j Z_{t-j}, \qquad \sum_{j=0}^{\infty} |\psi_j| < \infty, \tag{3.3.1}$$

where the $\{Z_t\}$ are independent and identically distributed with zero mean and variance $\sigma^2 > 0$. The other type of $\{\epsilon_t\}$ considered is stationary and satisfies a uniform mixing condition. Mixing conditions are notoriously difficult to check in practice. As $\{\epsilon_t\}$ in (3.3.1) is strictly stationary in t, mixing conditions for causal linear processes are much easier to verify, especially with Gaussian $\{Z_t\}$.

Conditions guaranteeing that least squares regression estimates are asymptotically normal can be extracted from Grenander and Rosenblatt (1957). Hannan (1961) proved a central limit theorem for least squares regression parameter estimators in a multiple linear regression model when the errors comprise a causal linear stationary time series. A very general discussion of this situation is given in Eicker (1967). Hannan (1970) used Eicker's ideas to modify the results in Hannan (1961). Anderson (1971) extended Hannan's (1961) results to settings where the errors $\{Z_t\}$ in (3.3.1) satisfy a Lindeberg-type condition. Unfortunately, these works are all phrased in frequency domain terminology and are difficult to interpret in the time domain, the setting most frequently used in modern analyses. Indeed, Fuller's (1996) main contribution was to translate various classical theorems into the time domain.

Most regression central limit theorems are based on Grenander's (1954) classical conditions for the regression variable φ_t in the multiple linear regression model

$$X_t = \sum_{i=1}^{p} \varphi_{t,i}\beta_i + \epsilon_t.$$

Grenander's three conditions, labelled as C1, C2, and C3, respectively, are

$$\text{(C1)} \quad \lim_{n\to\infty} \sum_{t=1}^{n} \varphi_{t,i}^2 = \infty, \quad \text{for} \ \ i = 1, \ldots, p;$$

$$\text{(C2)} \quad \lim_{n\to\infty} \frac{\varphi_{n,i}^2}{\sum_{t=1}^{n} \varphi_{t,i}^2} = 0, \quad \text{for} \ \ i = 1, \ldots, p; \qquad \text{and}$$

$$\text{(C3)} \quad \lim_{n\to\infty} \frac{\sum_{t=1}^{n-h} \varphi_{t,i}\varphi_{t+h,j}}{\sqrt{\sum_{t=1}^{n} \varphi_{t,i}^2 \sum_{t=1}^{n} \varphi_{t+h,j}^2}} \ \text{exists}, \qquad \text{for all} \ \ i, j = 1, \ldots, p \ \ \text{and} \ \ h \geq 0.$$

Condition C1 ensures that $\varphi_{t,i}$ grows unboundedly so that the regression equation component involving β_i is non-ignorable. Condition C2 precludes φ_n^2 from comprising an appreciable part of the sum of squares for large n and is perhaps reminiscent of Lindeberg conditions. Condition C3 stipulates that correlations between regression design columns for all sufficiently large n are well-defined. This can be interpreted in a law of large numbers sense: the sample correlations in the columns of the design matrix merely exist, implying time-constant dynamics in the limit.

Via the existence of limits in (C3), we set

$$r_h(i,j) = \lim_{n\to\infty} \frac{\sum_{t=1}^{n-h} \varphi_{t,i}\varphi_{t+h,j}}{\sqrt{\sum_{t=1}^{n} \varphi_{t,i}^2 \sum_{t=1}^{n} \varphi_{t+h,j}^2}}$$

and $R_h = [r_h(i,j)]_{i,j=1}^{p}$. We assume that R_0 is nonsingular for convenience. Under some very general assumptions on the distribution of the errors $\{\epsilon_t\}$, Grenander's conditions C1-C3 provide a wide arsenal for proving asymptotic normality of the least squares estimators of regression coefficients.

Returning to our setting of application, the model in (3.1.1) can be rewritten as

$$X_t = \mu + \alpha t + \sum_{i=2}^{k} \Delta_i \varphi_{t,i} + \epsilon_t, \tag{3.3.2}$$

where for $i = 2, \ldots, k$,

$$\varphi_{t,i} = \begin{cases} 1, & t \in Y_i \\ 0, & \text{otherwise} \end{cases}. \tag{3.3.3}$$

Now the columns in the regression design matrix corresponding to μ and α satisfy Grenander's conditions C1–C3; however the columns for the changepoint $\varphi_{t,i}$'s do not without stronger assumptions on the n_i's. Hence, one cannot use Grenander's results directly and some further analysis is needed. However, as $\hat{\alpha}_{\mathrm{OLS}}$ can be explicitly written as a function of $X_1, \ldots, X_n$ as seen in (3.2.1), there is hope. Indeed, the following result establishes asymptotic normality of $\hat{\alpha}_{\mathrm{OLS}}$ in considerable generality.

Theorem 3.3.1 Suppose that $\{\epsilon_t\}$ has the causal linear representation in (3.3.1) and that the regime lengths $\{n_i\}_{i=1}^{\infty}$ are realizations of an ergodic Markov chain $\{N_i\}_{i=0}^{\infty}$ on the state-space $\{1, 2, 3, \ldots\}$ with stationary measure $\vec{\pi} = \{\pi_\ell\}_{\ell=1}^{\infty}$. We assume that $\{N_i\}_{i=0}^{\infty}$ is independent of $\{X_t\}_{t=1}^{\infty}$ and has transition probability matrix $P = \{p_{ij}\}_{i,j=1}^{\infty}$. We take $N_0 \overset{D}{=} \vec{\pi}$ for convenience so as to render $\{N_i\}$ stationary. If $\vec{\pi}$ has a finite third moment in the sense that

$$\sum_{\ell=1}^{\infty} \ell^3 \pi_\ell < \infty, \tag{3.3.4}$$

then

$$\left(\sum_{i=1}^{k}\sum_{t\in Y_i}(t-\bar{t}_i)^2\right)^{\frac{1}{2}}(\hat{\alpha}_{\mathrm{OLS}}-\alpha)\xrightarrow{D}N(0,V) \tag{3.3.5}$$

as $n\to\infty$, where $V=\sum_{h=-\infty}^{\infty}w_h^*\gamma(h)$. Here, $w_h^*=\lim_{n\to\infty}w_{n,h}$, where $w_{n,h}$ is as in (3.2.4) and the limit exists almost surely in the realization of $\{N_i\}$.

It is worth noting that every causal ARMA series satisfies the short-memory assumption in Theorem 3.3.1. In fact, $\sum_{h=0}^{\infty}|\gamma(h)|<\infty$ is implied by (3.3.1). The 'Markov' sampling conditions in Theorem 3.3.1 imply that in the long run, a regime length ℓ has probability π_ℓ:

$$\lim_{i\to\infty}P[N_i=\ell]=\pi_\ell.$$

The randomness of the regime lengths will ensure that various limits exist in the ensuing analysis. It seems physically reasonable. Without such randomness, the analysis is very difficult. The finite third moment of $\vec{\pi}$ is a technical condition ensuring, amongst other things, that the following two limits exist:

$$\lim_{n\to\infty}\frac{k(n)}{n}=\frac{1}{\sum_{\ell=1}^{\infty}\ell\pi_\ell},$$

and

$$\lim_{n\to\infty}\frac{\sum_{i=1}^{k}\sum_{t\in Y_i}(t-\bar{t}_i)^2}{n}=\left(\frac{1}{\sum_{\ell=1}^{\infty}\ell\pi_\ell}\right)\left(\sum_{\ell=1}^{\infty}\frac{\ell(\ell+1)(\ell-1)}{12}\pi_\ell\right).$$

Finite third moment conditions arise frequently in renewal theory and regenerative process settings (cf. Ross 1996; Kalashnikov 1994). In view of the cubic structure in (3.2.2), this is not unexpected. The proof of Theorem 3.3.1 is rather technical and is given in the appendix at the end of this chapter. As a matter of notation in cases where the $\{N_i\}_{i=0}^{\infty}$ are random, and observed we abbreviate $\mathrm{Var}(\hat{\alpha}_{\mathrm{OLS}}|N_1=n_1,N_2=n_2,\ldots)$ to $\mathrm{Var}(\hat{\alpha}_{\mathrm{OLS}})$. We hope this causes no confusion as the limiting properties do not depend on the realization of $\{N_i\}_{i=0}^{\infty}$. However, all computations are conditional on the realization of $\{N_i\}_{i=0}^{\infty}$ and the analysis is properly phrased in terms of $\mathrm{Var}(\hat{\alpha}_{\mathrm{OLS}}|N_1=n_1,N_2=n_2,\ldots)$.

A large sample confidence interval for the trend estimator can be extracted from the above result. In particular,

$$\hat{\alpha}_{\text{OLS}} \pm z_{\alpha/2}\sqrt{\frac{V}{\sum_{i=1}^{k}\sum_{t\in Y_i}(t-\bar{t}_i)^2}} \tag{3.3.6}$$

is an approximate $(1-\alpha)\times 100\%$ confidence interval for the trend parameter α. Here $z_{\alpha/2}$ denotes the customary $(1-\alpha/2)$th quantile of the standard normal distribution. In the more practical setting where the autocovariances in $\{\epsilon_t\}$ are unknown, one can substitute estimates of $\gamma(\cdot)$ into (3.2.3) and use the interval

$$\hat{\alpha}_{\text{OLS}} \pm z_{\alpha/2}\widehat{\text{Var}}(\hat{\alpha}_{\text{OLS}})^{1/2}, \tag{3.3.7}$$

where $\widehat{\text{Var}}(\hat{\alpha}_{\text{OLS}})$ is an estimate of $\text{Var}(\hat{\alpha}_{\text{OLS}})$. Here a t-based margin may be more appropriate than the $z_{\alpha/2}$ quartile, but we will not pursue this slant here, preferring to quote Slustky's theorem and keep the analysis confined to asymptotics.

3.4 Periodic Simple Linear Regression Models with Multiple Changepoints

The previous section assumed that $\{\epsilon_t\}$ has zero mean and is stationary with short-memory autocovariance; in this section, the results are extended to periodically stationary time series with short-memory. A random sequence $\{Y_t\}$ with finite second moments is called a periodic series with period T if

$$\text{E}[Y_{t+T}] = \text{E}[Y_t]$$

and

$$\text{Cov}(Y_{t+T}, Y_{s+T}) = \text{Cov}(Y_t, Y_s)$$

for all integers t and s. Periodic series are also called periodically stationary, cyclostationary, or periodically correlated. The period T is taken as known and as the minimal integer satisfying the above periodicity equations (to avoid ambiguity).

One periodic version of the model in (3.1.1) (not the only) is

$$X_{mT+\nu} = \mu_\nu + \alpha_\nu(mT+\nu) + \delta_\nu^{(i)} + \epsilon_{mT+\nu}, \tag{3.4.1}$$

where ν is a season index satisfying $1 \le \nu \le T$; m is a cycle index satisfying $0 \le m \le d-1$; α_ν is the linear trend during season ν; $\{\delta_\nu^{(i)} = \Delta_{\nu,i}\}$ is a changepoint mean shift factor for season ν under regime i; and $\{\epsilon_{mT+\nu}\}$ is a zero mean periodic series with period T. The bookkeeping tracks $X_{mT+\nu}$ as the observation from season ν of cycle m and $d = \lfloor n/T \rfloor$ as the total observed number of cycles. To avoid trite work, we take d as an integer in what follows.

We comment that a periodic stationary time series with period T is not stationary in the usual covariance sense unless $T = 1$. However, periodic time series with period T are T-variate stationary series. The regression in (3.4.1) can be written as a Seemingly Unrelated Regression (SUR) model, which was studied by Zellner (1962).

Let $\vec{X}_\nu = \{X_{mT+\nu}\}_{m=0}^{d-1}$ denote the observed data during season ν; $\vec{\epsilon}_\nu = \{\epsilon_{mT+\nu}\}_{m=0}^{d-1}$ the errors during season ν; and $\vec{\beta}_\nu = (\mu_\nu, \alpha_\nu, \Delta_{\nu,2}, \ldots, \Delta_{\nu,k})'$ the parameters during season ν. Let D_ν denote the design matrix at season ν and note that D_ν has size $d \times (k+1)$ with

$$D_\nu = \begin{pmatrix}
1 & \nu & 0 & 0 & \ldots & 0 & 0 \\
1 & T+\nu & 0 & 0 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & (n_{1,\nu}-1)T+\nu & 0 & 0 & \ldots & 0 & 0 \\
1 & n_{1,\nu}T+\nu & 1 & 0 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & (n_{1,\nu}+n_{2,\nu}-1)T+\nu & 1 & 0 & \ldots & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & (\sum_{i=1}^{k-1} n_{i,\nu})T+\nu & 0 & 0 & \ldots & 0 & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
1 & (d-1)T+\nu & 0 & 0 & \ldots & 0 & 1
\end{pmatrix},$$

where $n_{i,\nu}$ is the number of observations in season under regime i. The SUR model can be expressed in multivariate form as

$$\vec{X} = D\vec{\beta} + \vec{\epsilon}, \tag{3.4.2}$$

where $\vec{X} = (\vec{X}_1', \vec{X}_2', \dots, \vec{X}_T')'$, $\vec{\epsilon} = (\vec{\epsilon}_1', \vec{\epsilon}_2', \dots, \vec{\epsilon}_T')'$, $\vec{\beta} = (\vec{\beta}_1', \vec{\beta}_2', \dots, \vec{\beta}_T')'$, and D is a block-diagonal matrix with D_ν in the νth place.

The OLS estimator of the trends $(\alpha_1, \dots, \alpha_T)'$ in model (3.4.2) is given by

$$\hat{\vec{\alpha}}_{\text{OLS}} = (J_T \otimes \vec{e}_2')(D'D)^{-1}D'\vec{X}, \tag{3.4.3}$$

where $J_T = (1, 1, \dots, 1)'$ is the T-dimensional one vector, $\vec{e}_2 = (0, 1, 0, \dots, 0)'$ is the $(k+1)$-dimensional vector with 1 in the second place and all other entries of zero, and $\otimes$ denotes Kronecker product.

The variance and covariance matrix of the OLS trend estimator can be obtained from (3.4.3) as

$$\text{Var}(\hat{\vec{\alpha}}_{\text{OLS}}) = (J_T \otimes \vec{e}_2')(D'D)^{-1}D'\Gamma_{\vec{\epsilon}}D(D'D)^{-1}(J_T' \otimes \vec{e}_2), \tag{3.4.4}$$

where $\Gamma_{\vec{\epsilon}} = [\Gamma_{ij}]_{i,j=1}^T$ is the covariance matrix of $\vec{\epsilon}$ and

$$\Gamma_{ij} = \begin{pmatrix} \gamma_{ij}(0) & \gamma_{ij}(1) & \dots & \gamma_{ij}(d-1) \\ \gamma_{ij}(-1) & \gamma_{ij}(0) & \dots & \gamma_{ij}(d-2) \\ \vdots & \vdots & \vdots & \vdots \\ \gamma_{ij}(-(d-1)) & \gamma_{ij}(-(d-2)) & \dots & \gamma_{ij}(0) \end{pmatrix}.$$

Notice that $\gamma_{\nu\nu}(\cdot)$ is the autocovariance of the season ν series $\{X_{mT+\nu}\}_{m=0}^d$ and $\gamma_{ij}(\cdot)$, $i \neq j$, is the cross-covariance function of $\{X_{mT+i}\}_{m=0}^d$ and $\{X_{mT+j}\}_{m=0}^d$. We remind the reader that $\gamma_{ij}(h) \neq \gamma_{ji}(h)$ in general.

In the SUR model (3.4.2), for each fixed season ν, the OLS estimator of α_ν can be constructed from $\vec{X}_\nu$ and has explicit form, similar in structure to (3.2.1):

$$\hat{\alpha}_{\nu,\text{OLS}} = \frac{\sum_{m=0}^{d-1} \xi_{\nu,m} X_{mT+\nu}}{T\sum_{m=0}^{d-1} \xi_{\nu,m}^2}, \tag{3.4.5}$$

where

$$\{\xi_{\nu,m}\} = \begin{pmatrix} 0 - \frac{n_{1,\nu}-1}{2} \\ 1 - \frac{n_{1,\nu}-1}{2} \\ \vdots \\ (n_{1,\nu}-1) - \frac{n_{1,\nu}-1}{2} \\ \vdots \\ 0 - \frac{n_{k,\nu}-1}{2} \\ 1 - \frac{n_{k,\nu}-1}{2} \\ \vdots \\ (n_{k,\nu}-1) - \frac{n_{k,\nu}-1}{2} \end{pmatrix} \tag{3.4.6}$$

is a d-dimensional vector for $\nu = 1, \ldots, T$.

Using linearity of $\hat{\alpha}_{\nu,\text{OLS}}$ in $\vec{X}_\nu$, the variance of $\hat{\alpha}_{\nu,\text{OLS}}$ is seen to be

$$\text{Var}(\hat{\alpha}_{\nu,\text{OLS}}) = \frac{\sum_{h=-(d-1)}^{d-1} \left(\frac{\sum_{m=0}^{(d-1-|h|)} \xi_{\nu,m}\xi_{\nu,m+|h|}}{\sum_{m=0}^{d-1} \xi_{\nu,m}^2} \right) \gamma_{\nu,\nu}(h)}{T^2 \sum_{m=0}^{d-1} \xi_{\nu,m}^2}. \tag{3.4.7}$$

Similarly, the covariance between $\hat{\alpha}_{i,\text{OLS}}$ and $\hat{\alpha}_{j,\text{OLS}}$ is

$$\text{Cov}(\hat{\alpha}_{i,\text{OLS}}, \hat{\alpha}_{j,\text{OLS}}) = \frac{\sum_{h=-(d-1)}^{d-1} \left(\frac{\sum_{m=0}^{(d-1-|h|)} \xi_{i,m}\xi_{j,m+|h|}}{\sqrt{\sum_{m=0}^{d-1} \xi_{i,m}^2 \sum_{m=0}^{d-1} \xi_{j,m}^2}} \right) \gamma_{i,j}(h)}{T^2 \sqrt{\sum_{m=0}^{d-1} \xi_{i,m}^2 \sum_{m=0}^{d-1} \xi_{j,m}^2}}. \tag{3.4.8}$$

Notice that for each season ν, application of the Theorem 3.3.1 immediately gives consistency and asymptotically normality of $\hat{\alpha}_{\nu,\text{OLS}}$ with minimal assumptions. We state this in Theorem 3.4.1 below.

Theorem 3.4.1 Suppose that $\{\vec{\epsilon}_t\} = \{(\epsilon_{1,t}, \ldots, \epsilon_{T,t})'\}$ is the T-variate stationary time series in the causal linear process

$$\vec{\epsilon}_t = \sum_{l=0}^{\infty} \Psi_l \vec{Z}_{t-l}, \tag{3.4.9}$$

where $\{\vec{Z}_t\}$ are independent and identically distributed with mean zero and invertible variance-covariance matrix Σ, and $\{\Psi_l\}_{l=0}^{\infty} = \{[\psi_l(i,j)]_{i,j=1}^{T}\}$ is a sequence of $T \times T$ matrices such that $\sum_{l=0}^{\infty} |\psi_l(i,j)| < \infty$, $i, j = 1, \ldots, T$.

Suppose that the regime lengths $\{n_i\}$ obey the same sampling assumptions imposed in Theorem 3.3.1 (the regime changes are non-periodic in particular). Then as $d \to \infty$,

$$T\left(\sum_{m=0}^{d-1} \xi_{\nu,m}^2\right)^{\frac{1}{2}} (\hat{\alpha}_{\nu,\text{OLS}} - \alpha_\nu) \xrightarrow{D} N(0, V_\nu), \tag{3.4.10}$$

where $V_\nu = \sum_{h=-\infty}^{\infty} r_h(\nu,\nu)\gamma_{\nu,\nu}(h)$ and

$$r_h(\nu,\nu) = \lim_{d\to\infty} \frac{\sum_{m=0}^{(d-1-|h|)} \xi_{\nu,m}\xi_{\nu,m+|h|}}{\sum_{m=0}^{d-1} \xi_{\nu,m}^2}, \qquad h = 0, \pm 1, \pm 2, \ldots. \tag{3.4.11}$$

Whereas $\hat{\alpha}_{\nu,\text{OLS}}$ is consistent and asymptotically normal, we question its asymptotic efficiency here. In particular, note that $\hat{\alpha}_{\nu,\text{OLS}}$ uses only data from season ν. The series values outside of season ν also contain some information about α_ν when the errors are correlated. In future work, we will address such efficiency issues, being content for the moment to construct any consistent estimator and quantify its asymptotic properties.

3.5 Appendix

Proof of Lemma 1. For a fixed sequence $\{n_i\}_{i=1}^{\infty}$, let $m_\ell(n)$ be the number of regimes before time n such that $n_i = \ell$:

$$m_\ell(n) = \#\{i : 1 \le i \le k(n) \quad \text{and} \quad n_i = \ell\}.$$

We suppress dependence of $k(n)$ and $m_1(n), m_2(n)$, *etc.* on n and use the abbreviations $k, m_1, m_2,$ *etc.* Note that $n = \sum_{i=1}^{k} n_i$.

Consider R defined by

$$R = \liminf_{k\to\infty} \frac{\sum_{i=1}^{k} {n_i}^3 - n}{k}.$$

Now if $R > 0$, then

$$\sum_{i=1}^{k} {n_i}^3 - \sum_{i=1}^{k} n_i \to \infty$$

as $k \to \infty$ and the lemma follows.

To show that $R > 0$, observe that

$$R = \liminf_{k\to\infty} \frac{\sum_{i=1}^{k} n_i({n_i}^2 - 1)}{k}. \qquad (3.5.1)$$

Since $m_1 + m_2 + \ldots + m_n = k$, (3.5.1) is merely

$$\begin{aligned} R &= \liminf_{k\to\infty} \frac{\sum_{\ell=1}^{n} m_\ell \ell(\ell^2 - 1)}{k} \\ &= \liminf_{k\to\infty} \frac{\sum_{\ell=1}^{n} m_\ell \ell(\ell + 1)(\ell - 1)}{k} \\ &= \liminf_{k\to\infty} \frac{\sum_{\ell=2}^{n} m_\ell \ell(\ell + 1)(\ell - 1)}{k}. \end{aligned}$$

Now for all $\ell \geq 2$, $\ell - 1 \geq 1$ and $\ell^2 \geq 1$; hence,

$$\begin{aligned} R &\geq \liminf_{k\to\infty} \frac{\sum_{\ell=2}^{n} m_\ell \ell^2}{k} \\ &\geq \liminf_{k\to\infty} \sum_{\ell=2}^{n} \frac{m_\ell}{k} \\ &= \liminf_{k\to\infty} \frac{(m_1 + \ldots + m_n) - m_1}{k} \\ &= 1 - \liminf_{k\to\infty} \frac{m_1}{k}. \end{aligned}$$

Hence, if $\liminf_{k\to\infty} \frac{m_1}{k} < 1$, then $R > 0$ and the lemma is proved. □

Proof of Theorem 3.2.1. Observe that

$$\sum_{i=1}^{k} \sum_{t \in Y_i} (t - \bar{t}_i)^2 = \sum_{t=1}^{n} \eta_t^2,$$

where $\{\eta_t\}$ is obtained from (3.2.5). From (3.2.3) $\mathrm{Var}(\hat{\alpha}_{\mathrm{OLS}})$ is

$$\mathrm{Var}(\hat{\alpha}_{\mathrm{OLS}}) = \frac{\gamma(0) + 2\sum_{h=1}^{n-1} w_{n,h}\gamma(h)}{\sum_{t=1}^{n} \eta_t^2}, \tag{3.5.2}$$

where $w_{n,h}$ is as in (3.2.4).

By the Cauchy-Schwarz inequality,

$$\begin{aligned} \left|\sum_{t=1}^{n-h} \eta_t \eta_{t+h}\right| &\leq \left(\sum_{t=1}^{n-h} \eta_t^2\right)^{\frac{1}{2}} \left(\sum_{t=1}^{n-h} \eta_{t+h}^2\right)^{\frac{1}{2}} \\ &\leq \sum_{t=1}^{n} \eta_t^2, \end{aligned} \tag{3.5.3}$$

and $|w_{n,h}| \leq 1$ for all n and h. Using this in (3.5.2) provides

$$\begin{aligned} \mathrm{Var}(\hat{\alpha}_{\mathrm{OLS}}) &\leq \frac{\gamma(0) + 2\sum_{h=1}^{n-1} |w_{n,h}||\gamma(h)|}{\sum_{t=1}^{n} \eta_t^2} \\ &\leq \frac{\gamma(0) + 2\sum_{h=1}^{n-1} |\gamma(h)|}{\sum_{t=1}^{n} \eta_t^2} \\ &\leq \frac{\sum_{h=-(n-1)}^{n-1} |\gamma(h)|}{\sum_{t=1}^{n} \eta_t^2}. \end{aligned} \tag{3.5.4}$$

As $\sum_{h=0}^{\infty} |\gamma(h)| < \infty$ by our short-memory assumption and $\sum_{t=1}^{n} \eta_t^2 \to \infty$ by Lemma 1 (recall that (3.2.6) holds), (3.5.4) shows that $\mathrm{Var}(\hat{\alpha}_{\mathrm{OLS}}) \to 0$ as $n \to \infty$. □

Proof of Theorem 3.3.1. Note that $\hat{\alpha}_{\mathrm{OLS}}$ in (3.2.1) can be written as

$$\hat{\alpha}_{\mathrm{OLS}} = \frac{\sum_{i=1}^{k} \sum_{t \in Y_i} (t - \bar{t}_i) X_t}{\sum_{i=1}^{k} \sum_{t \in Y_i} (t - \bar{t}_i)^2}, \tag{3.5.5}$$

since $\sum_{i=1}^{k} \sum_{t \in Y_i} (t - \bar{t}_i) \bar{X}_i = 0$. Thus,

$$\hat{\alpha}_{\mathrm{OLS}} = \frac{\sum_{t=1}^{n} \eta_t X_t}{\sum_{t=1}^{n} \eta_t^2},$$

where $\{\eta_t\}$ is defined as in (3.2.5); in particular,

$$\{\eta_t\} = \begin{pmatrix} 1 - \frac{n_1+1}{2} \\ 2 - \frac{n_1+1}{2} \\ \vdots \\ n_1 - \frac{n_1+1}{2} \\ \vdots \\ 1 - \frac{n_k+1}{2} \\ 2 - \frac{n_k+1}{2} \\ \vdots \\ n_k - \frac{n_k+1}{2} \end{pmatrix}. \tag{3.5.6}$$

is an n-dimensional vector.

Let

$$\begin{aligned} Y_n &= \left(\sum_{t=1}^{n} \eta_t^2\right)^{\frac{1}{2}} (\hat{\alpha}_{\text{OLS}} - \alpha) \\ &= \frac{\sum_{t=1}^{n} \eta_t \epsilon_t}{(\sum_{t=1}^{n} \eta_t^2)^{\frac{1}{2}}}. \end{aligned}$$

By Theorem 6.3.4 of Fuller (1996), we have

$$Y_n \xrightarrow{D} N(0, V)$$

as $n \to \infty$ provided that

$$\sum_{t=1}^{n} \eta_t^2 \to \infty, \tag{3.5.7}$$

$$\frac{\eta_n^2}{\sum_{t=1}^{n} \eta_t^2} \to 0, \qquad \text{as } n \to \infty \text{ and that the limit} \tag{3.5.8}$$

$$\lim_{n\to\infty} \frac{\sum_{t=1}^{n-|h|} \eta_t \eta_{t+|h|}}{\sum_{t=1}^{n} \eta_t^2} := g(h) \text{ exists for all } h = 0, \pm 1, \pm 2, \ldots, \tag{3.5.9}$$

where $V = \sum_{h=-\infty}^{\infty} g(h)\gamma(h) \neq 0$. Our work will verify that Fuller's three conditions hold in our setting.

Under the assumptions in Theorem 3.3.1, the regime lengths $\{n_i\}_{i=1}^{\infty}$ are sampled from the ergodic Markov Chain $\{N_i\}_{i=0}^{\infty}$ on the state-space $\{1, 2, 3, \ldots\}$ with stationary measure

$\{\pi_l\}_{l=1}^{\infty}$. Let $\vec{\pi} = (\pi_1, \pi_2, \ldots)'$ and take $N_0 \overset{D}{=} \vec{\pi}$ so that $N_i \overset{D}{=} \vec{\pi}$ for all $i \geq 1$. It follows that the limits involved in (3.5.7) — (3.5.9) do not depend on the realization of $\{N_i\}$ almost surely.

Further, the ergodic properties of Markov chains give

$$\begin{aligned} \frac{m_\ell(n)}{k(n)} &= \frac{\#\{i : 1 \leq i \leq k(n) \ \text{ and } \ N_i = \ell\}}{k(n)} \\ &\to \pi_\ell \end{aligned} \tag{3.5.10}$$

almost surely as $n \to \infty$. As $\pi_\ell > 0$ for all $\ell > 0$, (3.5.10) ensures that (3.2.6) in Theorem 3.2.1 holds. Hence, Lemma 1 applies and $\sum_{t=1}^{n} \eta_t^2 \to \infty$ as $n \to \infty$. It now follows that (3.5.7) holds.

Observe that

$$\frac{\eta_n^2}{\sum_{t=1}^{n} \eta_t^2} = \frac{\frac{(n_k-1)^2}{4}}{\sum_{t=1}^{n} \eta_t^2}. \tag{3.5.11}$$

Since n_k is finite almost surely and $\sum_{t=1}^{n} \eta_t^2 \to \infty$ by Lemma 1, (3.5.8) also holds.

For (3.5.9), note that

$$\lim_{n\to\infty} \frac{\frac{1}{n}\sum_{t=1}^{n-h} \eta_t \eta_{t+h}}{\frac{1}{n}\sum_{t=1}^{n} \eta_t^2} \tag{3.5.12}$$

exists if $\lim_{n\to\infty} n^{-1} \sum_{t=1}^{n-h} \eta_t \eta_{t+h}$ and $\lim_{n\to\infty} n^{-1} \sum_{t=1}^{n} \eta_t^2$ exist.

From (3.5.10) and (3.3.4), observe that $\lim_{n\to\infty} k(n)/n$ exists:

$$\begin{aligned} \frac{k(n)}{n} &= \frac{\frac{1}{k}\sum_{\ell=1}^{n} m_\ell}{\frac{1}{k}\sum_{\ell=1}^{n} m_\ell \ell} \\ &= \frac{1}{\sum_{\ell=1}^{n} \frac{m_\ell}{k} \ell} \\ &\to \frac{1}{\sum_{\ell=1}^{\infty} \pi_\ell \ell}, \\ &= \frac{1}{\mathrm{E}[N_0]}, \end{aligned} \tag{3.5.13}$$

where the limit existence follows by Proposition 3.3.1 of Ross (1996) and (3.3.4) ensures that $\mathrm{E}[N_0] < \infty$.

We now argue that $\lim_{n\to\infty} n^{-1}\sum_{t=1}^{n}\eta_t^2$ exists. For this, observe that

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n}\eta_t^2 &= \frac{1}{n}\sum_{\ell=1}^{n} m_\ell \ell(\ell+1)(\ell-1)/12 \\ &= \left(\frac{k}{n}\right)\left(\sum_{\ell=1}^{n}\left(\frac{m_\ell}{k}\right)\frac{\ell(\ell+1)(\ell-1)}{12}\right). \qquad (3.5.14)\end{aligned}$$

Since $k/n \to \mathrm{E}[N_0]^{-1}$, it suffices to prove that $\lim_{n\to\infty}\sum_{\ell=1}^{n}(\frac{m_\ell}{k})\frac{\ell(\ell+1)(\ell-1)}{12}$ exists. For this, we merely need that $\sum_{\ell=1}^{n}\ell^3 m_\ell/k$ converges. To see this latter point, we borrow an argument from Markov chain theory. Note that for each $M \geq 1$,

$$\begin{aligned}\liminf_{n\to\infty}\sum_{\ell=1}^{M}\left(\frac{m_\ell}{k}\right)\ell^3 &\geq \sum_{\ell=1}^{M}\left(\lim_{n\to\infty}\frac{m_\ell}{k}\right)\ell^3 \\ &= \sum_{\ell=1}^{M}\ell^3\pi_\ell \qquad (3.5.15)\end{aligned}$$

since $m_\ell/k \to \pi_\ell$ for each fixed ℓ. Now (3.5.15) implies that

$$\liminf_{n\to\infty}\sum_{\ell=1}^{\infty}\left(\frac{m_\ell}{k}\right)\ell^3 \geq \sum_{\ell=1}^{\infty}\ell^3\pi_\ell.$$

Also,

$$\begin{aligned}\limsup_{n\to\infty}\sum_{\ell=1}^{\infty}\left(\frac{m_\ell}{k}\right)\ell^3 &\leq \sum_{\ell=1}^{\infty}\left(\limsup_{n\to\infty}\frac{m_\ell}{k}\right)\ell^3 \\ &= \sum_{\ell=1}^{\infty}\ell^3\pi_\ell.\end{aligned}$$

Hence, under condition (3.3.4)

$$\begin{aligned}\lim_{n\to\infty}\sum_{\ell=1}^{n}\left(\frac{m_\ell}{k}\right)\ell^3 &= \sum_{\ell=1}^{\infty}\ell^3\pi_\ell \\ &< \infty,\end{aligned}$$

and

$$\lim_{n\to\infty}\sum_{\ell=1}^{n}\left(\frac{m_\ell}{k}\right)\frac{\ell(\ell+1)(\ell-1)}{12} = \sum_{\ell=1}^{\infty}\frac{\ell(\ell+1)(\ell-1)}{12}\pi_\ell.$$

.

Hence, from (3.5.14) $\lim_{n\to\infty} n^{-1}\sum_{t=1}^{n}\eta_t^2$ exists and equals

$$\left(\frac{1}{\sum_{\ell=1}^{\infty}\ell\pi_\ell}\right)\left(\sum_{\ell=1}^{\infty}\frac{\ell(\ell+1)(\ell-1)}{12}\pi_\ell\right)=\left(\frac{1}{\mathrm{E}[N_0]}\right)\left(\frac{\mathrm{E}[N_0^3]-\mathrm{E}[N_0]}{12}\right).$$

To argue that $\lim_{n\to\infty} n^{-1}\sum_{t=1}^{n-h}\eta_t\eta_{t+h}$ exists for each $h\geq 1$, first, consider the case $h=1$. Tedious manipulations provide the explicit form

$$\frac{\sum_{t=1}^{n-h}\eta_t\eta_{t+h}}{n}=1+\frac{\sum_{t=1}^{n}\eta_t^2}{n}-\frac{\frac{1}{4}(\sum_{i=1}^{k}n_i^2+\sum_{i=1}^{k-1}n_in_{i+1})+\frac{1}{4}(n_1+n_k)-\frac{1}{4}}{n}. \qquad (3.5.16)$$

Hence, $\lim_{n\to\infty} n^{-1}\sum_{t=1}^{n-1}\eta_t\eta_{t+1}$ exists if $\lim_{k\to\infty} k^{-1}\sum_{i=1}^{k-1}N_iN_{i+1}$ exists. By Theorem 1.1 and Theorem 1.3 of Billingsley (1961),

$$\begin{aligned}\frac{\sum_{i=1}^{k-1}N_iN_{i+1}}{k} &\xrightarrow{a.s.} \mathrm{E}[N_0N_1]\\ &= \sum_{s=1}^{\infty}\sum_{m=1}^{\infty}sm\pi_s p_{sm}.\end{aligned} \qquad (3.5.17)$$

Hence, $\lim_{n\to\infty} n^{-1}\sum_{t=1}^{n-1}\eta_t\eta_{t+1}$ exists with

$$\begin{aligned}&\lim_{n\to\infty}\frac{\sum_{t=1}^{n-1}\eta_t\eta_{t+1}}{n}\\ &= 1+\left(\frac{1}{\sum_{\ell=1}^{\infty}\ell\pi_\ell}\right)\left(\sum_{\ell=1}^{\infty}\frac{\ell(\ell+1)(\ell-1)}{12}\pi_\ell-\frac{\sum_{\ell=1}^{\infty}\ell^2\pi_\ell}{4}-\frac{\sum_{s=1}^{\infty}\sum_{m=1}^{\infty}sm\pi_s p_{sm}}{4}\right)\\ &= 1+\left(\frac{1}{\mathrm{E}[N_0]}\right)\left(\frac{\mathrm{E}[N_0^3]-\mathrm{E}[N_0]}{12}-\frac{\mathrm{E}[N_0^2]}{4}-\frac{\mathrm{E}[N_0N_1]}{4}\right).\end{aligned}$$

For $h\geq 2$, the arguments that $\lim_{n\to\infty} n^{-1}\sum_{t=1}^{n-h}\eta_t\eta_{t+h}$ exist are analogous and every bit as tedious.

Putting the above facts together allows us to finally conclude that the limit in (3.5.9) exists. Thus, Fuller's three conditions are satisfied and $Y_n \xrightarrow{D} N(0,V)$ as $n\to\infty$ where $V=\sum_{h=-\infty}^{\infty}g(h)\gamma(h)$.

□

3.6 References

[1] Anderson, T.W. (1971). *The Statistical Analysis of Time Series*, John Wiley, New York.

[2] Billingsley, P. (1961). *Statistical Inference for Markov Processes*, The University of Chicago Press.

[3] Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods*, Springer-Verlag, New York.

[4] Easterling, D.R. and Peterson, T. (1995). A new method for detecting undocumented discontinuities in climatological time series, *International Journal of Climatology*, **15**, 369-377.

[5] Eicker, F. (1967). Limit theorems for regressions with unequal and dependent errors, In: *Proceeding of The Fifth Berkeley Symposium: Mathematical Statistics and Probability (Berkeley California 1965/1966)*, **1** Statistics, 59–82, University of California Press, Berkeley.

[6] Fuller, W. (1996). *Introduction to Statistical Time Series*, Second Edition, John Wiley, New York.

[7] Grenander, U. (1954). On the estimation of regression coefficients in the case of autocorrelated disturbances, *Annals of Mathematical Statistics*, **25**, 252–272.

[8] Grenander, U. and Rosenblatt M. (1957). *Statistical Analysis of Stationary Time Series*, John Wiley, New York.

[9] Hannan, E.J. (1961). A central limit theorem for systems of regressions, *Proceedings of the Cambridge Philosophical Society*, **57**, 583–588.

[10] Hannan, E.J. (1970). *Multiple Time Series*, John Wiley, New York.

[11] Hinkley, D.V. (1969). Inference about the intersection in two-phase regression, *Biometrika*, **56**, 495-504.

[12] Hinkley, D.V. (1971b). Inference in two-phase regression, *Journal of the American Statistical Association*, **66**, 736-743.

[13] Kalashnikov, V.V. (1994). *Topics on Regenerative Processes*, CRC Press, Boca Raton.

[14] Lund, R.B. and Reeves, J. (2002). Detection of undocumented changepoints: A revision of the two-phase regression model, *Journal of Climate*, **15**, 2547-2554.

[15] Mitchell, J.M.Jr. (1953). On the causes of instrumentally observed secular temperature trends, *Journal of Applied Meteorology*, **10**, 244–261.

[16] Quandt, R.E. (1958). The estimation of the parameters of a linear regression system obeys two separate regimes, *Journal of the American Statistical Association*, **53**, 873-880.

[17] Quandt, R.E. (1960). Test of the hypothesis that a linear regression system obeys two separate regimes, *Journal of the American Statistical Association*, **55**, 324-330.

[18] Ross, S.M. (1996). *Stochastic Processes*, John Wiley and Sons, New York.

[19] Solow, A.R. (1987). Testing for climate change: an application of the two-phase regression model, *Journal of Climate and Applied Meteorology*, **26**, 1401-1405.

[20] Vincent, L.A. (1998). A technique for the identification of inhomogeneities in Canadian temperature series, *Journal of Climate*, **11**, 1094-1104.

[21] Zellner, A. (1962). An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias, *Journal of the American Statistical Association*, **57**, 348-368.

Chapter 4

An Update of United States Temperature Trends[1]

[1] Lu, Q., Lund, R., and Seymour, L. Submitted to *Journal of Climate,* 7/15/2004.

Abstract

This paper updates the temperature trend study of the contiguous 48 United States in Lund *et al.* (2001) for data observed during the last four years. Better methods for handling missing observations and model parsimony are among the improvements. The number of stations with usable data has now increased from 359 to 969, thereby improving the accuracy of the reported spatial patterns in the trends. The methodological improvements and additional four years of data produce slightly smaller trend estimate standard errors at fixed stations. Warming is again found in the Northeast, West, Northern Midwest and cooling in the Southeast; overall, the trends here suggest more warming than the Lund *et al.* (2001) study. Finally, the estimated temperature trends are regressed on several covariates, revealing a significant negative correlation between temperature trends and precipitation.

Key Words and Phrases: Head-Banging Algorithm; Linear Trend; Spatial Autocorrrelation; Standard Error; Temporal Autocorrelation; Time Series.

4.1 Introduction

This paper updates the United States temperature trend study in Lund *et al.* (2001) for recent data and for improvements in the methods of analysis. United States temperature changes have been previously studied in Diaz and Quayle (1980) and Karl *et al.* (1995) (amongst many others). Global temperature change studies (e.g. Ellsaesser *et al.* 1986, Hansen *et al.* 1999, 2001, and the references therein) also provide insight into United States temperature trends.

The Lund *et al.* (2001) study was a statistically detailed examination of temperature changes in the United States. It made three fundamental analysis contributions. First, the reported standard errors of the trend estimates accounted for temporal correlations in the data. Second, seasonal aspects were considered in that monthly rather than yearly series were examined. Third, the analysis allowed for the crucial effect of changepoints at times where changepoints were known to occur. Many (unfortunately, not all) changepoint times are explicitly noted in the station history logs; such records permit for adjustment of this factor (cf. Karl and Williams 1987; Karl *et al.* 1990).

However, there were some drawbacks in the Lund *et al.* (2001) study. One is that the spatial coverage was not good — only 359 of 1221 stations in the network were deemed usable due to data quality (missing data in particular). For instance, only 2 stations in Oklahoma, 4 stations in Mississippi, and 10 stations in Texas were in this study. The updated study contains 34 Oklahoma stations, 29 Mississippi stations, and 39 Texas stations. The primary reason driving the improved spatial coverage lies in the revised methods for handling missing data. As the number of stations with good quality data has now increased from 359 to 969, the spatial coverage of climate change in the United States is now significantly more complete and the resulting analysis more reliable.

The other main improvement in this update lies with parsimony issues for handling changepoints. The Lund *et al.* (2001) study expended a large number of changepoint parameters in its mathematical regression model: one for each different changepoint and season (month). For example, in a monthly series (period 12) with 6 changepoints per century, which is the average number of documented changepoints seen over all 969 stations in this study, the regression model employed used $84 = 24 + 12 \times (6 - 1)$ parameters. In contrast, the periodic simple linear regression model used here assumes the same changepoint mean shift response over all seasons, thereby reducing the number of parameters to a more parsimonious number of $29 = 24 + (6 - 1)$.

The rest of the paper is organized as follows. We first describe the sources of data and quality restrictions imposed upon series entering this study. The methods used to obtain the trends and their uncertainty margins are then narrated. A case study of one of the 969 stations is then presented for illustration. Spatial contour maps of the smoothed trend estimates are then presented and discussed. Finally, we regress the estimated trends on factors such as precipitation and elevation in an attempt to better understand them.

4.2 The Data

The data set used in this study is taken from the United States Historical Climatology Network (USHCN). The USHCN consists of 1221 stations in the 48 contiguous United States; this data is raw but adjusted for biases due to the time of observation. Karl *et al.* (1990) describes the USHCN in more detail.

Data quality is an issue due to site changes (changepoints) and missing observations. We take a changepoint as a change of station location, station instrument, or station shelter. The dates of the changepoints that are known were noted and rounded to the nearest month. Stations having two changepoints within four months were discarded; in comparison, Lund *et al.* (2001) discarded series with two or more changepoints occurring within three years. The driving improvement here is that the regression model introduced in the next section is more parsimonious and can accommodate a smaller number of observations taken from one 'regime'. As before, a minimum of 75 years of record is required to enter the study and stations missing 5% or more of their observations during their period of record were discarded, or the starting date of the series was advanced if possible so as to make the series meet the above constraints.

After these data quality restrictions were imposed, 969 stations remained for study. Figure 4.1 graphically depicts the location of these stations; the spatial coverage is excellent. The duration of record at each station is variable, with the average series containing 103 years of data. The starting month of every series is rounded to the 'nearest viable January'; the last observation is taken during December of 2000. The earliest starting year is 1812 (New Bedford, MA); the latest starting year is 1926 (8 stations).

Any missing observations that remain were infilled with a model-based Expectation-Maximization (EM) algorithm (cf. Dempster *et al.* 1977). In the first step, missing values are replaced with regression-based interpolations; then an ARMA model was chosen for the residuals of the regression fit via the AICC statistic (cf. Brockwell and Davis 1991). Predictions for the missing values were computed as one-step-ahead predictions based on the chosen ARMA model. The process is iterated until the ARMA parameters and predictions converge. Such an EM procedure can be used to infill series with many consecutive missing observations; in comparison, stations with six or more consecutive points missing were discarded in Lund *et al.* (2001). Infilling is needed for automation of computation; specifically, it would be virtually impossible to individually handle all possible missing data configurations that arise in the 969 stations.

4.3 Statistical Methods

4.3.1 Periodic regression model

Consider one fixed temperature station. As the data are monthly in structure, our methods center on the periodic simple linear regression model under multiple changepoints with period $T = 12$:

$$X_{mT+\nu} = \mu_\nu + \alpha_\nu(mT + \nu) + \delta_{mT+\nu} + \epsilon_{mT+\nu}, \tag{4.3.1}$$

where the series $\{\epsilon_{mT+\nu}\}$ is zero mean random error with periodic temporal autocovariances as elaborated upon below. The notation here uses $X_{mT+\nu}$ as the observed monthly mean temperature during the νth month of year m, where ν is a monthly index satisfying $1 \leq \nu \leq T$ and m is a yearly index satisfying $0 \leq m \leq d - 1$. Time is scaled at each station so as to make $m = 0$ the first year in the data record. The data record length is $n = dT$, where d is the total number of years of data; to avoid trite work, we take d as an integer.

For parameter interpretations, μ_ν is the average temperature during month ν in the absence of trend ($\alpha_\nu = 0$) and changepoints ($\delta_{mT+\nu} \equiv 0$); α_ν is the month ν linear trend, which will be our focus later, or the average temperature change rate during month ν in the absence of changepoints; and $\{\delta_{mT+\nu}\}$ is a changepoint mean shift factor. In particular, the changepoints have the structure

$$\delta_{mT+\nu} = \begin{cases} \Delta_1, & 1 \leq mT+\nu < \tau_1 \\ \Delta_2, & \tau_1 \leq mT+\nu < \tau_2 \\ \vdots & \vdots \\ \Delta_k, & \tau_{k-1} \leq mT+\nu \leq n \end{cases}, \tag{4.3.2}$$

where $\tau_1 < \tau_2 < \ldots < \tau_{k-1}$ are the months of the known changepoint times and $k-1$ is the total number of changepoints in the record (hence there are k different 'regimes'). One could regard the $k-1$ changepoints as inducing a k-phase simple linear regression, extending the two-phase setup in Lund and Reeves (2002). Until the first changepoint time, the mean effect in the regression is Δ_1; between the first and second changepoint times, the mean effect becomes Δ_2, and so forth. For parameter identifiability, we take $\Delta_1 = 0$ (else the Δ_i's and μ_ν's become confounded).

In general, the model in (4.3.1) is a parsimonious version of the model used in Lund *et al.* (2001) in that changepoint mean shifts in (4.3.1) are required to be the same over varying seasons. We consider such structure for three fundamental reasons. First, imposing equivalent changepoint effects across the seasons seems physically reasonable. A change of station instrumentation, for example, should not produce *radically* different responses during different seasons. The changepoint effects for a fixed station are also now easy to interpret: Δ_k is the mean shift, as measured against the first regime, of series values from the kth regime. Second, the regression parameter numbers are reduced from $2 \times T + (k-1) \times T$ in the Lund *et al.* (2001) study to $2T + k - 1$ here. This is a very large reduction in settings with a large k. One expects more efficient trend estimates than those in Lund *et al.* (2001) (smaller standard errors).

The seasonal parametrization in (4.3.1) has advantages over yearly analyses. For instance, it allows one to address uniformity of temperature change over different seasons. Indeed, some authors suspect that temperature warming is most rapid during winter due to decreased nightly radiational cooling, the latter attributed to increasing carbon dioxide (see Callendar 1961, Madden and Ramanathan 1980, and Jones *et al.* 1982 for early references).

For brevity's sake (12 maps is excessive), we partition the monthly trends into the four seasons Winter, Spring, Summer, and Fall. Winter is taken as December, January, and February (DJF); Spring as March, April, and May (MAM); Summer as June, July, and August (JJA); and Fall as September, October, and November (SON). A trend for Spring, for example, is obtained by adding the trends during March, April, and May: $\hat{\alpha}_{\mathrm{SPR}} = \hat{\alpha}_3 + \hat{\alpha}_4 + \hat{\alpha}_5$. Trends for the other three seasons, denoted by $\hat{\alpha}_{\mathrm{WIN}}, \hat{\alpha}_{\mathrm{SUM}}$, and $\hat{\alpha}_{\mathrm{FAL}}$ for Winter, Summer, and Fall, respectively, are obtained analogously. An annual or yearly trend estimate, denoted by $\hat{\alpha}_{\mathrm{YR}}$, is obtained by adding all monthly trends:

$$\hat{\alpha}_{\mathrm{YR}} = \sum_{\nu=1}^{T} \hat{\alpha}_\nu = \hat{\alpha}_{\mathrm{WIN}} + \hat{\alpha}_{\mathrm{SPR}} + \hat{\alpha}_{\mathrm{SUM}} + \hat{\alpha}_{\mathrm{FAL}}.$$

For ease of interpretation and a standard basis of comparison, all trend estimates are converted into degrees Celsius per century. This entails multiplying seasonal trends by 400 and annual trends by 100.

4.3.2 Inference for the trend parameters

To estimate α_ν for each month ν, we use the method of ordinary least squares (OLS). The OLS estimators of the α_νs cannot be explicitly derived in closed form akin to (5) in Lund *et al.* (2001). The regression in (4.3.1), however, does have the general linear model representation

$$\vec{X} = D\vec{\beta} + \vec{\epsilon},$$

where $\vec{X} = (X_1, \ldots, X_n)'$ is the observed series ($'$ indicates matrix transpose), D is the $n \times (2T + k - 1)$ dimensional design matrix

$$D = (D_1 | D_2 | D_3), \tag{4.3.3}$$

$\vec{\beta} = (\mu_1, \ldots, \mu_{12}, \alpha_1, \ldots, \alpha_{12}, \Delta_2, \ldots, \Delta_k)'$ is the $2T + k - 1$ dimensional parameter vector, and $\vec{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)'$ contains all regression errors. The component matrices D_1 and D_2 in

(4.3.3) are $n \times T$ dimensional with (i,j)th entry of zero except that $(D_1)_{mT+\nu,\nu} = 1$ and $(D_2)_{mT+\nu,\nu} = mT + \nu$ for $0 \leq m \leq d-1$ and $1 \leq \nu \leq T$. The components in D_3 are zero except for $(D_3)_{\ell,j} = 1$ when $\ell \in [\tau_i, \tau_{i+1})$ and $j = i$ as $1 \leq i \leq k-1$ (take $\tau_k = n+1$ here).

An estimator of $\vec{\beta}$ is obtained from the classical least squares formula

$$\hat{\vec{\beta}} = (D'D)^{-1}D'\vec{X}. \tag{4.3.4}$$

The estimator $\hat{\alpha}_\nu$ is contained in the $(T+\nu)$th component of $\hat{\vec{\beta}}$.

The trend estimators $\hat{\vec{\alpha}} = (\hat{\alpha}_1, \ldots, \hat{\alpha}_T)'$ are unbiased for any zero mean errors $\{\epsilon_t\}$. OLS estimators are asymptotically most efficient in time-homogeneous settings (cf. Grenander 1954); one does not expect drastic suboptimality in changepoint and/or periodic settings. The variance/covariance matrix of the trend estimators is obtained from (4.3.4):

$$\text{var}(\hat{\vec{\beta}}) = (D'D)^{-1}(D'\Gamma D)(D'D)^{-1},$$

where $\Gamma = \text{var}(\vec{\epsilon}) = E[\vec{\epsilon}\vec{\epsilon}']$ is the covariance matrix of $\{\epsilon_t\}$. Observe that $\text{var}(\hat{\vec{\beta}})$ is a $(2T + k - 1) \times (2T + k - 1)$ matrix. The standard error of $\hat{\alpha}_\nu$ is the square root of the $(T+\nu)$th diagonal component of $\text{var}(\hat{\vec{\beta}})$ (or more precisely an estimate of this quantity).

Ignoring autocorrelations in $\{\epsilon_t\}$ by taking Γ to be a multiple of the identity matrix underestimates the true variabilities in $\hat{\alpha}_\nu$ (cf. Bloomfield and Nychka 1992, Lund *et al.* 1995 for discussion). To estimate Γ, a periodic autoregressive moving-average (PARMA) model was fitted to the residuals of the regression fit. PARMA models are flexible short-memory periodic time series models. Their development in climatological settings is described in detail in Lund *et al.* (1995, 2001) and the references therein.

4.3.3 Spatial smoothing

For each station in the study, the above methods provide a trend estimate and standard error for each of the four seasons and the entire year. To summarize these trends, we spatially smooth them by longitude and latitude with the weighted head-banging algorithm described in Hansen (1991) and Mungiole *et al.* (1999). Briefly, the weighted head-banging algorithm is a nonparametric local averaging smoother that is especially adept with rough (highly variable) spatial fields. Head-banging methods effectively preserve edge features in the spatial field while simultaneously downweighting outliers. The algorithm variant we use weights the trends inversely to their standard error before averaging; hence, more questionable trends have less overall influence on the end result. This is but one reason to pursue accurate standard errors for the individual trends. The head-banging smoothing parameters used here were 15 triples with 30 nearest neighbors for each contour map. Selection of appropriate values of head-banging parameters is important, but not overly crucial in interpreting general spatial structure of the trends. One will get a feel for smoothing aspects by comparing Figures 4.3 and 4.4 below.

Application of the head-banging algorithm yields a spatially smoothed trend estimate at each station longitude and latitude in the study; such smoothing accounts for spatial autocorrelations in the trend estimates nonparametrically. As an end step, the ESRI ARCGIS software was applied to the head-banging smoothed trends. Here, the Inverse Distance Weighted interpolation method with 12 neighbors was applied to display the head-banging smoothed trends in the contour maps. The end-product is reasonably attractive while preserving general structure.

4.4 A Comparison

To gain some feel for the methods, especially in regard to the different statistical models, we consider the station located at Chula Vista, California in a case study. The data record at this station through 1996 is plotted in Figure 1 of Lund *et al.* (2001). This series has seen three documented changepoint times since 1919; the effects of each changepoint are elaborated upon in Lund and Reeves (2002). Annual trend estimates and their standard errors are listed below.

Annual Chula Vista Trend Estimates in °C/Century

Model	Annual Trend	Standard Error
Neglecting Changepoints	2.896	0.7184
Lund *et al.* (2001)	1.126	0.6581
Model (4.3.1)	0.932	0.6117

The annual trend estimate of 0.932°C/Century based on (4.3.1) is 17% smaller than the 1.126°C/Century reported in the Lund *et al.* (2001). The standard error has decreased from 0.6581°C/Century to 0.6117°C/Century, suggesting slightly improved accuracy. Comparisons at other stations also suggest some efficiency gains; specifically, an average standard error of 0.8226°C/Century was obtained by Lund *et al.* (2001) whereas that here is 0.7646°C/Century. It should be stressed again that the numbers quoted for an analysis where changepoints are neglected are invalid, but it is interesting to see how far off they indeed are.

4.5 Overall Results

This section studies the estimated trends at all 969 stations. Figure 4.2 shows boxplots of the raw trend estimates during each season. Observe that the median line is positive in each plot, suggesting overall warming in every season. The variability of the estimated trends is greatest during Winter, minimal during Fall and Summer. There are a few outlying trends in each season, of course in part attributed to the large sample size of 969. The average trend estimate over all stations is 1.167°C/Century for Winter, 0.971°C/Century for Spring, 0.658°C/Century for Summer, and 0.515°C/Century for Fall. The season with the largest temperature variability (Winter) also shows the most warming. The average annual trend estimate over all 969 stations is 0.8278°C/Century. Overall, these results suggest more warming than those reported in Lund *et al.* (2001), where the average annual trend was 0.7253°C/Century.

Figure 4.3 is a GIS contour plot of the raw annual estimated trends without application of the head-banging smoother. The plot is included for smoothing feel, and is color coded, with red representing warming and blue cooling. One sees a speckled structure to the plot, indicating a rough spatial field, with cooling stations relatively more prevalent in the Southeastern United States and warming stations relatively more frequent in the Southern Rockies, Northern Midwest, New England, and Oregon.

Figure 4.4 is a head-banging smoothed version of the annual trends in Figure 4.3. Figure 4.4 conveys the general structure of Figure 4.3 with much of the noise variability smoothed away. Overall, much of the West appears to be warming while a status quo climate, with perhaps slight cooling, is evident in the Southeast and Ohio River Basin. This is less cooling than that reported in Lund *et al.* (2001), but it is somewhat more widespread. Some cooling in the Northern Rockies that did not appear in the Lund *et al.* (2001) study is also evident. As before, rapid warming is apparent in Maine, Southern Arizona, and Northern Minnesota. Notice the scale in the plot is not symmetric: cooling rates are not less than -1.5°C/Century whereas some warming rates exceed 2.5°C/Century.

Figures 4.5 — 4.8 present spatially smoothed contours of trend estimates for Winter, Spring, Summer, and Fall, respectively. A consistent feature in these plots is the slight cooling in the Southeast and Ohio River Basin and the warming in the Four Corners Region and Northern Midwest. The Dakotas show rapid Winter warming as does Southern Arizona and Southern California. One can discern additional structure by examining the plots in detail. We do not encourage extrapolation and/or local inference in the plots beyond very general patterns; the usual disclaimer on interpreting local variations in a rough spatial field applies.

4.6 Explanation of the Trends

The maps in Figures 4.4 — 4.8 here, along with analogous maps in Lund *et al.* (2001), suggest that the Southeastern United States and Ohio River Basin have cooled slightly over the period of record and that the Four Corners Region, Northern Midwest, much of the West, and the Northeast are warming. It would be informative to explain these patterns if possible.

We now regress the updated annual trends on the factors precipitation (average yearly), altitude of recording station, and longitude and latitude in an effort to explain the temperature changes. In particular, we fit the multiple regression model

$$\hat{\alpha}_i = a + b_1 P_i + b_2 A_i + b_3 \text{Long}_i + b_4 \text{Lat}_i + Z_i, \tag{4.6.1}$$

where $\hat{\alpha}_i$ is the estimated annual temperature change ($\hat{\alpha}_{\text{YR}}$) in °C/Century at station i, P_i is the annual average precipitation (in inches), A_i is the altitude of the station (in feet above or below mean sea level), Long_i is the longitude (in degrees), and Lat_i is the latitude (in degrees). An overall mean trend change of a is put into the model. The series $\{Z_i\}$ is assumed to be zero mean random error.

In fitting the model in (4.6.1), only precipitation is a significant explanatory factor of the estimated temperature trends at the 5% significance level. The elevation, longitude, and latitude factors were in fact highly insignificant. The p-value for the precipitation factor is, however, very small — less than 0.0001 — indicating that precipitation and temperature change are indeed correlated. The estimated regression coefficient corresponding to precipitation is $\hat{b}_1 = -1.758$.

The negative association between the precipitation and temperature change is not unexpected. Precipitation has been previously linked to temperature change (cf. Ellsaesser *et al.* 1986). An inspection of Figure 4.3 shows that most cooling stations reside in areas with heavier precipitation (for examples, the blue speckles in the mountainous stations in the Rockies). This logic does not commute as there are many stations with relatively heavier precipitation rates that are warming (New England for example); however, the widespread warming in the dry Desert Southwest and Great Plains is evident in Figure 4.4.

It would have been interesting to include forestation *changes* as an additional factor in (4.6.1); however, no consistent measure of this quantity was readily available over the same time at which the temperatures were collected. We leave this issue as well as exploration of other factors to others more knowledgeable in these areas.

Acknowledgments. National Science Foundation support from grants DMS 00071383 and DMS 0304407 is acknowledged. Comments from a referee substantially improved this manuscript.

4.7 References

[1] Bloomfield, P. and Nychka, D. (1992). Climate spectra and detecting climate change, *Climate Change*, **21**, 275–287.

[2] Brockwell, P.J. and Davis, R.A. (1991). *Time Series: Theory and Methods*, Springer-Verlag, New York.

[3] Callendar, G.S. (1961). Temperature fluctuations and trends over the earth, *Quarterly Journal of the Royal Meteorological Society*, **87**, 1–12.

[4] Dempster, A.P., Laird, N.M., and Rubin, D.B. (1977). Maximum likelihood from incomplete data via the EM algorithm, *Journal of the Royal Statistical Society, Series B*, **39**, 1–38.

[5] Diaz, H.F. and Quayle, R.G. (1980). The climate of the United States since 1895: spatial and temporal changes, *Monthly Weather Review*, **108**, 249–226.

[6] Ellsaesser, H.W., MacCracken, M.C., Walton, J.J., and Grotch, S.L. (1986). Global climate trends as revealed by the recorded data, *Reviews of Geophysics*, **24**, 745-792.

[7] Grenander, U. (1954). On the estimation of regression coefficients in the case of autocorrelated disturbances, *Annals of Mathematical Statistics*, **25**, 252–272.

[8] Hansen, K.M. (1991). Head-banging: robust smoothing in the plane, *IEEE Transactions on Geoscience and Remote Sensing*, **29**(3), 369–378.

[9] Hansen, J., Ruedy, R., Glascoe, J., and Sato, M. (1999). GISS analysis of surface temperature change, *Journal of Geophysical Research*, **104**, 30997–31022.

[10] Hansen, J., Ruedy, R., Sato, M., Imhoff, M, Lawrence, W., Easterling, D., Peterson, T., and Karl, T. (2001). A closer look at United Stated and global surface temperature change, *Journal of Geophysical Research*, **106**, 23947–23963.

[11] Jones, P.D., Wigley, T.M.L., and Kelly, P.M. (1982). Variations in surface air temperature: part 1. northern hemisphere, 1881–1980, *Monthly Weather Review*, **110**, 59–70.

[12] Karl, T.R., Knight, R.W., Easterling, D.R., and Quayle, R.G. (1995). Trends in U.S. climate during the twentieth century, *Consequences: The Nature and Implications of Environmental Change*, **1**, 3-12.

[13] Karl, T.R. and Williams, C.N., Jr. (1987). An approach to adjusting climatological time series for discontinuous inhomogeneities, *Journal of Climate and Applied Meteorology*, **26**, 1744-1763.

[14] Karl, T.R., Williams, C.N., Quinlan, F.T., and Boden, T.A. (1990). In United States Historical Climatology Network (USHCN) Serial Temperature and Precipitation Data, Environmental Sciences Division, Publication No. 3404, Carbon Dioxide Information and Analysis center, Oak Ridge National Laboratory, Oak Ridge, TN, 389pp.

[15] Lund, R.B., Hurd, H.L., Bloomfield, P., and Smith, R.L. (1995). Climatological time series with periodic correlation, *Journal of Climate*, **8**, 2787-2809.

[16] Lund, R.B. and Reeves, J. (2002). Detection of undocumented changepoints: A revision of the two-phase regression model, *Journal of Climate*, **15**, 2547-2554.

[17] Lund, R.B., Seymour, L., and Kafadar, K. (2001). Temperature trends in the United States, *Environmetrics*, **12**, 673-679.

[18] Madden, R.A. and Ramanathan, V. (1980). Detecting climate change due to increasing carbon dioxide, *Science*, **209**, 763–768.

[19] Mungiole, M., Pickle, L.W. and Simonson, K. (1999). Application of a weighted head-banging algorithm to mortality data maps, *Statistics in Medicine*, **18**, 3201–3209.

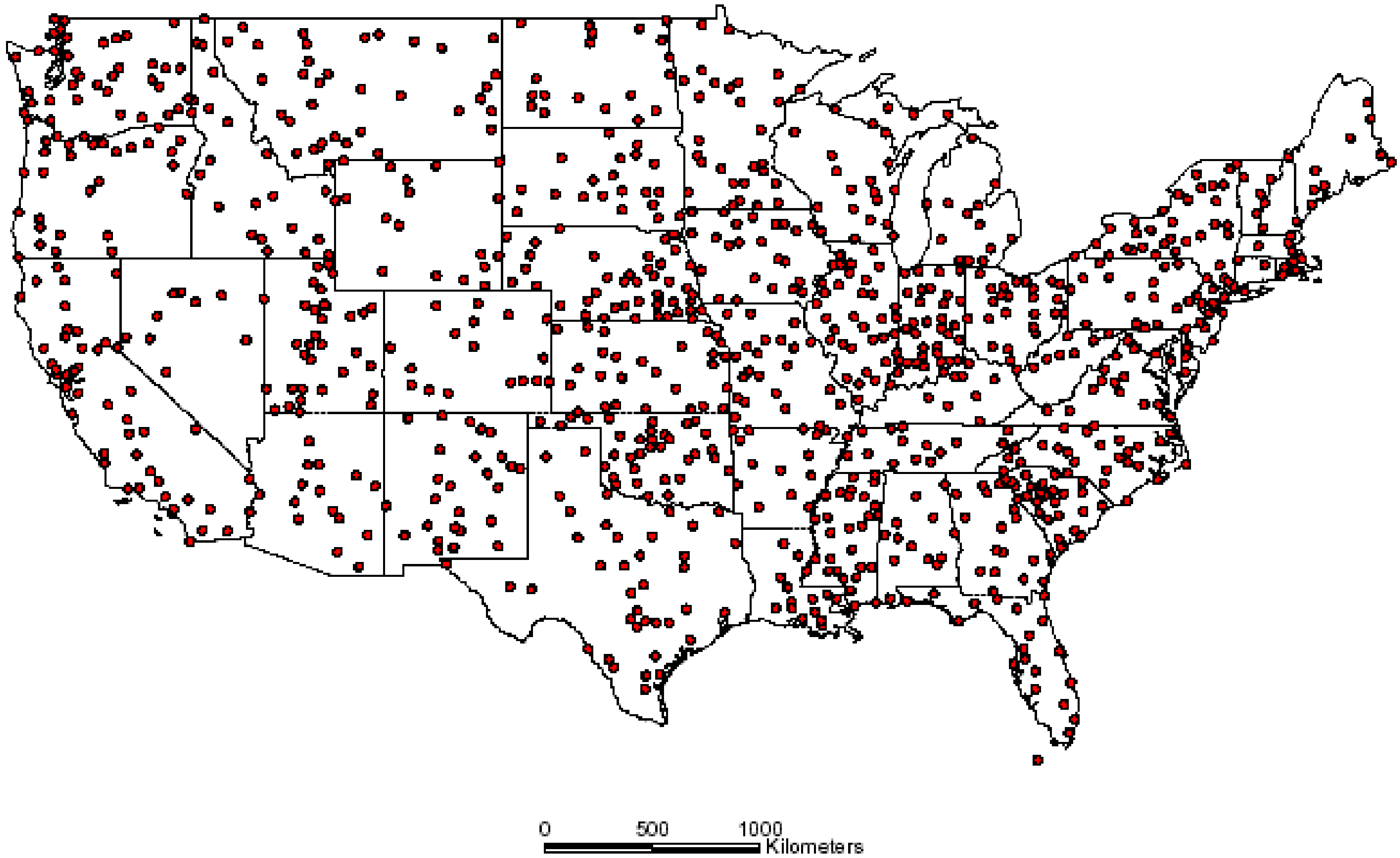

Figure 4.1: Station Locations

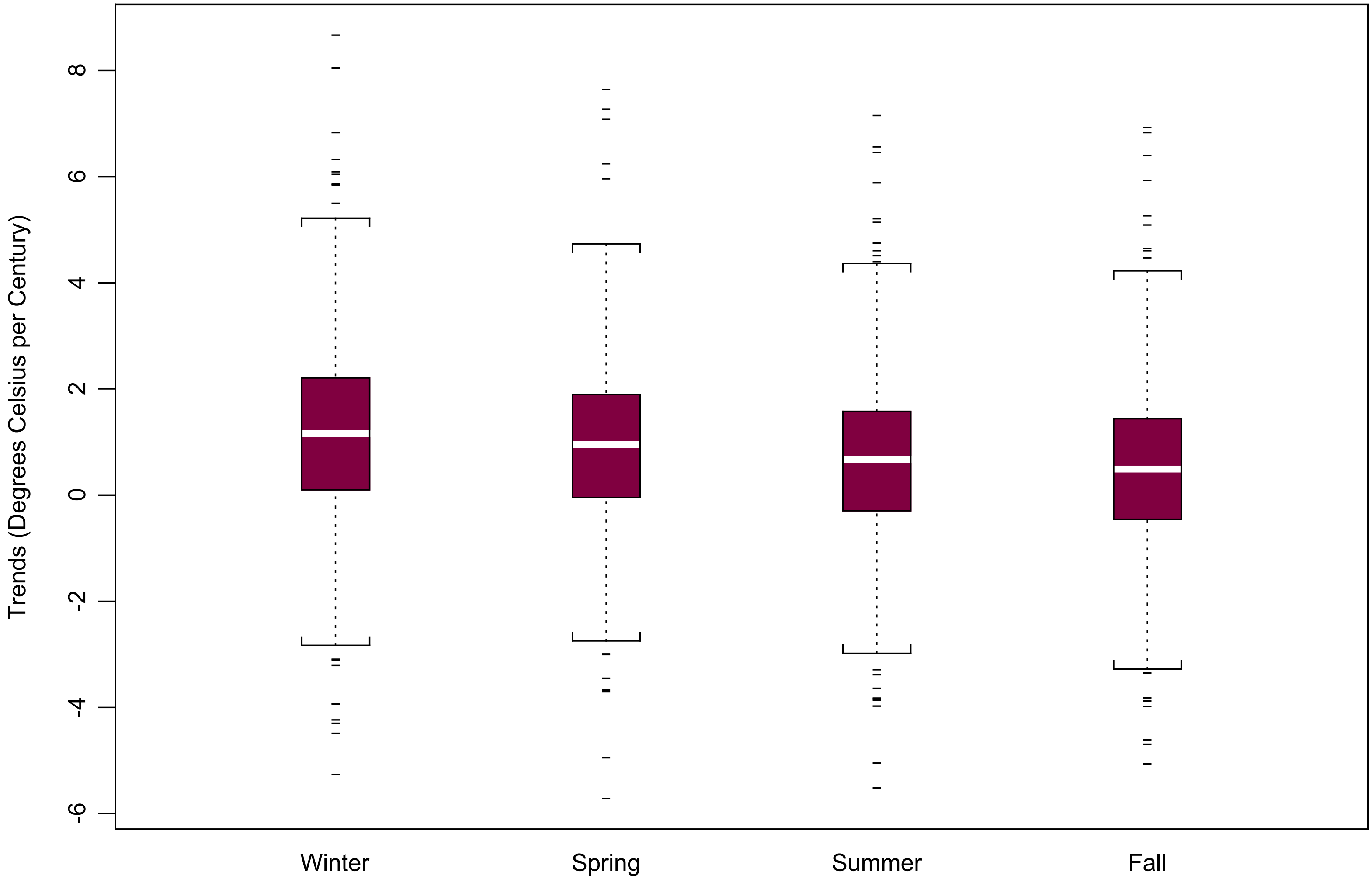

Figure 4.2: Boxplot of Trends by Season

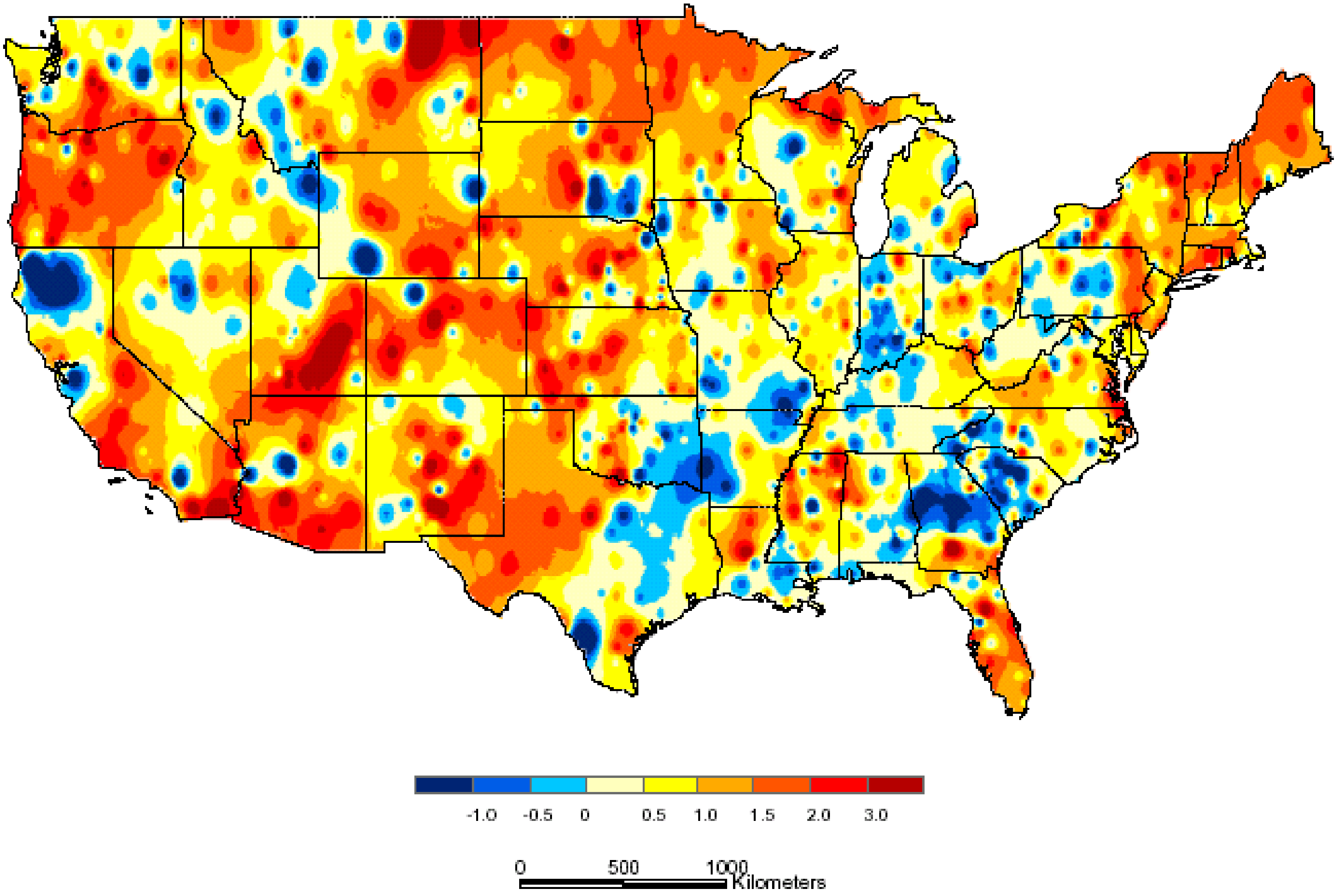

Figure 4.3: Raw Annual Trends

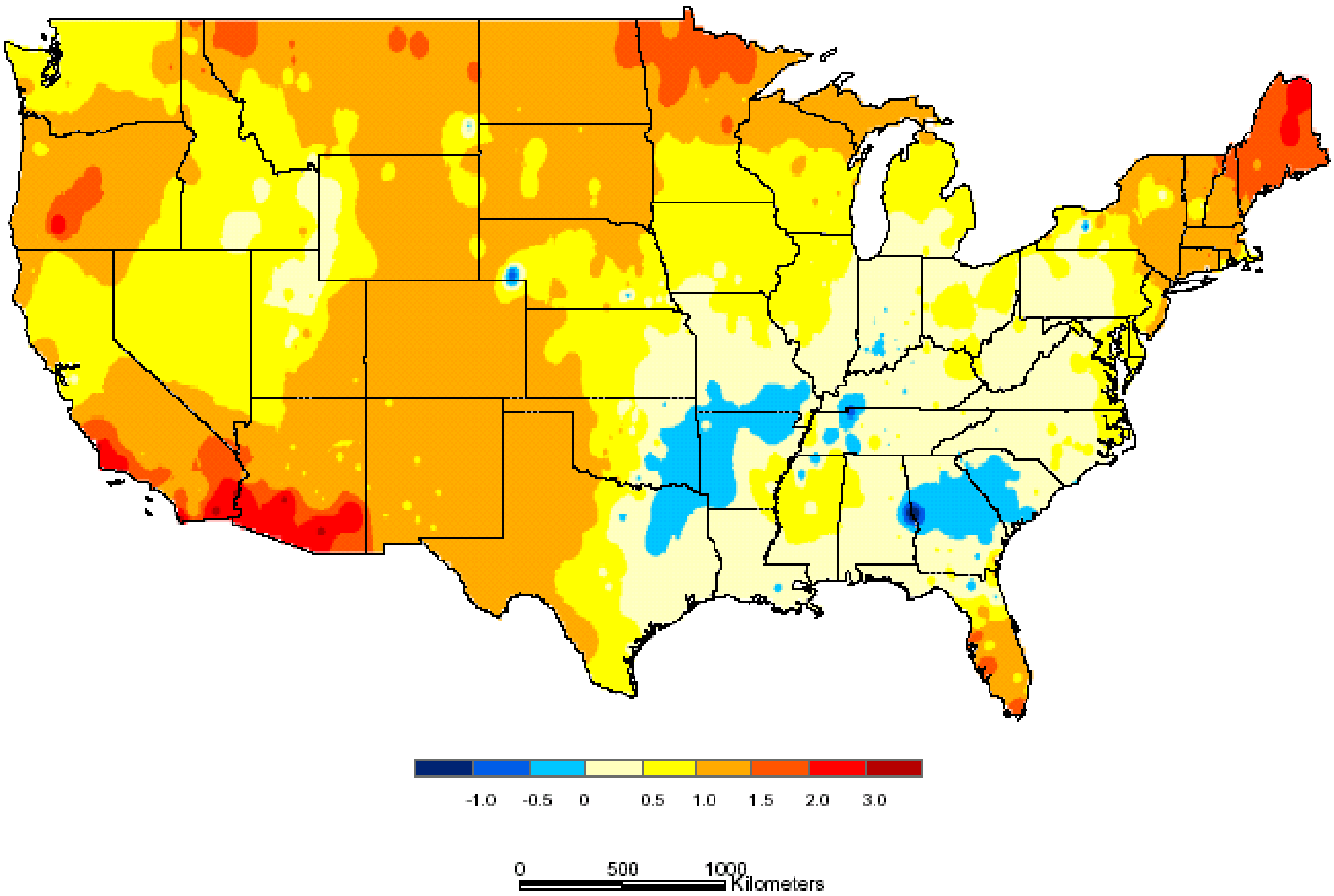

Figure 4.4: Smoothed Annual Trends

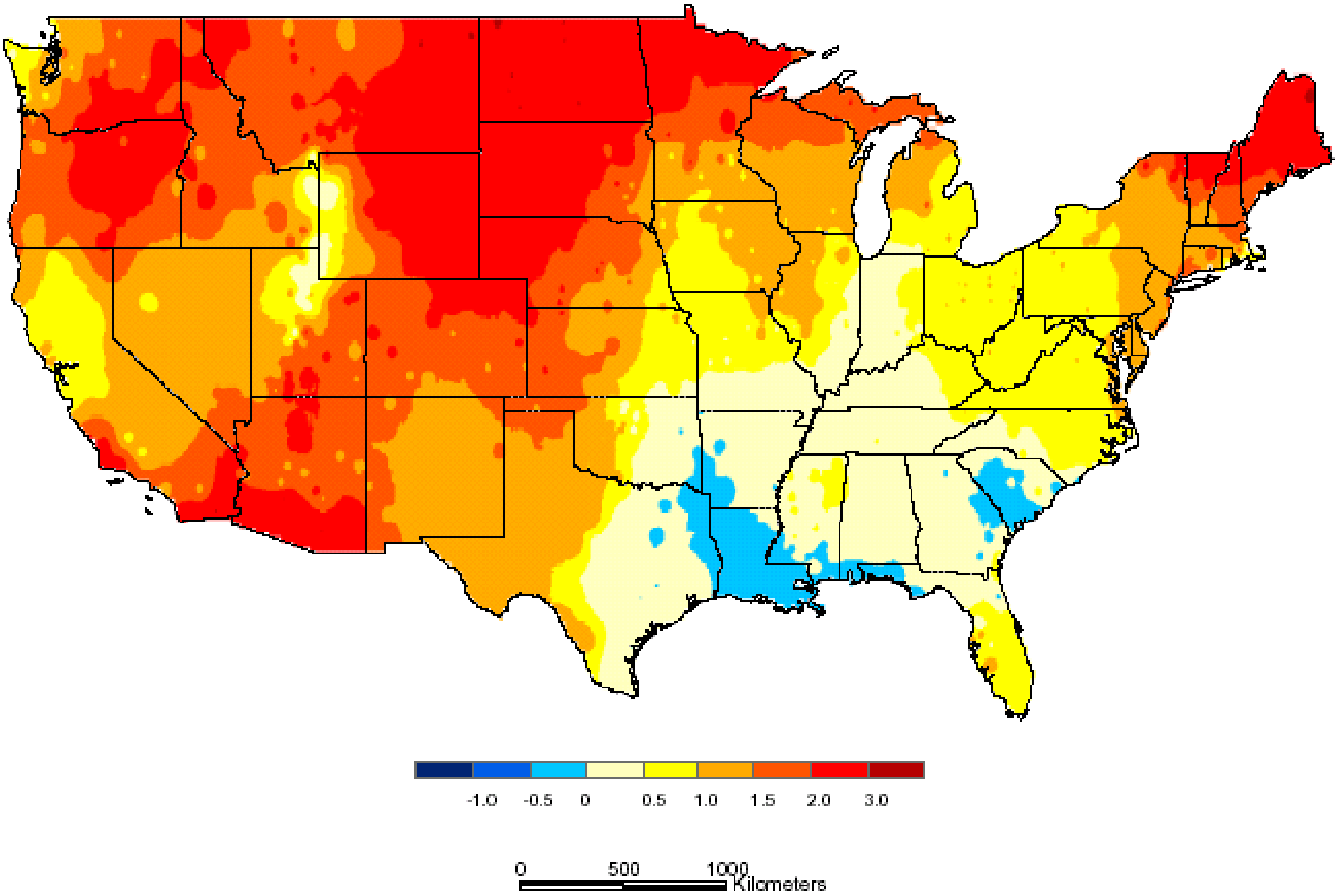

Figure 4.5: Smoothed Winter Trends

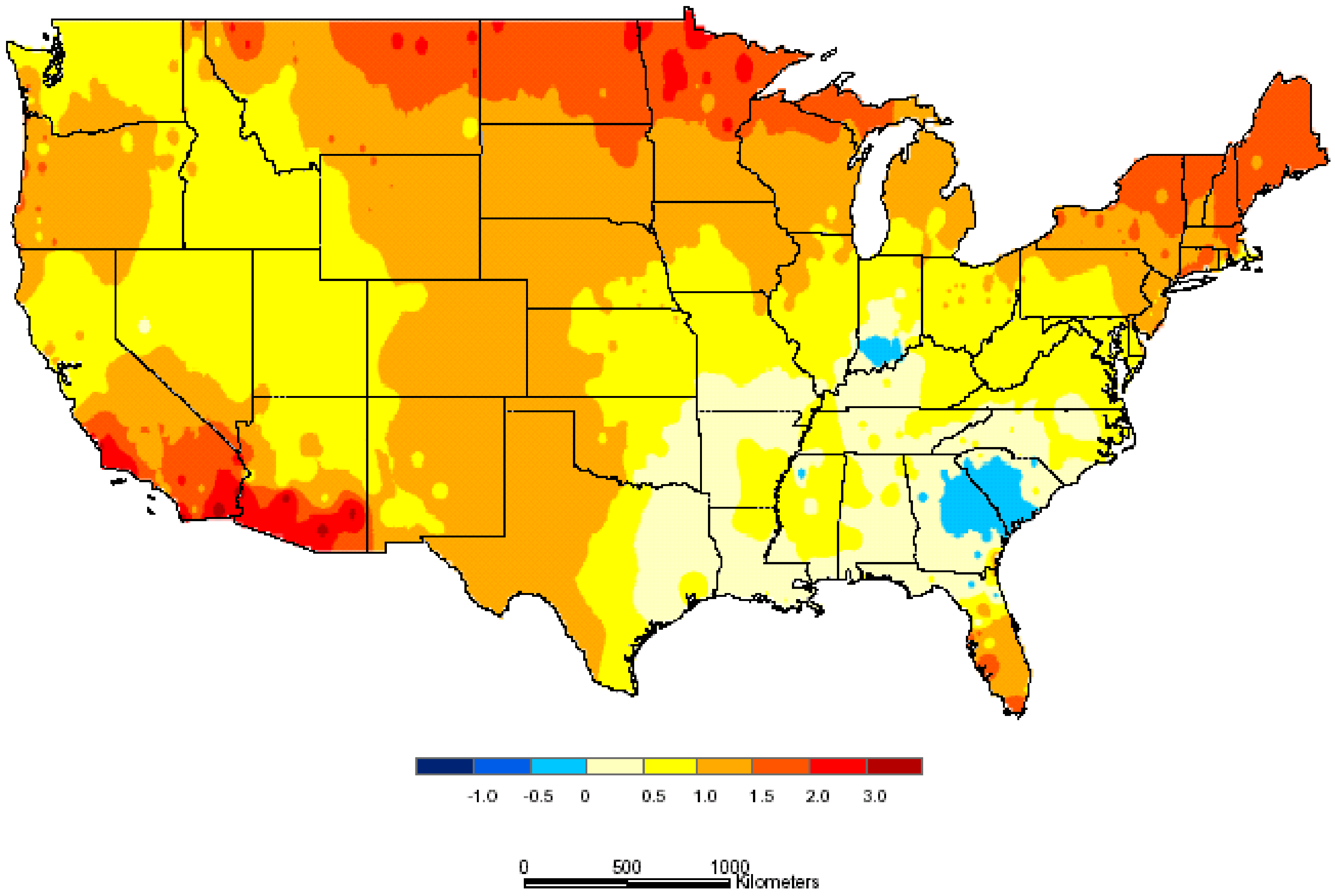

Figure 4.6: Smoothed Spring Trends

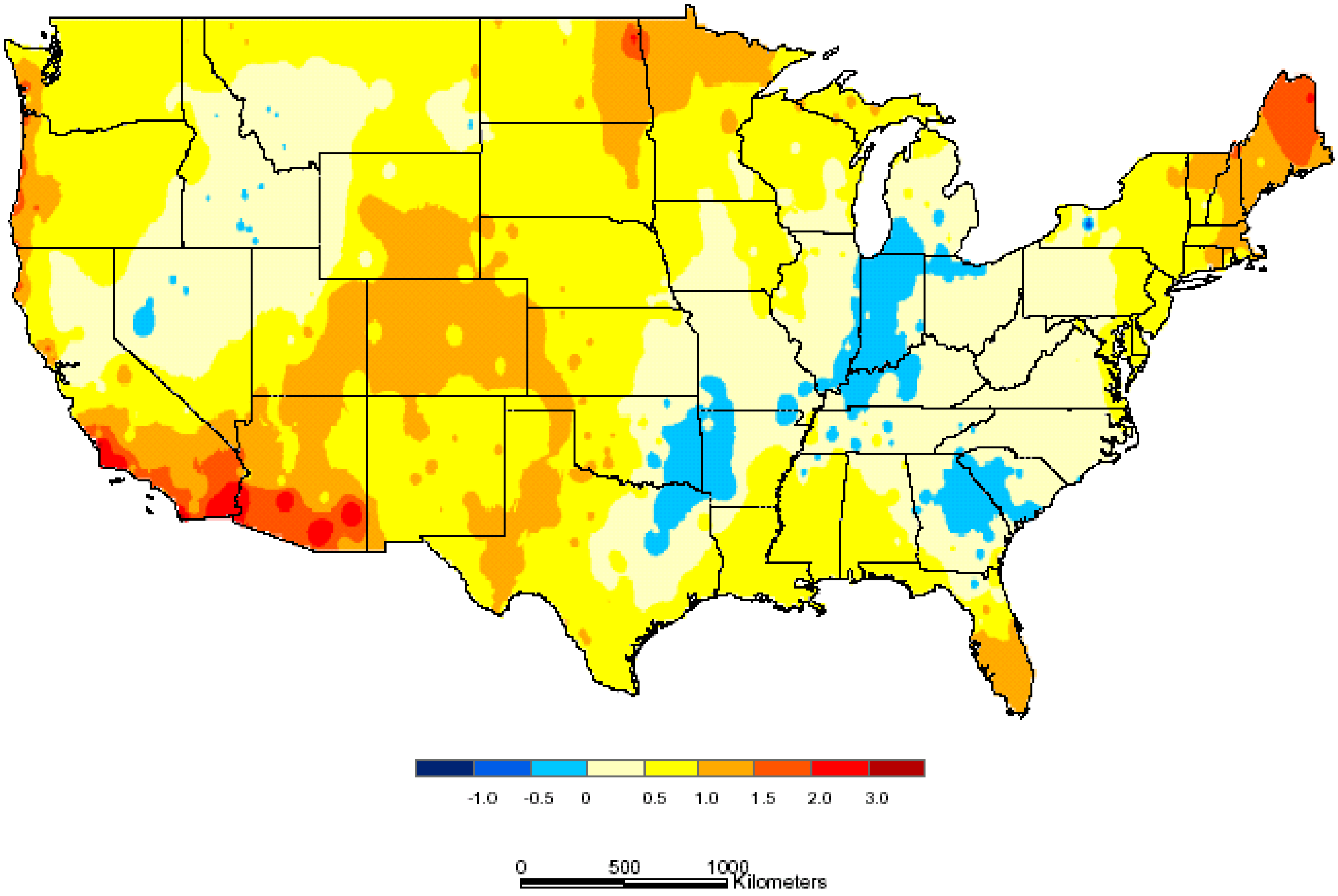

Figure 4.7: Smoothed Summer Trends

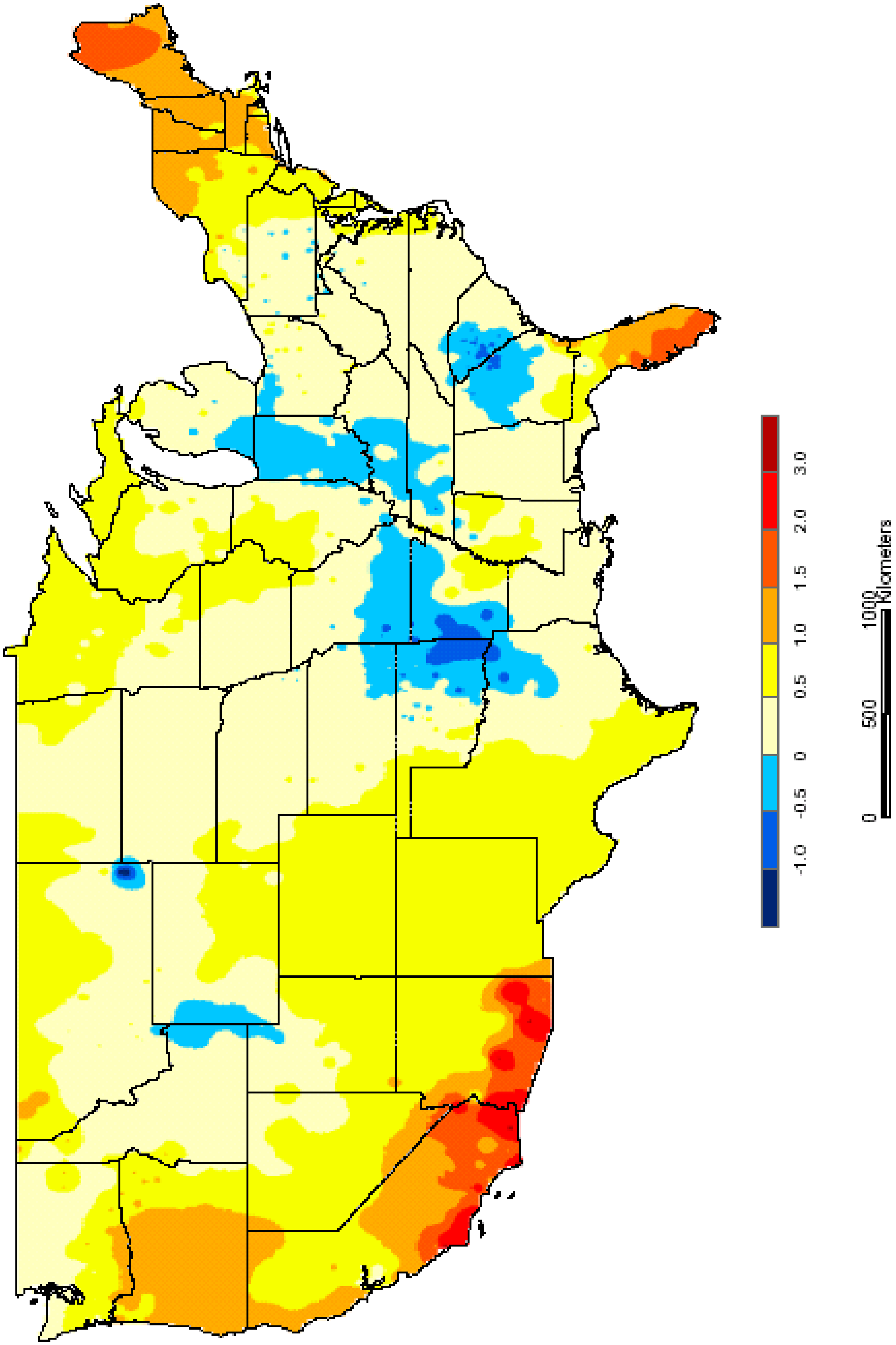

Figure 4.8: Smoothed Fall Trends

Chapter 5

Future Work

5.1 Simple Linear Regression in periodic settings

As we have seen, the simple linear regression

$$X_t = \mu + \alpha t + \epsilon_t \tag{5.1.1}$$

is useful in many situations with correlated $\{\epsilon_t\}$.

The OLS estimator of the trend α, denoted by $\hat{\alpha}_{\text{OLS}}$, is unbiased for any mean zero $\{\epsilon_t\}$ and has the simplistic explicit form

$$\hat{\alpha}_{\text{OLS}} = \frac{\sum_{t=1}^{n}(t-\bar{t})(X_t-\bar{X})}{\sum_{t=1}^{n}(t-\bar{t})^2},$$

where $\bar{t} = (n+1)/2$ and $\bar{X} = n^{-1}\sum_{t=1}^{n} X_t$ are the time and observation averages. Lee and Lund (2002) derive explicit expressions for $\text{Var}(\hat{\alpha}_{\text{OLS}})$ under autocorrelation structures commonly encountered in time series practice.

It is well known that the OLS estimator of α does not have the smallest variance among all unbiased linear estimates unless $\{\epsilon_t\}$ is white noise (uncorrelated with a constant variance). Specifically, a generalized least squares estimator of α (also called a weighted least squares or BLUE estimate), denoted by $\hat{\alpha}_{\text{GLS}}$, has a smaller variance than $\hat{\alpha}_{\text{OLS}}$. However, as has been pointed out by Grenander (1954), if $\{\epsilon_t\}$ is stationary and has a spectral density that is strictly positive at all frequencies, then

$$\lim_{n\to\infty} \frac{\text{Var}(\hat{\alpha}_{\text{OLS}})}{\text{Var}(\hat{\alpha}_{\text{GLS}})} = 1. \tag{5.1.2}$$

This justifies use of OLS estimators in lieu of generalized least squares estimators in substantial generality.

The analogy of (5.1.1) in a periodic setting is

$$X_{mT+\nu} = \mu_\nu + \alpha_\nu(mT+\nu) + \epsilon_{mT+\nu}, \tag{5.1.3}$$

where $\{\epsilon_t\}$ is a mean zero periodic time series with known period T. The OLS estimator of α_ν is

$$\hat{\alpha}_\nu = \frac{\sum_{m=0}^{d-1}(X_{mT+\nu} - \bar{X}_\nu)(mT+\nu-\bar{t}_\nu)}{\sum_{m=0}^{d-1}(mT+\nu-\bar{t}_\nu)^2}, \tag{5.1.4}$$

where $\bar{X}_\nu = d^{-1}\sum_{m=0}^{d-1} X_{mT+\nu}$ and $\bar{t}_\nu = \sum_{m=0}^{d-1}(mT+\nu)$ are the observation and time averages. Here, $d = \lfloor n/T \rfloor$ is the total observed number of cycles of data. We take d as an integer to avoid trite work.

As mentioned in Chapter 3, we question the optimality of OLS trend estimators in a periodic environment. Whereas we do not believe that the OLS estimators in (5.1.4) are radically inefficient, we do not believe they are asymptotically most efficient either. The tradeoff needs to be understood and quantified. Future work will study this issue.

5.2 Detection of undocumented changepoints

This dissertation focuses on a simple linear regression with multiple changepoints when the changepoint times are known. However, not all changepoint times are documented in practice.

To test a null hypothesis of no changepoint in (3.1.1) against an alternative of one or more changepoints, we consider a variant of (3.1.1) that allows for a single changepoint at an unknown time c:

$$X_t = \begin{cases} \mu_1 + \alpha_1 t + \epsilon_t, & 1 \le t \le c \\ \mu_2 + \alpha_2 t + \epsilon_t, & c < t \le n \end{cases}. \tag{5.2.1}$$

When $\{\epsilon_t\}$ is mean zero independent random error with constant variance, this two-phase linear regression model has been studied by many authors (cf. Hinkley 1969 and 1971b; Solow 1987; Easterling and Peterson 1995; Vincent 1998; Lund and Reeves 2002).

For a fixed $c \in \{1, 2, \ldots, n\}$, let $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\hat{\mu}_1$, and $\hat{\mu}_2$ be OLS estimates of parameters in (5.2.1). Under the null hypothesis of no changepoint, a regression F-statistic

$$F_c = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/2}{SSE_{\text{full}}/(n-4)} \tag{5.2.2}$$

should be small for each c when there is no changepoint. If a changepoint occurred at time c, F_c, and hence $F_{\max} = \max_{1 \le c \le n} F_c$, should be statistically large. Lund and Reeves (2002) list null hypothesis (no changepoint) percentiles and point out that Hinkley's (1969, 1971b) claim that the null hypothesis distribution of $F_{\max}$ was, under the constraint that the two regression lines meet at time c, approximately an $F_{3,n-4}$ distribution is incorrect and leads to an overestimation of undocumented changepoints.

However, the theory and methods in Lund and Reeves (2002) need to be extended to autocorrelated and periodic settings. We propose to derive a mathematical theory for the asymptotic distribution of $F_{\max}$ as $n \to \infty$ so as to obviate the need for simulation in each separate situation encountered.

5.3 References

[1] Easterling, D.R. and Peterson, T. (1995). A new method for detecting undocumented discontinuities in climatological time series, *International Journal of Climatology*, **15**, 369-377.

[2] Grenander, U. (1954). On the estimation of regression coefficients in the case of autocorrelated disturbances, *Annals of Mathematical Statistics*, **25**, 252–272.

[3] Hinkley, D.V. (1969). Inference about the intersection in two-phase regression, *Biometrika*, **56**, 495-504.

[4] Hinkley, D.V. (1971b). Inference in two-phase regression, *Journal of the American Statistical Association*, **66**, 736-743.

[5] Lund, R.B. and Reeves, J. (2002). Detection of undocumented changepoints: A revision of the two-phase regression model, *Journal of Climate*, **15**, 2547-2554.

[6] Solow, A.R. (1987). Testing for climate change: an application of the two-phase regression model, *Journal of Climate and Applied Meteorology*, **26**, 1401-1405.

[7] Vincent, L.A. (1998). A technique for the identification of inhomogeneities in Canadian temperature series, *Journal of Climate*, **11**, 1094-1104.