Mathematical and Statistical Modelling for Air Quality Management

by

Jun Bai

A thesis submitted to the
Australian National University
for the degree of Doctor of Philosophy

April 1990

PREFACE

The following are publications prepared during the period of this course and undertaken as joint research with Anthony J. Jakeman, Michael McAleer, John A. Taylor of ANU, and Gerald Miles of the National Capital Planning Authority.

(1) **Bai, J. and Taylor, J.A.** (1986), Estimation of the parameters and upper percentiles of statistical distributions applicable to air quality data, part I: The gamma distribution. CRES Working Paper, No 1986/39, Australian National University, pp. 131.

(2) **Taylor, J.A., Jakeman, A.J. and Bai, J.**(1986), A Monte Carlo study of estimation of the upper percentiles of the three-parameter gamma distribution using methods of moments and maximum likelihood. CRES Working Paper, No 1986/34, Australian National University, pp. 16.

(3) **Jakeman A.J., Bai, J. and Taylor, J.A.** (1988), On the variability of the wind speed exponent in urban air pollution models. Atmospheric Environment, 22 (9), 2013-2019.

(4) **Bai, J., Jakeman A.J. and Taylor, J.A.** (1988), Statistical distribution modelling: Function, methods and application to air quality management. Mathematics and Computers in Simulation, 30, 3-9.

(5) **Taylor, J.A., Jakeman, A.J. and Bai, J.** (1988), Modelling for air quality management. 12th World IMACS Congress on Scientific Computation, 18-22.

(6) **Miles, G.H., Jakeman, A.J. and Bai, J.** (1989), A method for predicting the future extremes of urban air pollution from vehicle emission, meteorology and historical concentrations. Proceedings of the Simulation Society of Australia Conference, 418-423.

(7) **Bai, J., Jakeman, A.J. and McAleer, M.** (1989), A new approach to maximum likelihood estimation of the three-parameter gamma and Weibull distributions. Working Paper in Economics and Econometrics No. 191, Australian National University, November 1989, pp. 26. (also submitted to The Australian Journal of Statistics)

(8) **Bai, J., Jakeman, A.J. and McAleer, M.** (1989), The estimating the percentiles of some misspecified non-nested distributions. Working Papers in Economics and Econometrics No. 193, Australia National University, December 1989. (also submitted to Communications in Statistics).

(9) **Bai, J., Jakeman A.J. and Taylor, J.A.** (1990), Percentile estimation of the three-parameter gamma and lognormal distribution: methods of moments versus maximum likelihood. Mathematics and Computers in Simulation, 32, 164-169.

(10) **Bai, J., Jakeman, A.J. and McAleer, M.** (1990), The effects of misspecification in estimating the percentiles of some two- and three-parameter distributions. Mathematics and Computers in Simulation, 32, 194-199.

(11) **Bai, J., Jakeman, A.J. and McAleer, M.** (1990), Discrimination between nested two- and three-parameter distributions: An application to models of air pollution. Working Paper in Economics and Econometrics No. 197, Australian National University, March 1990, pp. 23. (also submitted to Technometrics)

(12) **Bai, J., Jakeman, A.J. and McAleer M.** (1990), Discrimination procedure for fitting nested and non-nested distributions to environmental quality data. Working Papers in Economics and Econometrics No. 200, Australian National University, March 1990, pp. 58. (also submitted to Environmentrics)

(13) **Jakeman, A.J., Bai, J. and Miles, G.H.** (1990), Prediction of non-stationary extremes of one-hour average urban CO concentrations. (submitted to Atmospheric Environment)

The text of these papers has been closely followed in Chapters 3 to 12 of this thesis. Unless otherwise acknowledged in the text, the remainder of this thesis represents the original research of the author.

Jun Bai

April 1990

ACKNOWLEDGEMENTS

I would like to thank my supervisory panel of Anthony J. Jakeman, Michael McAleer, Michael Hutchinson and Rodney Simpson for their advice and general guidance. I am particularly indebted to Tony Jakeman and Michael McAleer for their direction, encouragement, and kindly giving of their time to the development of the study and formation of this thesis. Special thanks also go to John Taylor for his kindly advice on technical issues.

I am also grateful to the following for their assistance: Gerald Miles of the National Capital Planning Authority, and G. Moynihan of the Canberra branch of the National Climate Centre, Australian Bureau of Meteorology. They provided the air quality data and the meteorological data recorded in Canberra, respectively.

I have received welcome assistance from Tony Bayes in the management and organisation of the data sets used in the thesis, from Mark Greenaway of CRES for equipment operation and production of graphics, and from Kathy Handle of the Computer Services Center for advice on available computer software and techniques.

Many thanks to Henry Nix, Director of CRES, for his profound concern, and encouragement throughout my candidature. Also thanks to my friends and colleagues at CRES who provided an interesting and stimulating environment within which this research was undertaken.

I am very grateful to Ian Taylor and Wendy Chen for careful proof-reading of the final draft of this thesis.

Thanks go to my family and friends, especially to Alejandra Benedictos for her support during my period of intensive research.

This work would not have been possible without the assistance of a Commonwealth Postgraduate Research Award.

ABSTRACT

Air quality management requires the development of relationships between the frequency distribution of ambient pollutant concentrations at sites of concern and emissions, meteorological and other forcing variables. In this thesis, mathematical models are devised to achieve a range of goals within this basic objective. In Australia, much recent effort has been devoted to both statistical and hybrid deterministic-statistical distribution modelling approaches which have extensive potential but previously have been given only limited attention in the literature.

These are the approaches investigated in this thesis, although it should be emphasised that the so-called statistical models should more appropriately be labelled frequency or probability distribution models as they are parametric forms of probability density functions. There are six parameterisations used in the thesis and these are the two- and three-parameter versions of the gamma, Weibull and lognormal distributions.

Basically, each of the two approaches treated has its own assumptions and the choice of approach depends on the goal, the validity of the assumptions and the data available. Each approach must also invoke a methodological infrastructure. Generally, this involves the use of adequate techniques for identification of an appropriate parametric form to represent a given pollutant data set and for estimation of the associated parametric values and the ensuing errors.

This thesis contributes to the development of the requisite techniques and illustrates their application. It includes the construction of a new generalised information criterion (GIC) for discriminating among candidate parametric forms that may be nested or non-nested. The thesis develops a new algorithmic approach to maximum likelihood estimation (ML) which complements the general or traditional ML approach and can provide more computationally efficient parameter estimates. The algorithms allow maximisation of the likelihood function in an extended parameter space that is important in air quality applications where the traditional approach does not apply. The thesis also demonstrates that simple computational methods of estimation such

as methods of moments can suffice if the objective is merely to summarise the data and/or to allow high variance estimates. It shows how to construct error models that allow the calculation of the minimum errors in percentiles to be expected when fitting samples from different probability distributions. The thesis also evaluates errors of misspecification which arise when the wrong parametric form of distribution is selected. All of these tools are then combined to illustrate the practical use of a comprehensive procedure to identify a suitable parametric form which represents a given pollutant over the years at single and multiple sites. It should also be mentioned that much of the new technology developed for the statistical aspects of the thesis can also be used for application of extreme value theory in statistics since the same identification and estimation tools are required.

The last two chapters of the thesis involve the use of deterministic models. Chapter 11 shows how to combine a statistical distribution model of short-term carbon monoxide concentrations in an urban area with a deterministic model that predicts average pollutant concentrations from vehicle patterns and basic meteorology. The application shows how the deterministic and statistical components can be successfully combined when the attributes of predictive interest, that is the annual maxima of 1-hour carbon monoxide concentrations, represent samples from a long-term non-stationary process. Chapter 12 investigates the inadequacy of the standard box model of urban diffusion and illustrates the need for an additional connective parameter as exponent on the wind speed variable. The resulting generalised box model would then be suitable for hybrid modelling since it could be capable of reasonable predictions of long-term mean concentrations. It would be complementary to the hybrid model developed in the previous chapter since it would require time series of wind measurements rather than seasonal or annual means of a wider range of meteorological variables.

Contents

Part I

INTRODUCTION

Chapter 1

A Systems Approach to Air Quality Management

1.1 Introduction

The atmosphere is an important shared resource whose quality needs protection. The deterioration of atmospheric environmental quality, related to changes in the chemical, physical and biological nature of air brought about by industrial, agricultural and social activities, has become a threat to many plant and animal communities, and to the human race. A systems approach to atmospheric environment quality control is urgently required to provide an effective means of preserving for future generations some semblance of the biological order of the world and to improve the deteriorating standard of urban public health (Metcalf and Pitts, 1972).

Appendix 1 addresses some basic issues raised in air quality management. A brief historical review in section A1.1 provides an overall picture of past air pollution and reveals the major problems. Air pollution problems are discussed separately, in terms of the different scales of air pollution dispersion, local or regional, and continental or global, although the latter is not a major topic in this thesis. From the review of air pollution history it is clear that direct energy use, industry and agriculture are three dominant contributors to air pollutant concentrations caused by human activities. They are strongly responsible for pollution of the atmospheric environment. Among global pollution problems, the greenhouse effect and depletion of the ozone layer, which

have become of more concern in recent years, are discussed.

Appendix 1 also introduces the important concepts which underlie any study of air quality. These concepts include definitions for the terms atmosphere, clean air and global background concentrations. In order to better understand air quality impact assessment, most air pollutants and their characteristics are summarised in the appendix, and some of them will be chosen for modelling in this thesis.

In recent decades, there have been great improvements in control technology, in understanding of atmospheric processes and interactions among air-borne pollutants. There have also been accompanying developments in administrative and legislative instruments for regulation and abatement of those pollutants. To develop management of our atmospheric resources further, more scientific information is needed. For example, de Nevers et al. (1977) have argued that a complete and closed information cycle might be required. The information would include emission data, air quality monitoring and air quality modelling results. If such associated analytical tools are widely available, their use may strongly influence decisions on any proposed or existing projects. Unfortunately, there is considerable uncertainty at present about all three factors in this information cycle leading to model results of limited accuracy and validity. Therefore, each of the factors must be improved in order to develop a predictive scheme which can be used with confidence.

As a crucial aspect in the information cycle, air quality modelling has attracted a great deal of attention. Air pollution models may be constructed for diverse purposes. They can be used to describe the physical and chemical nature of air pollution processes, and to assist our scientific understanding of complex atmospheric behaviour. Alternatively, modelling techniques may be developed to provide information as aids in air quality management, and to be used in decision making for city, regional and other project planning. Generally, the former approach has tended to involve considerable mathematically complexity. For the latter, models should be relatively simple to use, and be constructed within the context of the avaliable data and the required form of

model output.

Atmospheric systems are, in the terminology of Young (1978), 'badly defined' in that the relationships that describe their behaviour are complex and not easily amenable to exploration through planned experimentation. Theoretically in such systems both causal and non-causal relations exist. For example, many deterministic models attempt to present an ideal pattern of the causal linkage among dependent and independent variables, such as in the basic equations of molecular diffusion. However, inherent in these systems are statistical features (e.g. Pasquill and Smith, 1983). Atmospheric behaviour is essentially the consequence of an hybrid deterministic-statistical system. Therefore theory and techniques for both deterministic and statistical modelling approaches are required. It would be useful to develop a specific system methodology which can be applied to such a hybrid system.

The aim of this thesis is to develop tools for the construction of simple but effective models to predict ambient pollutants which reduce the difficulties arising from limited meteorological information, sparse pollutant monitoring data and the stochastic nature of turbulent diffusion processes. The models developed in this thesis have certain desirable properties; they contain as few parameters as are necessary for satisfactory performance and the output of the models allows direct comparison with air quality standards. The models provide an indication of the uncertainty associated with model prediction. In addition, the modelling tools developed here can be widely used in many other areas such as hydrology, water pollution management, reliability and life-testing (Yevjevich, 1972 and Bain, 1978).

Systematic methods for identification of the best parameterization of probability distribution models, estimation of their associated parameters and ensuing errors have received little attention. This places limitations on the utility of probability distribution models and hybrid models. In addition, hybrid models have not been constructed with the aim of minimising problems imposed by autocorrelation and non-stationarity in pollutant concentration data. This thesis aims to provide some useful concepts,

methods and applications to address these requirements.

1.2 Air Quality Management

"Air quality management is the regulation of the amount, location, and time of pollutant emissions to achieve some clearly defined set of ambient air quality criteria. It includes the evaluation of various sets of emission control schedules to determine consequences to air quality and the formulation of alternative emission control schedules to meet air quality goals subject to some other constraint, e.g. technology feasibility or minimum cost" (de Nevers et al., 1977).

The definition implies that the following data and knowledge be available: a statement of air quality criteria, goals or standards; estimates of pollutant emissions; observations of ambient air pollutant concentrations; models for atmospheric dispersion; and models for characterization of the frequency distribution of air pollutant concentrations.

1.3 Air Quality Standards

Ambient air quality standards are used in many countries to protect public health and welfare. In some countries, such as the United States, they are cast in legislation, while in others, such as Australia and the United Kingdom, they are used more as guidelines.

In general, air quality goals or standards tend to be written in two ways:

a. Air quality standards can be prescribed as long-term mean levels that ideally are not to be exceeded; or

b. Air quality standards can be prescribed as short-term levels not to be exceeded or only to be exceeded a small percentage of time in a given time period.

These air quality standards involve three factors: the period of prescription (usually annual), the permitted time at a specified level and the quantity of air pollutant con-

centration. In many air quality standards the latter two factors are covered i.e. large dose in a short interval, or repeated small doses over a long period. These two types of air quality standards have arisen from evaluation of the exposure-response relationship, since air pollutants may have both long-term and short-term effects. Accordingly, air quality standards are usually classified into primary standards and secondary standards. Primary standards are concerned with the accepted maximum level of a pollutant whereas secondary standards are concerned with the mean of concentrations. Both are intended to protect public welfare.

On the other hand, it should be noted that most current air quality standards are effectively stated in terms of the frequency with which a specified concentration may be exceeded for a given averaging, or sample time. For example, the United States Environmental Protection Agency (EPA) has a primary short-term standard for sulphur dioxide which is 14 parts per hundred million (pphm) for a 24-hour average (sampling interval) and this figure must not be exceeded more than once per year. Therefore, an important goal in the evaluation of compliance with air quality standards is the estimation of the upper percentiles of the frequency distribution of pollutant distribution. Important quantities are the maximum and second maximum, and sometimes the 98 percentile of the annual frequency distribution.

Note also that, in addition to interest in the upper percentiles of the air pollutant concentrations, there is a need to accurately estimate the entire range of pollutant concentrations. Such information may be applicable for the integration of damages sustained (Jakeman and Simpson, 1987). For example, materials may deteriorate gradually as a result of low concentrations occurring with high frequency. Lower concentrations may also become more problematical where a synergistic combination of pollutants occurs. Hence, if the frequency distribution of pollutant concentration within an area of interest were predicted, then emission control strategies could be assessed to enable air quality standards to be met and damages to be minimised, subject to cost-benefit analysis.

Air quality standards can differ by region, state or country, as shown in Table 1.1. For example, developing countries, such as China, generally have a lower level of air quality. The pollutants listed in Table 1.1 are those of most concern in this thesis and will be compared with the output of air quality modelling exercises in Chapter 11.

1.4 Thesis Outline

The thesis can be considered in four parts. Part I contains this introductory chapter and a review chapter. Part II contains chapters on parametric estimation methods for probability distributions, associated percentile errors and empirical models of those errors. In Part III, the focus is on methods of discrimination among distributions and errors in misspecifying a distribution. Part IV applies many of the tools developed and results obtained to problems in air quality management. It also contains the concluding chapter.

The remaining part of Part I is Chapter 2 which reviews developments in deterministic, statistical and hybrid deterministic-statistical distribution modelling. A variety of models, their functions, advantages and limitations are also assessed.

Chapters 3 to 9 constitute Parts II and III. They present the modelling techniques which have been developed in the thesis for the analysis and prediction of air quality. Chapter 3 is concerned with parameter estimation for the following frequency or probability distributions: the two- and three-parameter gamma, Weibull and lognormal distributions. The general or traditional maximum likelihood estimation method is re-examined. To overcome the problem of non-existence of a solution for certain parent parameter values when the traditional formulation is applied, a new approach to maximum likelihood estimation is proposed. This method is accurate and computationally efficient, and is particularly useful for fitting air pollution data where the parent parameter values are such that the traditional formulation fails.

In Chapter 4, a comparison of different methods for percentile estimation, especially the upper percentiles, is discussed for all six distributions. When fitting the observed or

sample upper percentiles it is seen that the method of moments can be more accurate than the method of maximum likelihood. Using Monte Carlo experiments, the bias (BIAS) and relative root-mean-square error (RRMSE) are calculated for each method in order to calculate the theoretical and empirical departure from the underlying distribution. In Chapter 5 response-surface techniques are adopted to develop some simple empirical models for predicting RRMSE. Based on extensive experiments, empirical models for the three-parameter gamma, Weibull and lognormal distributions are constructed. They can be easily used in air quality applications to compute the variability of percentile estimates against type of distribution, parameter values and sample size. This chapter also represents an attempt to develop the response-surface techniques for data from computer simulations.

Chapter 6 is the first chapter in Part III. It considers the problem of discrimination among nested distributions. Many well-known hypothesis tests and information criteria are considered for discriminating between the two- and three-parameter gamma, Weibull and lognormal distributions. To re-examine their performance both simulation experiments and observational data are used. The results from the simulations show that the performance of the tests and information criteria depend on the type of distribution and the range of parameters. The intended use of the distribution is a very important consideration for selecting an appropriate criterion. Chapter 7 assesses the effects of misspecification in estimating the percentiles of the two- and three- parameter nested distributions, where the emphasis is placed on the upper percentiles. Conventional wisdom regarding underfitting or overfitting may not be a good guide to selecting a distribution. The consequence of such misspecification could cause substantially larger errors.

Chapter 8 considers discrimination among non-nested distributions. Some existing problems with the way discrimination is approached in the literature are addressed and the standardized procedure, which tests the null hypothesis against an alternative with some well-known discrimination criteria, is re-examined. The emphasis is to discriminate among a set of distributions. For example, discrimination is required

among six distributions used in this thesis. In order to complement the weakness of the relevance of existing criteria for air quality management applications, a generalized information criterion (GIC) is constructed. As in chapter 7, Monte Carlo experiments are employed to examine their performance over a range of parameters and to assess the effects of misspecification in estimating upper percentiles of non-nested distributions.

Chapters 10 to Chapter 12 constitute the main body of Part IV where the focus is on applying the modelling techniques. In Chapter 10 the estimation and discrimination techniques are used on air pollution data collected in Melbourne, Australia. The purpose of the investigation is to discriminate among the appropriate distributions and estimate the parameters of the distributions. For different averaging times, pollutants may change distributional form.

Chapter 11 describes a hybrid method to determine the seasonal extremes of 1-hour average CO concentrations from vehicle patterns and emissions, basic meteorological measurements and historical records of ambient concentrations. The method links the output of a deterministic Gaussian plume line source model with knowledge of a suitable parametric form of the probability density function (pdf) of peak 1-hour CO concentrations. The deterministic model requires only average emission and meteorological data as input, although the approach outlined can be extended to include more complex deterministic models with more detailed dynamic input information. Knowledge of the pdf of ambient concentrations is gained from past data by applying goodness-of-fit tests based upon maximum likelihood estimation and its accuracy is assessed by examining prediction performance for the extremes of interest. Problems of autocorrelation and non-stationarity in the distribution of pollutant concentrations are minimised by restricting attention to the winter season and to the evening peak concentration. The method is used to predict 1-hour maxima of CO concentrations for winter seasons in Canberra, Australia, although it applies to other extremes at other time averages, such as 8-hour averages, and to other pollutants dispersed predominantly from mobile sources.

Chapter 12 investigates the seasonal variability in the wind speed exponent of the generalised box model for 24-hour average total suspended particulates (TSP), β-scattering and CO data collected in Canberra, Australia. The results confirm that it may be erroneous to assume that the exponent of the horizontal wind speed in the box model is equal to -1, at least in urban airsheds. The sensitivity of predicted concentration by a box model is high with respect to the exponent value. Therefore, calibration for just a single season is likely to lead to much better predictions in future years than simply assuming an exponent value of -1, as is the present practice. This chapter also demonstrates that more sophisticated, but algorithmically straightforward, techniques for parameter estimation can be used. These are based on flexible but simple parameterizations of the parameter changes over time. This chapter also indicates, under Gauss-Markov process assumptions, that Kalman filtering and smoothing algorithms are very useful in handling the data sets related to the box model formulation. Such techniques can be easily adopted when constructing a hybrid model to predict air pollutant concentrations.

TABLE 1.1
Ambient air quality standards in some countries

Pollutant	Average Time	United States (Primary)	Germany (West)	Canada	Japan	Sweden	China	Australia (Victoria)
Carbon Monoxide CO	24-h				10		60(mg/m^3)	
	8-h	9(b)	10(mg/m^3)	13(a)	20			20(a)
	1-h	35(b)	0(mg/m^3)	30(a)				60(a)
Nitrogen Dioxide NO_2	Year	0.05		0.05	0.01			
	24-h		0.05	0.16(d)	0.02			0.15(a)
	1-h		1(mg/m^3)	0.30(a)				0.25(a)
Ozone O_3	1-h	0.12	0.06				0.2(mg/m^3	
Sulphur Dioxide SO_2	Year	0.03						
	24-h	0.14(b)	0.14(mg/m^3)	0.06(a)	0.04	0.10	025(mg/m^3)	0.06(a)
	1-h		0.40(mg/m^3)	0.34(a)	0.10	0.25		0.34(a)
Total Suspended Particulate TSP	Year	0.075(mg/m^3)						
	24-h	0.26(mg/m^3)	0.10(mg/m^3)	0.10(mg/m^3)		0.50(mg/m^3)		
	1-h		0.48(mg/m^3)(b)	0.20(mg/m^3)	0.10(mg/m^3)			

Note: The unit of pollutants is 'ppm' unless otherwise indicated in the table. (a) indicates the detrimental level; (b) not to be exceeded more than once per year; (c) the annual mean; (d) the maximum tolerable value.
Source: SPCC Quarterly Air Quality Monitoring Report, Vol 8, No. 2, 1982; Newill (1976), Simpson (1989).

Chapter 2

Review of Air Quality Modelling

2.1 Introduction

Mathematical modelling for air pollution management has progressed considerably in recent years. Models are widely available for the prediction of both short and long-term mean ambient pollutant concentrations. Many models are applied to assess the impact on the atmospheric environment arising from both new and existing industries, to calculate urban pollution levels and global background concentrations.

This chapter provides a review of the major modelling approaches available for air quality management of ambient concentrations. Three key approaches to the modelling of air pollutant concentrations in the atmosphere are examined, namely deterministic, statistical and hybrid. Emphasis is placed on pollutants with inert, or relatively inert, properties. Attention is given to assessing the performance of each modelling approach and comparing their advantages and limitations.

The term 'deterministic' refers to models formulated on a physical basis and is concerned with mechanical outcomes. They can be constructed according to hypothesized causality among driving factors and defined by one or several mathematical functions. Such models attempt to provide a description and explanation of the dispersion process in the atmosphere.

The term 'statistical' refers to a subset of stochastic models that are constructed on a non-causal or phenomenological basis. These models are calibrated by statistical

methods and the output is the probability of pollutant concentrations. The statistical models addressed in the thesis are parametric forms of probability distribution functions, including the two- and three-parameter gamma, Weibull and lognormal distributions.

The term 'hybrid' refers to models comprising both deterministic and statistical components. Such models attempt to combine the best features of each approach and can be used to predict the frequency distribution of pollutant concentrations for direct comparison with air quality criteria. Hybrid models can provide a measure of uncertainty associated with model predictions. Since they predict the frequency distribution of air pollutant concentrations, such models allow the development of strategies for emission control with respect to ambient mean levels and extreme events.

2.2 Deterministic Models for Air Pollution Concentrations

Deterministic modelling is the traditional approach applied to the prediction of air pollutant concentrations. Initial contributions to modelling atmospheric dispersion were those by Taylor (1915), Scrase (1930), Sutton (1932) and Giblett et al. (see Pasquill and Smith (1983)). Since then, the number of deterministic models developed has grown rapidly, and the literature contains a plethora of deterministic approaches for a wide range of physical circumstances.

There have been many reviews of deterministic models, including Lamb and Seinfeld (1975), Eschenroeder (1975), Johnson et al. (1976), Hanna (1978), Drake et al. (1979), Turner (1979), Simpson and Hanna (1981), Hanna (1982), and Geraghty and Ricci (1984). The task of the review here is to examine critically the existing literature, elicit some pertinent conclusions and clarify useful future directions. The overview of deterministic models also provides a basis for selecting an appropriate model for a given application, and presents evidence to demonstrate the need for other modelling approaches. Only major references are given as the reader can refer to the publications

cited in the reviews noted above.

2.3 Deterministic Models for Air Pollution Concentrations

Deterministic modelling is a traditional approach applied to the prediction of air pollutant concentrations. Initial contributions to modelling atmospheric dispersion were those by Taylor (1915), Scrase (1930), Sutton (1932) and Giblett et al. (see Pasquill and Smith (1983)). Since then, the number of deterministic models developed has grown rapidly, and the literature contains a plethora of deterministic approaches for a wide range of physical circumstances.

There have been many reviews of deterministic models. These include Lamb and Seinfeld (1975), Eschenroeder (1975), Johnson et al. (1976), Hanna (1978), Drake et al. (1979), Turner (1979), Simpson and Hanna (1981), Hanna (1982) and Geraghty and Ricci (1984). The task of the review here is to critically examine the existing literature, elicit some pertinent conclusions and clarify useful future directions. The overview of deterministic models also provides a basis for selecting an appropriate model for a given application, and presents evidence to demonstrate the need for other modelling approaches. Only major references are given as the reader can refer to the publications cited in the reviews noted above.

2.3.1 Gaussian Plume Models

In the literature, the Gaussian plume model is the most well-known and widely applied air quality model for describing the dispersion of most primary pollutants. The Gaussian plume model was first applied to point sources assuming homogeneous and stationary turbulent atmospheric conditions. The name of the model derives from the description of the shape of the dispersion from a source which, on average, was apparently found to be a close approximation to the Gaussian form in both the horizontal and vertical directions (Pasquill, 1974). The Gaussian plume model has been extended

to describe the dispersion of pollutants from line, area and urban sources (see Chock (1978) and Burt and Slater (1977)).

Gaussian plume models were originally derived from the theory of molecular diffusion and heat conduction but they can also be regarded as a special case of the general mass conservation equation. Huang (1979) has derived a generalised non-Gaussian diffusion model for a turbulent shear flow relating K-theory to statistical theory. Thus, the conventional Gaussian dispersion model can be regarded as a special case of Huang's model.

Gaussian plume models often assume the following:

1. meteorological conditions are steady, with wind speed and direction kept constant, and no inversion layer;

2. the initial concentration is assumed to be zero, emissions are constant, and the pollutant is inert;

3. there is no downwind diffusion, the diffusion coefficients in the cross-wind and vertical directions vary only with downwind distance and are constant in the diffusion domain, and there is no absorption or generation of pollutants by the ground.

The typical Gaussian plume model recommended by the United States Environmental Protection Agency (EPA) is expressed as (Turner, 1964)

$$\chi(x,y,z) = \frac{Q}{2\pi\sigma_y\sigma_z\bar{u}} exp(-\frac{y^2}{2\sigma_y^2})\{exp[-\frac{(H-Z)^2)}{2\sigma_z^2}] + exp[-\frac{(H+Z)^2)}{2\sigma_z^2}]\} \qquad (2.1)$$

where χ is concentration at the point (x,y,z), Q the steady source strength (mass emission rate), $\bar{u}$ the mean wind speed, σ_y and σ_z the standard deviations in concentration in the crosswind (y) and vertical (z) directions, respectively, and H the effective height of emission. Wind is assumed to be in the x-direction. A feature of this model is its simplicity and requirement of little data input in comparison with other models. Based on this formulation, many well-known Gaussian plume models are used and recommended by the EPA for impact assessments.

Equation (2.1) can be used directly to calculate ambient pollutant concentrations. The key procedure is the estimation of H and the σ values because even modest errors in both estimates may yield a 50% total error (Weber, 1976). Calculation of effective height H is given by

$$H = h + \Delta H \tag{2.2}$$

where h is the physical height of the source and ΔH is the plume rise.

The σ values depend on the turbulence characteristics of the flow. The most common method of estimating σ is to use the Pasquill-Gifford curves (Turner, 1970), which classify the turbulent state of the atmosphere into six categories A to F. Alternatives include the stability curves from Singer and Smith (1966), and McElroy and Pooler (1968), and the interpolation formulae of Briggs (1974).

2.3.2 K-Theory Models

A sophisticated model, based upon K-theory, has appeared widely in the literature (Hanna et al., 1982). K-theory assumes that there is similarity between atmospheric turbulence and molecular diffusion. The major physical assumption is that the turbulent flux of material is proportional to the mean concentration gradient. In the x-direction, the proportional relation can be given as

$$\overline{U'\chi'} = -K_x \frac{\partial \overline{\chi}}{\partial x}$$

where a prime indicates fluctuation about the mean.

On the basis of the above gradient assumption, the K-theory model can be expressed as

$$\frac{\partial \chi}{\partial t} + U\nabla\chi = Q + \frac{\partial}{\partial x}(K_x \frac{\partial \overline{\chi}}{\partial x}) + \frac{\partial}{\partial y}(K_y \frac{\partial \overline{\chi}}{\partial y}) + \frac{\partial}{\partial z}(K_z \frac{\partial \overline{\chi}}{\partial z}) \tag{2.3}$$

where χ is the species pollutant concentration, U the velocity, Q a source term, and K_x, K_y and K_z are the appropriate diffusivities in the x, y and z directions (see e.g.

Carras, 1989).

Generally, a K-theory model requires a numerical solution to provide predictions of the ambient concentration of a particular species as a function of position and time. However, to obtain a satisfactory result, K-theory assumes that the largest eddies responsible for plume dispersion are much smaller than the dimensions of the plume. For convective conditions, the mixing height typically varies from 500m to 2000m. The size of these largest eddies is approximately 1.5 times the mixing height. Models based on K-theory will sometimes fail to work well (Carras, 1989).

2.3.3 Box Models

Based on different assumptions, a variety of box models has been developed for prediction of air pollutant concentrations. They can be divided into single box and multi-box models, and can also be considered to apply to area emission sources, line sources, and even point sources. In addition, according to the basic assumptions of the box model approach, box models can be divided into two types. One type of box model assumes that the pollutants are unlikely to disperse as far as the inversion layer. This assumption is likely to prevail if the area considered is small and the wind speed is not too low. The other type of box model assumes that vertical dispersion is affected by the inversion layer, which usually occurs for stagnant wind conditions and large areas.

The box model can be derived from simple physical considerations as shown, for example, by Simpson and Hanna (1981). However, some box models often incorporate ideas inherent in the Gaussian plume approach. Thus, models such as the Atmospheric Turbulence and Diffusion Laboratory (ATDL) model are sometimes referred to as Gaussian models.

The Atmospheric Turbulence and Diffusion Laboratory (ATDL) Model

The popular ATDL model was proposed by Gifford (1970, 1973), Gifford and Hanna (1971, 1973), and Hanna (1971, 1973). This model is applied to urban area sources for

stable non-reacting pollutant species. Emissions are assumed to be uniform over each grid square. The essential idea is based on integration in the up-wind direction of a cross-wind infinite line-source diffusion formula, namely the simple power law

$$Z(x) = ax^b \tag{2.4}$$

which only involves vertical diffusion, where $Z(x)$ is σ_z, the standard deviation in Gaussian dispersion, x is the downwind distance, and a and b are parameters dependent on atmospheric stability. Lateral dispersion is neglected so that area sources are treated by the narrow-plume assumption. Based on these considerations, the ATDL model is described by the following formula

$$\chi_0 = (\frac{2}{\pi})^{\frac{1}{2}} \frac{(\Delta x/2)^{1-b}}{a(1-b)u} \{Q_0 + \sum_{i=1}^{n} Q_i[(2i+1)^{1-b} - (2i-1)^{1-b}]\} \tag{2.5}$$

where χ_0 is the pollutant concentration at ground level, u is the mean wind speed, Δx is the source inventory grid spacing, Q_i are the source strengths in the $n+1$ upwind source boxes, and $i = 0, 1, \ldots, n$. The total ambient air quality then follows by combining the contributions from equation (2.5) with the point source contribution Q_0 and the background concentrations.

The ATDL model uses typical values of grid spacing Δx from 1 km to 10 km. In equation (2.5), the term $(\Delta x/2)^{1-b}$ varies by less than a factor of 5 over the range of stabilities encountered in a given city. This model has been shown to yield comparable predictions with other more complex models in a wide range of urban environments by Eschenroeder (1975), Hanna and Gifford (1977), Daly and Steele (1976), and Gualdi and Tebaldi (1982).

Simple Box Models

The ATDL model is generally regarded as the first step towards obtaining a simple box model. A further step considers the source terms Q_i. When dealing with smooth area source distributions, it is noted that the variations in the source term outside the

receptor grid square do not greatly influence χ_0. This means that the coefficients of the Q_i terms are certainly less than that for the Q_0 term. Then the assumption is made that all source terms are equal to Q_0 and the simplified model based on equation (2.5) can be written as (see Hanna (1971))

$$\chi_0 = \frac{CQ_0}{U} \tag{2.6}$$

where

$$C = (\frac{2}{\pi})^{\frac{1}{2}}[(2n+1)\frac{\Delta x}{2}]^{1-b}[a(1-b)]^{-1}. \tag{2.7}$$

The term a(1-b) varies very slowly for the stability range normally encountered over cities, and is considered approximately constant for broad stability categories. C values have been assigned values of 50, 200 and 600 for unstable, neutral and stable conditions, respectively, for total suspended particulate (TSP) levels (Hanna, 1971). This simplified model suggests that the pollutant concentrations depend mainly on source strength and wind speed, and are virtually independent of the inversion height which may enter the calculations via the stability category (Simpson and Hanna,1981).

However, it has been found that the relationship between the pollutant concentration and wind speed may vary from time to time. Benarie (1978) examined equation (2.6) and revised it to include a parameter for the exponent of wind speed in order to incorporate such changes in the relationship. Thus, the simple box model can be generalized as

$$\chi = CQU^e \tag{2.8}$$

where e is the wind speed exponent which may be a seasonally varying climatological characteristic for a given city. Further discussion and development of the model will be given in Chapter 12. Basically, equations (2.6) and (2.8) imply that wind speed and pollutant concentration are inversely related as matched pairs of observations when e is -1. However, at best this inverse relationship is very broad under certain conditions.

Chapter 12 demonstrates the variability in this relationship for TSP and other pollutant concentrations in Canberra.

To take advantage of the causal relationship between wind speed and concentration, Daly and Steele (1976) and Simpson et al. (1983) assumed that the simple inverse relationship might exist between opposite percentile values of wind speed and air pollutant concentrations. Simpson et al. (1983) showed that simple probabilistic arguments can be used to convert equation (2.6) to a more general form

$$\chi_p = \frac{T}{U_{100-p}} \tag{2.9}$$

where χ_p is the air pollutant concentration corresponding to the p-percentile ordinate of the air pollution cumulative frequency distribution, U_{100-p} the wind speed corresponding to the $(100-p)$ - percentile ordinate of the wind speed cumulative frequency distribution, and T is a constant. The constant is derived from the relationship between χ_p and U_{100-p} for each sampling station under consideration over some percentile range for which T is approximately constant. The T term is the emission parameter Q multiplied by C, the latter parameter depending on atmospheric stability.

Without using direct knowledge of the source strength, Knox and Lange (1974), Benarie (1976) and Simpson et al. (1983) proposed the following simple formula for calculating T, namely

$$T = U_{50}\chi_{50} \tag{2.10}$$

where the right-hand terms are the medians of distributions of wind speed and pollutant concentrations, respectively. The simple model of Simpson et al. (1983) has been successfully applied to TSP and acid gas data, and $T = U_{100-p}\chi_p$ was found reasonably constant over the 30-70 percentile range (Simpson et al, 1985). Thus, such a model can be a good representation of the relationship of the frequency distributions of wind speed and pollutant concentration for at least the 30-70 percentile range.

Simpson et al. (1983) also showed that when combined with the assumption of

a lognormal distribution of pollutant concentrations, from which it follows that the distribution of wind speed is also lognormal, then equation (2.6) becomes

$$\chi_p = \frac{T}{a_u}(\beta_u)^{z_p} \tag{2.11}$$

where β_u is the geometric standard deviation, a_u the geometric mean of the wind speed data, and z_p is the standard variable corresponding to the percentile p. This model was developed to yield estimates of the entire distribution of pollutant concentration (Simpson et al., 1983) and has been used to forecast worst case pollution scenarios for particulates due to urban industrial development (Simpson and Jakeman, 1985). Nicholson (1975) proposed the simple box model to predict street level concentrations of traffic CO emissions. Leahy (1975) has used a moving box model for hourly ground-level concentrations of nitrogen oxides (NO_x) at Edmonton, Canada. Smith (1976) also developed a simple box model incorporating a simple relationship between the mixing layer depth and the horizontal dimension of the box to obtain good results for climatological averages of pollutant concentrations of SO_2 for a number of English cities.

Multi-box Models

The principle of multi-box models is to use the stepwise movement of a box to describe a curvilinear path over the ground. It is dependent on the choice of adequate wind direction to trace the trajectory. Such a procedure is known as a back-tracing or reverse trajectory method and is commonly used in weather forecasting and atmospheric research.

Multi-box models have been widely used for different pollutants and in various locations. MacCracken et al. (1971) developed a multi-box model to simulate hourly CO data in California. Gifford and Hanna (1971) have also used this method to model SO_2 concentrations in Milan, Italy. In Japan, the multi-box model has been used by several authors. Shiozawa et al. (1973) estimated SO_2 concentrations in Tokyo. Ishikawa

et al. (1973) used a two-layer multi-box model in Osaka. Funabashi (1973) also employed real-time filtering for prediction purposes. Benarie (1980) discussed how to use a multi-box model to calculate CO concentrations. Dabbert et al. (1973) suggested a model, named APRAC, to predict concentrations of inert, vehicle-generated pollutants. The derivation of APRAC is quite similar to that of Gifford (1973). Ragland (1973) provided a steady-state implicit finite-difference matrix solution for an array of $n \times n$ boxes in the vertical plane. Calculated concentration patterns in the x-z plane appear similar to those obtained by Egan and Mahoney (1972), who employed the mass continuity equation directly. Kontnik (1974) designed a multi-layer box model to account for both light winds and non-uniform wind fields, which essentially moves material along the wind directions, with the addition of a proper amount of vertical mixing. Hameed (1974) has compared a simple version of the ATDL model with a more complex one by Randerson (1968) in studying a two-hour SO_x episode in Nashville, Tennessee. Hameed (1974) found that the simpler model yields comparable results to the more complex one.

Rollback Models

A lack of meteorological information or knowledge of the relationship between meteorological data and pollutant dispersion can cause great difficulties in the prediction of air pollutant concentrations. In these cases, the rollback model can be used to provide a simple method to assess source pollutant emissions required to satisfy air quality criteria. The rollback model may be considered as belonging to the same family as the box model (Benarie, 1980), so that it is sometimes called the receptor-oriented symmetrical counterpart of the box model. A basic physical assumption of the rollback model is that the pollutant concentrations are directly proportional to emissions according to some simple relationship. Without consideration of meterological parameters and local effects, such as the terrain, the emission control requirements can be determined as proportional to the amount by which the peak pollutant concentration exceeds the desired air quality standards.

The simplest form of the rollback model is of the type

$$\chi = \chi_b + Re \tag{2.12}$$

where χ is the pollutant concentration due to emissions with rate e, and χ_b is a measure of background pollutant concentration. R is the constant of proportionality which includes all the effects relevant to the meteorology and area source distribution, and can be determined as (de Nevers and Morris, 1975)

$$R = (\frac{\chi_{max} - \chi_b}{e}) \tag{2.13}$$

where χ_{max} is the highest pollutant concentration in the region of interest. According to the selected air quality standard, the allowable emission rate for a new χ_{max} can be obtained from

$$e_{max} = \frac{e(\chi_{sta} - \chi_b)}{(\chi_{max} - \chi_b)} \tag{2.14}$$

where χ_{sta} is the designated air quality standard specified for the pollutant being considered. The required reduction from a peak of pollutant concentration can be obtained from

$$P = 100\frac{\chi_{max} - \chi_{sta}}{\chi_{max} - \chi_b} \tag{2.15}$$

where P is the percentage reduction required (Schuck and Papetti, 1973).

A generalised formulation of the simple rollback model is given by (Chang and Weinstock, 1973, 1974)

$$\chi_i = \chi_b + \sum_{i=1}^{n}\sum_{j=1}^{m} C_{ij}e_j \tag{2.16}$$

where χ_i is the concentration at receptor i, e_j is the emission rate for source j, and R_{ij} is the source-receptor interaction for source j and receptor i. The C_{ij} may be calculated from a moving box model or a Gaussian plume model.

The simple rollback model has been used successfully to examine the motor vehicle emission goals for standards governing CO, NO_x and hydrocarbons (Barth (1970) and Schuck and Papetti (1973)). The technique has also been employed to describe photochemical smog effects in terms of the primary pollutant concentrations (Hamming et al, 1973). de Nevers and Morris (1975) extended the basic technique to apply to multiple sources, different stack heights, different source-to-receptor distances and wind direction frequencies. Szepesi (1977) specified source-receptor functions as Gaussian for point and area sources. Peterson and Moyers (1980) developed a model for the case where continuous measurements of ambient concentrations and emissions are available and recorded over time intervals corresponding to air quality standards. Georgopoulos and Seinfeld (1982) recommended the use of the mean values $E(\chi_{max})$ and $E(\chi_{sta})$ instead of χ_{max} and χ_{sta} in rollback calculations, which has the advantage of allowing for the conservation of mass of non-reactive pollutants.

The nonlinearity of atmospheric processes limits the usefulness of the rollback model, as does its lack of spatial resolution. Horie and Overton (1974) noted that the higher the percentile value of concentration considered as the desired air quality goal, the greater the uncertainty in the emissions reductions calculated by the rollback technique. When using the model to predict the rate of growth of air pollution due to urban development, it must be assumed that the distribution of sources is unchanging with time. Therefore, the rollback model may be used for regional analysis of areas with many well distributed sources of various types and as a first step approach or screening model to obtain a crude picture of future trends.

2.3.4 Performance and Validation of Deterministic Models

From the preceding discussion above, it is seen that deterministic models vary in descriptive and computational complexity. They can be simple (e.g. the Hanna-Gifford ATDL model), intermediate (e.g. Gaussian plume model) or complex (e.g. numerical models based on K-theory). Numerous investigations have shown that, under certain conditions, a simple modelling approach may perform quite well when compared

with more complex models for estimation of ambient pollutant concentrations resulting from the dispersion of pollutants in an airshed. Simpson and Hanna (1981) argued that the advection effects of the atmosphere dominate horizontal diffusion effects for long periods. Therefore, the vertical diffusion is relatively less important and it can be accommodated by simple Gaussian or box assumptions. On the other hand, complex models such as the K-diffusion model merely involve different assumptions to handle what is commonly regarded as a highly stochastic problem, and also require a numerical solution which may introduce computational errors.

The best feature of deterministic models is that they can be used for approximation of the causal link between the variables, such as those describing emissions, meteorological conditions and terrain, and the dependent pollutant concentrations. These models have improved our understanding of the nature of pollutant dispersion in the atmospheric environment and describe the physical processes of pollutant dispersion. In practice, most applicable deterministic models are useful at best for predicting the mean of pollutant concentrations (see e.g. Jakeman et al., 1988). Many deterministic models can predict long-term means of pollutant concentrations for a wide range of physical circumstances with reasonable accuracy. Such models retain sensitivity to variations in both mean emission strengths and meteorological variables, such as wind speed and wind direction. Thus, deterministic models are generally best suited to estimation of mean pollutant concentration under mean conditions.

Deterministic modelling encounters two major practical difficulties with respect to model performance. First, the deterministic models are not capable of predicting extreme pollutant levels especially well, and many air quality standards require this knowledge. Second, by their very nature, deterministic models cannot characterize the uncertainty in model predictions.

It has been found, for example, by Simpson and Hanna (1981) in an examination of the predictive ability of various deterministic models, that the values of the extremes of the distribution are unlikely to be very accurate. Pasquill and Smith (1983) argue

that it is the special nature of the meteorological conditions and other circumstances which combine occasionally to form the worst pollution episodes, and it is difficult to model such extreme occurrences. Hanna (1982) refers to 'natural variability' as the turbulent fluctuations in wind velocity which may occur over time periods ranging from microseconds to years. Obviously, the existence of natural variability limits strongly the estimation accuracy of air pollutant concentrations using deterministic models. Venkatram's (1983, 1984) analysis reveals that the expected deviation of observations from predictions becomes large when the sampling time is not much greater than the time scale controlling diffusion. From the study of Hanna (1982) and the theoretical analysis of Venkatram (1984), it is often stated that the accuracy of predictions of existing deterministic models for ensemble means is approximately of order 2.

The accuracy and application of deterministic models is often restricted due to the lack of essential meteorological or topographical information being available, particularly for complex models. Enhancing the data collection substantially raises the costs of model development, which may be prohibitive in many circumstances. Therefore, a simple but functional deterministic model is normally very important in practice for air quality management problems.

In conclusion, a wide range of deterministic models is available in the literature (and in computer packages) for impact assessments of air quality. Such models are most useful in predicting concentration values around the mean or median of pollutant concentrations over time periods no shorter than one hour. The highly stochastic nature of turbulent diffusion is a problematic constraint. Deterministic models are unlikely to perform well in estimating the upper percentiles of the distribution of air pollutant concentrations. Different approaches are required to circumvent these problems and to satisfy the demands of air quality management.

2.4 Statistical Modelling

While the main emphasis has traditionally been placed on development of deterministic models for describing the physical behaviour of atmospheric dispersion of air pollutants, the statistical properties of air pollutant concentrations are important because of the complexities which arise in the physically-based analysis of atmospheric turbulence. Although predictions of air pollutant concentration might be obtained from a sophisticated model, large departures will still be expected when compared with observed air quality data. The statistical description of turbulent flow is an essential tool in representing the fluctuations of a variety of meteorological and emission quantities. Air pollutant concentrations are inherently random variables in nature.

Since air pollutant concentrations are normally measured sequentially over time, and averaged over successive non-overlapping time periods of equal length, air quality data consist of statistical time series which can be written as (Georgopoulos and Seinfeld, 1982)

$$\chi_1(t_1), \chi_2(t_2), \ldots, \chi_n(t_n); \qquad t_1 < t_2 < \cdots < t_n \tag{2.17}$$

where the sampling period is known as the averaging time τ, defined as $\tau = t_2 - t_1 = t_3 - t_2 = \ldots = t_n - t_{n-1}$.

Since atmospheric systems extend up to the large scales associated with disturbances of the general circulation, the properties of air quality data depend considerably on the sampling duration. That is, the length of the averaging time will affect the degree of correlation of successive data points (Georgopoulos and Seinfeld, 1982). For a given pollutant and measurement site, there is an inverse proportional relationship between the averaging time and the degree of autocorrelation, such that the longer is the averaging time, the less is the autocorrelation. Furthermore, the properties of air quality data also depend on the specific place and time at which the observations are made. Due to the effects of variable terrain, of diurnal heating and nocturnal cooling of the ground, and of the continually changing large-scale pattern of air flow, air

pollutant concentrations are neither homogeneous nor stationary (Pasquill and Smith, 1983). Based on the statistical characteristics of air quality data, there are several alternative statistical approaches which have been applied to air quality assessment.

2.4.1 Probability Distribution Modelling

The assessment of environmental impact for air quality management in terms of air quality standards is based on a probability curve of concentrations measured over a fixed averaging time at locations of interest. Such an assessment requires specification of at least the mean and upper percentiles of the frequency distribution of pollutant concentrations. Thus, probability distribution modelling plays an important role in control and management of air pollution, and is of particular interest in this thesis.

The graphical nature of air pollutant frequency concentrations over a given averaging time can be viewed with the aid of an histogram. A typical histogram of air quality data tends to be unimodal and skewed to the right (Benarie, 1980). Quite often the histograms of air pollution concentration appear to be inversely "J" shaped, having a peak value of frequency near the lower concentrations and a gradual but long tail. Based on such intuitive information, many skewed distributions have been developed in the statistical literature and have been demonstrated to be useful for fitting air pollution data. Benarie (1980) has enumerated distributions such as: Poisson (Benarie (1980)); negative binomial (Prinz and Stratman, 1966); Weibull (Barlow, 1971; Curran and Frank, 1975; Tsukatani and Shoyi, 1977); exponential (Barry, 1971; Scriven, 1971; Curran and Frank, 1975); gamma and beta (Lynn, 1972; Graedel et al., 1974); lognormal (Mage, 1975; Larsen, 1977a,b).

Probability distribution models complement deterministic models (Jakeman et al., 1988) but, unlike deterministic models, they provide solutions which are not certain, as they are described within the framework of probability. They are especially useful in cases where the output of natural systems cannot be expressed satisfactorily as a fixed function of input variables, whether the reason be inadequate understanding

of the system of interest, inadequate data or inherent randomness of the observations. These problems are circumvented by modelling the observations in a simple parametric manner.

The uses of efficiently parameterised probability distribution models have been summarised by Ott and Mage (1979), Bai et al. (1988), and Jakeman and Taylor (1989). Such models can provide a simple description and summary of a set of data as a member of a general class of distributions. The reduction of masses of data to more manageable quantities with relatively few parameters provides administrative benefits, for example, in storage and transmission, by converting long records of data to a simple parametric form which retains the basic information needed for future reference. The models can be used to filter the effects of noise inherent in raw observations by interpolation or extrapolation. For example, they can fill in gaps created by missing information that is random or uniform, smooth measurement, sampling error, or unrepresentative events.

As argued in Bai et al. (1988), statistical inferences can be drawn from the parameterisation taking into account the properties of the methods and the raw data used. Hypotheses related to the population can be tested in order to reach certain conclusions. Statistical models can express uncertainties or tolerances, and the variability of the system can be quantified. These models can also be applied to design, analyse and assess sampling methods or data bases. Finally, Jakeman et al. (1988) show how statistical models can be augmented with deterministic models to obtain hybrid models with the desirable properties of each modelling approach for prediction under a wide range of conditions. These hybrid models will be discussed in a later section.

A major limitation of probability distribution modelling is that calibrated models cannot be expected to be valid under conditions other than those which existed during data collection. Prediction under wide-ranging conditions is restricted because such models contain no explicit relation between parametric form, or at least between parametric values and the major causal factors. On the positive side, these statistical models can be designed to predict all events well, being based closely on observations. Because

observations are usually assumed to be independent and identically distributed (iid), uncertainty in parameters and hence model properties can be characterised easily by appropriate estimation methods. The iid assumption, however, restricts the straightforward identification of statistical models when observations are autocorrelated. As with time series modelling, it is possible to induce independence and stationarity by variable transformation or selection of appropriate subsets of observations for analysis. Even without adjustment for its effect, some autocorrelation can still be tolerated since it does not affect the estimate of the mean, only the estimates of the variance (see Jakeman et al. (1986) for further details).

Probability distribution modelling has been successfully applied in air quality applications. Larsen (1964, 1969, 1973, 1974) developed the so-called statistical model for predicting maximum air pollutant concentrations across an airshed from limited sets of observations, in conjunction with a single continuous monitoring site. Jakeman and Taylor (1989) summarise the applications of probability distribution modelling. Further developments of this capability will be discussed in the hybrid modelling section of this chapter.

2.4.2 Stochastic Modelling

Air quality data are essentially autocorrelated and non-stationary statistical time series, although the degree of autocorrelation and non-stationary can vary considerably. In recent years, stochastic modelling of diffusion has become increasingly popular. A simple form of stochastic time series modelling is the linear rollback model, described as being based on the assumption that pollutant concentrations are proportional to emissions. Similar concepts can also be used to construct source-oriented models which establish transfer functions between the distribution of pollutant emissions and concentrations for atmospheric dispersion. These transfer functions can be obtained empirically through an appropriate mathematical inversion technique from accumulated data on joint distributions of air pollution and emission, and can be used for predicting air pollutant concentrations (Benarie, 1980). To determine the functions adequately, a

large amount of data is necessary and the assumption that there is a consistent source-receptor relationship in the region under study is essential. Successful applications of this method were illustrated by Meisel (1976), Meisel and Teener (1976), and Breiman and Meisel (1976).

Some stochastic modelling approaches apply the random walk to atmospheric dispersion. Initial work can be traced to Einstein (1905), who first used the familiar 'drunkard's walk' to simulate molecular diffusion. Recent approaches adopt the Markovian assumption to treat eddy diffusion as a continuous process. In the simplest case of one-dimensional homogeneous stationary turbulence, the random walk equation can be written as

$$V(t+\Delta t) = W(\Delta t)V(t) + V'(t) \tag{2.18}$$

where W is the Lagrangian correlation function and $V'(t)$ is a random velocity drawn from a Gaussian distribution with zero mean and standard derivation σ_v (Pasquill and Smith, 1983). Smith (1968) used equation (2.18) in a study of conditioned particle motion in homogeneous turbulence, and Hall (1975) applied the same method to simulate sea spray droplet motions and their resulting distribution in the surface layer of the atmosphere. Significant contributions to Markovian modelling include those of Hanna (1978), Reid (1979), Durbin and Hunt (1980), Lamb (1982), Ley (1982), and especially Wilson et al. (1981) and Thomson and Ley (1982).

An alternative to the Markovian assumption is to use the well-established Kalman (1960) filtering technique to predict air pollutant concentrations. The underlying model is known as the state-space model. This method specifies an optimal estimate of the state in a time-varying dynamic system with additive Gaussian noise. The estimate obtained at each time step is optimal in, for example, the mean square sense, based on all the observations up to that time. The Kalman filter is applied to models of the form

$$y_k = Z_k b_k + e_k, \qquad k = 1, 2, \cdots \tag{2.19}$$

where y_k is the $n_k \times 1$ vector of observations available at time k, Z_k an $n_k \times q$ known matrix and b_k the $q \times 1$ state vector of the estimates. The state vector is allowed to change through time in accordance with the state equation

$$b_{k+1} = T_{k+1} b_k + w_{k+1}, \qquad k = 1, 2, \cdots \tag{2.20}$$

where T_{k+1} is the transition matrix, and e_k and w_k are independently distributed multivariate normal random variables (Sallas and Harvile, 1981).

The Kalman filtering technique has been applied to air pollution forecasting by Takamatsu et al. (1971) by using the basic Gaussian plume concept to formulate the state equation. Wells and Lau (1971), and Bankoff and Hanzevack (1973, 1975), also used the technique for numerical integration of the mass transport balance equation.

2.5 Hybrid Modelling

The atmospheric environment is regarded as a complex system which requires both deterministic and statistical modelling techniques. Until recently, these two approaches had undergone parallel but separate developments. As discussed before, each approach has its own advantages and weaknesses, with augmentation of these two approaches providing additional improvements.

Some insights were shown in the 1970's where it was recommended that deterministic and statistical models be combined. Eschenroeder (1975) suggested that estimation of the parameters of a lognormal distribution could use the distribution of the deterministic model output. Benarie (1976) recommended a possible link between the distribution of wind speeds and pollutant observations through a simple inverse relationship between percentiles. Simpson et al. (1983) combined the ATDL model with Larsen's lognormal statistical model for estimating the maximum concentration. Since

then, the hybrid approach has been successfully extended, developed and used in many applications.

The hybrid modelling approach adopts systematic methods in model construction for predicting air pollutant concentrations (Jakeman et al., 1987), and is based on the notion that the air pollution system is made up of both deterministic and stochastic components. The mean conditions may be determined reasonably well by fundamental laws but many extreme conditions can only be estimated statistically. The approach sets out guidelines for relating the deterministic component of the system to the stochastic component, thereby deriving relationships between the extreme pollutant events and the driving forces in the system, such as emissions and meteorology. There are no fundamental laws for predicting extreme events, but there are empirical relationships which vary in different situations. However, many applications have shown that such relationships are predictable if extensive data sets are available (Jakeman et al., 1987). Alternatively, such relationships can be assumed and the sensitivity to different assumptions can be investigated.

The major steps in the application of hybrid modelling are as follows. First, information related to the sources, pollutant type, meteorology, the topography between source and monitor, and the historical pollutant concentration data are required as input. From this information, selection among deterministic models and different parametric distributions can be made in order to obtain the correct deterministic and statistical components. Selection of deterministic models can be made from a range of available model types generally relevant to source type and available emission and meterological data. Choice of a final model is based on the performance in predicting percentile concentrations, but when the performance is equal among models, parsimony and the level of input information available may be considered as final determining factors.

Second, from input variables such as emission and meteorological data, the selected deterministic model will be used to predict pollutant concentrations in the middle

percentiles (e.g. means, medians), where the model invariably performs the best and has the greatest mechanistic reliability.

Third, the middle percentiles obtained from the output of deterministic models are fitted to the statistical distribution chosen at the preceding step and parameter estimates are generated.

Finally, the more extreme events can be predicted from the estimated distribution. The output of hybrid models can thus be compared with air quality standards.

In practice, hybrid models have been applied with considerable success. Their numerous advantages over deterministic and statistical models include:

(a) provision of relatively good predictions over the entire distribution of pollutant concentrations, in particular the upper percentiles; and

(b) a causal link between source and receptor via the deterministic model component which relates pollutant concentrations to emissions and meteorological data, thereby retaining sensitivity to variations in inputs when meteorological conditions vary or emission control strategies change (e.g. changing stack heights and pollutant emission levels).

When employing hybrid modelling, the following assumptions apply:

(a) the historical pollutant data should be sufficiently informative to develop a distributional type for the statistical component;

(b) the distributional form for a given pollutant at a specific site should remain consistent with changes in emissions and meteorological conditions;

(c) the deterministic form should be commensurate with the input information available and capable of predicting reliably the required range of percentiles for the distribution of pollutant concentrations; and

(d) hybrid modelling is presently confined to predicting the frequency distribution of pollutant concentration over time, and pollutants have been restricted to relatively inert types. Hybrid models are not used to predict a particular pollutant concentration in time without related frequencies.

It should be noted that the first two limitations above arise from the major underlying assumption of hybrid modelling. Changes in emissions and meteorology from year to year basically do not affect the form of the distribution but merely the values of the parameters of the distribution. According to this presupposition, historical data can be used to identify the distributional form of the statistical component. Obviously, sufficiently informative data sets are necessary.

Part II

ESTIMATION

Chapter 3

A New Approach to Maximum Likelihood Estimation of the Three-Parameter Distributions

3.1 Introduction

Continuous univariate distributions, such as the gamma and Weibull, have received much attention since they first appeared in the literature. They have been used extensively in many areas such as reliability and life-testing (e.g. Mann et al. (1974) and Bain (1978)), hydrology (Stedinger (1980)), and air quality management (Jakeman and Taylor (1989), Jakeman et al. (1986)), where natural laws can be modelled quite successfully. Three parameter distributions involving the shape, scale and location of the distribution have been considered to be reasonably satisfactory, containing sufficient flexibility and sensitivity to fit real data while avoiding the problem of over-parameterization. For this reason, these distributions have been examined by many authors, such as Harter and More (1965), Johnson and Kotz (1970) and Cohen and Whitten (1982).

However, there remain some well-known problems in the literature when employing the maximum likelihood (ML) method to estimate the parameters of the gamma and Weibull distributions. A theoretical difficulty occurs when the shape parameter is less than or equal to unity, since one of the three first-derivative equations often used to maximize the likelihood function is not valid in these cases.

Some effort has been made to overcome this weakness. For example, Johnson and Kotz (1970, p. 185) suggested that ML should not be used for the gamma distribution if the shape parameter is less than 2.5 and gave details of the method of moments as an alternative procedure. Cohen and Norgaard (1977) also discussed this theoretical difficulty and developed a modified ML method which introduced the method of moments formula to avoid the failure of ML. It can be shown that such methods involving the method of moments will have a high degree of deviation from the parent distribution and will be outperformed when compared with ML over a suitable range of the shape parameter (Bai et al. (1988), and Bai and Taylor (1986)).

Cheng and Amin (1983) provided a method called the maximum product of spacings (MPS) estimation. For a distribution function F with parameter $\boldsymbol{\theta}$ and random observations x_i, their method uses the transformation $y_i = F(x_i, \boldsymbol{\theta})$, $i = 0, 1, \ldots, n+1$ to transform the sample into the interval $(0, 1)$ and maximize the geometric mean of the spacings, $D_i = y_i - y_{i-1}$, instead of the maximum likelihood function itself. Unfortunately, their method is very difficult to use because the derivative equations of MPS are complicated and implicit. In particular, if the distribution has no explicit form of the cumulative function, such as in the case of the gamma distribution, the transformation of the data itself is implicit and the derivative equations become more complicated and are difficult to solve. Estimation using MPS also appears to be less accurate than the ML method. Further discussion and comparison of this method with the ML method are given in the appendix.

It is known that, in the definition of the gamma and Weibull distributions, the range of the shape parameter is greater than zero. If the range of the shape parameter is less than or equal to unity, ML technically fails in its performance because its three first-derivative equations cannot be used. Some simple techniques can be used to overcome this difficulty while still following the principle of ML.

The primary motivation for this chapter is to propose a general methodology which can provide satisfactory ML estimates consistently and efficiently. This methodology

can be used, in principle, for all distributions with shape, scale and location parameters, and is especially useful when ML breaks down. Such a methodology also has a computational advantage in various applications since the normal ML procedure sometimes has difficulties in converging, particularly when the sample size is not especially large. In the following sections, estimation of both the gamma and Weibull distributions is examined by this method but more emphasis is placed on the former, together with the other three alternative methods for purposes of comparison. Monte Carlo simulations are used to assess the performance of each method. Some real pollutant data from Melbourne, Australia are used for purposes of empirical illustration.

3.2 A General Approach to Maximum Likelihood Estimation of the Three Parameter Gamma and Weibull Distributions

The likelihood principle in its general form selects parameters of the distribution over the admissible range to make the likelihood function as large as possible (see e.g. Kendall and Stuart (1979)). The probability density functions of the random variable for the three- parameter gamma and Weibull distributions are of the form:

$$f(x) = \frac{1}{\beta\Gamma(\alpha)}(\frac{x-\gamma}{\beta})^{\alpha-1}exp[-(\frac{x-\gamma}{\beta})] \tag{3.1}$$

$$f(x) = \frac{\alpha}{\beta}(\frac{x-\gamma}{\beta})^{\alpha-1}exp[-(\frac{x-\gamma}{\beta})^{\alpha}] \tag{3.2}$$

respectively, where β represents the scale parameter, α the shape parameter, γ the location parameter and Γ the gamma function. In the above equations, $\beta > 0$, $\alpha > 0$ and $\gamma < x < \infty$. If $x_1, \ldots, x_n$ is a random sample of n observations, then the logarithmic likelihood functions for the gamma and Weibull distributions are:

$$logL = -n\alpha log\beta - nlog\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^{n} log(x_i-\gamma) - \sum_{i=1}^{n}(\frac{x_i-\gamma}{\beta}) \tag{3.3}$$

$$logL = nlog\alpha - n\alpha log\beta + (\alpha-1)\sum_{i=1}^{n} log(x_i-\gamma) - \sum_{i=1}^{n}(\frac{x_i-\gamma}{\beta})^{\alpha} \tag{3.4}$$

respectively. Taking partial derivatives with respect to the parameters γ, β and α yields the following first-order conditions:

Gamma:

$$\frac{\partial logL}{\partial \gamma} = -(\alpha - 1)\sum_{i=1}^{n}(\frac{1}{x_i - \gamma}) + \frac{n}{\beta} = 0 \tag{3.5}$$

$$\frac{\partial logL}{\partial \beta} = -\frac{n\alpha}{\beta} + \sum_{i=1}^{n}(\frac{x_i - \gamma}{\beta^2}) = 0 \tag{3.6}$$

$$\frac{\partial logL}{\partial \alpha} = -nlog\beta - n\Psi(\alpha) + \sum_{i=1}^{n} log(x_i - \gamma) = 0 \tag{3.7}$$

Weibull:

$$\frac{\partial logL}{\partial \gamma} = -(\alpha - 1)\sum_{i=1}^{n}(\frac{1}{x_i - \gamma}) + \frac{\alpha}{\beta^\alpha}\sum_{i=1}^{n}(x_i - \gamma)^{\alpha-1} = 0 \tag{3.8}$$

$$\frac{\partial logL}{\partial \beta} = -\frac{n\alpha}{\beta} + \frac{\alpha}{\beta^{\alpha+1}}\sum_{i=1}^{n}(x_i - \gamma)^\alpha = 0 \tag{3.9}$$

$$\frac{\partial logL}{\partial \alpha} = -\frac{n}{\alpha} - nlog\beta + \sum_{i=1}^{n} log(x_i - \gamma) - \sum_{i=1}^{n}(\frac{x_i - \gamma}{\beta})^\alpha log(\frac{x_i - \gamma}{\beta}) = 0 \tag{3.10}$$

where $\Psi(\alpha)$ is the digamma function, given as

$$\Psi(\alpha) = \frac{\partial log\Gamma(\alpha)}{\partial \alpha}.$$

The usual approach to ML estimation is to simultaneously solve the derivative equations (5) to (7) and (8) to (10) for the parameters of the two distributions. It is clear that this method is not valid if $\alpha \leq 1$ because the terms $\frac{n}{\beta}$ and $\sum_{i=1}^{n}(\frac{1}{x_i-\gamma})$ in (5), and $\sum_{i=1}^{n}(\frac{1}{x_i-\gamma})$ and $\frac{\alpha}{\beta^\alpha}\sum_{i=1}^{n}(x_i-\gamma)^{\alpha-1}$ in (8), are always positive values. Obviously, it is then not possible to satisfy the first-order conditions for ML estimation simultaneously. For further details of estimation when $\alpha \leq 1$, see e.g. Cohen and Norgaard (1977).

3.3 A New Approach to Maximum Likelihood Estimation

In contrast to the derivative method for finding the extreme value of a given function, it is possible to consider an optimization approach to maximise directly the likelihood function and search for its extreme value over the permissible range of parameters. Although not reported in this chapter, extensive experiments show that the optimization method of searching in three dimensions for three parameters can provide quite satisfactory estimates for the three-parameter gamma and Weibull distributions without limiting the range of the shape parameter whenever the maximum of the likelihood function exists. However, this three-parameter optimization method is not computationally efficient and also consumes considerable CPU time.

A computationally more efficient method may be obtained by combining the optimization of the likelihood function with an associated simplification and reduction from a three- to a two-parameter distribution. The main difference between the three- and two-parameter distributions lies in the location parameter, which the latter sets equal to zero. The curve for both distributions is principally determined by the shape and scale parameters. Therefore, a three-parameter distribution can be treated as a two-parameter distribution after using the linear transformation given by

$$x_i' = x_i - \gamma \qquad (i = 1, 2, \ldots, n) \tag{3.11}$$

which simply transforms the original variable x to a new variable x'. Two-parameter gamma and Weibull distributions can be estimated without any difficulties over the full range of the shape parameter by using the derivative equations. Therefore, using the transformation in (11) to eliminate the location parameter, the derivative equations can now be expressed as follows:

Gamma:

$$\frac{1}{n}\sum_{i=1}^{n} log x' - log\overline{x'} = \Psi(\alpha) - log\alpha \tag{3.12}$$

$$\overline{x'} = \alpha\beta \tag{3.13}$$

Weibull:

$$\alpha = [(\sum_{i=1}^{n} x_i'^{\alpha} log x_i')(\sum_{i=1}^{n} x_i'^{\alpha})^{-1} - \frac{1}{n}\sum_{i=1}^{n} log x_i']^{-1} \tag{3.14}$$

$$\beta = (\frac{1}{n}\sum_{i=1}^{n} x_i'^{\alpha})^{\frac{1}{\alpha}}. \tag{3.15}$$

The key to this method is to use the optimization procedure to search over the entire admissible range of γ and then to obtain the shape and scale parameters by using (12), (13) and (14), (15) for the gamma and Weibull distributions, respectively. The upper bound should be less then the first observation according to the definition that $x_i > \gamma$ for $i = 1, 2, \ldots, n$. This method can be used for estimation over the entire range of parameters without restricting the range of the shape parameter. From our simulation experiments, it is found that this method can also avoid the convergence difficulties which are sometimes encountered in solving the derivative equations.

3.4 Improved General Approaches for the Gamma Distribution

If prior knowledge can be used to show that the distribution is bell-shaped, it is worthwhile considering an improvement in the derivative method which requires a three-dimensional parametric search. Rearrangement of equations (5) and (6), assuming a known value for γ, yields the direct analytic solutions for β and α in terms of γ as follows:

$$\alpha = \sum_{i=1}^{n} (\frac{x_i - \gamma}{n\beta}) \tag{3.16}$$

$$\beta = \sum_{i=1}^{n} (\frac{x_i - \gamma}{n}) - \frac{n}{\sum_{i=1}^{n} (x_i - \gamma)^{-1}}. \tag{3.17}$$

The use of equations (16) and (17) requires only a one-dimensional search for the location parameter γ. There are two ways of proceeding with this numerical search,

namely by substituting (16) and (17) into the derivative equation (7) or into the likelihood function (3). In both cases, conditional on the estimate of γ, the estimates of α and β are determined according to (16) and (17). The solution can be attained in the former case by reaching an arbitrary permissible tolerance level, or in the latter case by finding a local maximum of the likelihood function. Of course, unlike the method outlined in the previous section, these two methods of obtaining maximum likelihood estimates are still subject to the restriction that estimates of β and γ cannot be obtained where the estimate of α is less than or equal to unity.

3.5 Simulation Experiments

Monte Carlo simulations are undertaken to assess the performance of the estimation methods discussed in the previous sections. The parameters are taken over a range of possible values which might arise in practice. Results are reported in tabular form for shape parameters taking the values { 0.3, 0.5, 0.8, 1, 1.5, 2, 4, 6 } for the gamma distribution, and values { 0.3, 0.5, 0.8, 1, 1.5, 2, 3, 4 } for the Weibull distribution. These values span a large range of shapes which may arise in the analysis of life-testing and reliability, or air pollution concentrations (see Jakeman et al. (1986)). For the gamma distribution, the curve becomes symmetrical as the shape parameter exceeds 6 while, for the Weibull distribution, negative skewness will result as the shape parameter exceeds 4. These extreme cases are not of interest here. The scale and location parameters are set at unity in all cases in order to enable an emphasis to be placed on difficult situations. Figures 3.1 and 3.2 indicate the form of the probability density functions for some selected parameter values. For all parameter sets, one thousand experiments are conducted. The main sample size used is $n = 365$, which is chosen as it represents a common case: namely a year of 24-hourly average observations. In order to examine the effects of varying the sample size on the method described in the previous section, a range of sample sizes from 50 to 1000 (which covers most applicable situations) is also used.

In this Monte Carlo study the performance criteria recommended for assessing air quality distributions are chosen, namely the relative bias (BIAS) and the relative root mean square error (RRMSE) (Fox, 1981). The criteria are defined as follows:

$$BIAS(q) = \frac{1}{N}\sum_{i=1}^{N}(\frac{\hat{q}_i - q}{q}) \tag{3.18}$$

$$RRMSE(q) = [\frac{1}{N}\sum_{i=1}^{N}(\frac{\hat{q}_i - q}{q})^2]^{0.5} \tag{3.19}$$

where N is the number of Monte Carlo experiments, q is the true value of the parameter or the percentiles of the underlying distribution, and $\hat{q}_i$ is the estimate of the parameter for the i'th experiment.

The random sample generators used for the Monte Carlo study are DRNGAM for the gamma distribution and DRNWIB for the Weibull distribution, available as subroutines in the International Mathematical and Statistical Library (IMSL) in version 1.0 (April, 1987). The same seed number (1234) is used to obtain the first random sample of the first 1000 simulations. Varying the initial seed produces similar results to those reported in the chapter.

Tables 3.1 to 3.6 list the MEAN, BIAS and RRMSE for estimation of the parameters of the gamma and Weibull distributions for different sample sizes, different values of the location parameter and different methods of estimation. In Table 3.1, results from using four methods for estimating the gamma distribution when $n = 365$ are given, namely method 1 (the new ML approach of Section 3), method 2 (the general ML approach of Section 2), methods 3 and 4 (the improved general ML approaches of Section 4, the former solving the third equation to obtain γ and the latter by maximising the likelihood function). In this table, the shape parameter is limited to $\alpha > 1$ because decreasing the value further will not be valid for methods 2, 3 and 4. It can be seen that these four methods perform reasonably closely in terms of estimating all three parameters of the distribution. The similar performances can be investigated in terms of the mean and standard deviation of the maximised log-likelihood function over 1000

simulation experiments. For the gamma distribution with shape parameter taking the value 2, the mean and standard deviation of the maximised log-likelihoods for methods 1, 3 and 4 are the same to four decimal places, namely - 573.90 and 15.68, respectively, while for method 2 they are - 573.91 and 15.71, respectively.

Considering CPU time, methods 3 and 4 are the most efficient computationally, followed by method 1 and finally method 2. The differences in computational efficiency increase with the value of the shape parameter. A conservative comparison of the efficiency of the four methods is therefore to consider the CPU time taken when fitting random samples from a parent gamma distribution with the shape parameter taking the value 2. Over 1000 experiments on samples of size 365, the estimation time for methods 1 to 4 are 7(min):36(secs), 10:13, 3:24 and 4:02, respectively. The rankings in terms of computational efficiency are broadly similar for many of the other experiments.

Table 3.2 reports only the results for method 1 when the shape parameter is equal to or less than unity in which case the other three methods are invalid. It is clear from the table that method 1 provides quite accurate results and the estimates are consistently close to the true values of all three parameters of the underlying distribution. When the true value of the shape parameter is less than unity, the estimates of the location parameter are always equal to unity, with BIAS and RRMSE both zero (or very close to zero). In order to investigate the performance of the new ML approach, Table 3.3 reports the results of estimating the gamma distribution when varying the sample size from 50 to 1000. The results in the table demonstrate that this method is capable of working well over the full range and produces consistent results following asymptotic theory. Accuracy improves as the sample size is increased, although there is little to be gained after $n = 300$ for the gamma distribution. The estimates of the location parameter are again very close to unity, with zero BIAS and RRMSE, for sample sizes in excess of 100 when the shape parameter is 0.8.

In Table 3.4, methods 1 and 2 are used for the Weibull distribution. The results are similar to those for the gamma distribution in that all three parameter values are

accurately estimated. As in the case for Table 3.2, Table 3.5 reports the results for estimating the Weibull distribution using method 1 only and the estimates are quite close to the true values of the parameters. Following along the lines of the gamma distribution in Table 3.3, the results in Table 3.6 also show that the new ML approach seems to perform extremely well for different sample sizes, with little improvement beyond $n = 200$ for the Weibull distribution. The results for estimates of the location parameter for the Weibull distribution when the shape parameter is less than unity (see Tables 3.5 and 3.6) are very similar to those obtained for the gamma distribution (see Tables 3.2 and 3.3).

3.6 Fitting Real Data

As illustrative examples, two data sets are chosen for air pollutant measurements of 24-hourly nitrogen dioxide concentrations in Melbourne, Australia. These two data sets present two diverse cases: "bell" shaped and reverse "J" shaped samples for the same pollutant and same year but measured at different stations. It is obvious that there is no prior knowledge to indicate whether the shape parameter of the parent distribution is less than, equal to or greater than unity, and it is uncertain whether the general approach of ML can be used or might break down. These two data sets demonstrate that the new approach of ML is necessary in order to enable fitting over the entire range of parameters without restricting the range of the shape parameter. In the first data set, the sample size is 271 from the Museum monitoring station in 1979; four methods for estimating the gamma distribution and two methods for the Weibull distribution are employed. The results are listed in Table 3.7. In the second data set, the sample size is 317 from the Alphington station in 1979. Only the new approach of ML for the gamma and Weibull distributions is applied and the estimated parameters are also listed in Table 3.7, with the gamma distribution fitting better than the Weibull distribution. Figures 3.3 and 3.4 show the fit to the data sets by the new ML approach. The results show that the new ML approach produces identical estimates to those of the

general approaches for the first data set, with the gamma distribution being superior to the Weibull distribution. The estimates are also satisfactory for the second data set, with the estimated shape parameter being less than unity and the ML estimate of the location parameter being equal to the minimum observed value of the data (as required). On the basis of these estimates, it is clearly useful to apply the new ML approach in general situations.

3.7 Discussion of the Maximum Product of Spacings (MPS) Estimation Method

The MPS estimation method is designed as a general method of providing consistent estimators by Cheng and Amin (1983). From their paper, this method can be applied to any univariate distribution, and is especially useful for non-regular cases such as the gamma and Weibull distributions when ML fails. They proposed that this method retains the desirable properties of ML estimation and, importantly, that the MPS method estimates the parameters of distributions under much more general conditions than the ML method. In some situations, the MPS estimator can be a function of sufficient statistics whereas the ML estimator is not.

As discussed in previous sections, the ML principle can be applied to a wide range of estimation problems under very general conditions. The ML method may be successfully used in many continuous univariate distributions by implementing the general approach of solving the derivative equations. In particular cases such as the gamma and Weibull distributions, a simple useful technique can be employed to overcome the difficulties encountered by the general method following the ML principle.

In comparison with the ML method and the new approach suggested in this chapter, the MPS method is computationally more complex and it can be less accurate.

3.7.1 MPS Estimation of the Gamma Distribution

For purposes of illustration, the gamma distribution will be used to describe the estimation method. The derivation of the estimation equations is as follows.

If $x_1 < x_2 < \ldots < x_n$ is a random sample of n observations from the gamma distribution, transform all the data into the unit interval (0,1) by using the transformation $y_i = F(x_i,\theta), i = 0, 1, \ldots, n+1$. The difference given by

$$D_i = y_i - y_{i-1} = \int_{x_{i-1}}^{x_i} f(x,\theta)dx \qquad (i = 1, 2, \ldots, n+1) \tag{3.20}$$

is called the (uniform) spacing of the sample $\{y\}$. The MPS method maximizes the geometric mean of the spacings

$$\begin{aligned} H &= logG = log(\prod_{i=1}^{n+1} D_i)^{\frac{1}{n+1}} = -\alpha log\beta - log\Gamma(\alpha) \\ &\quad + \frac{1}{n+1}\sum_{i=1}^{n+1} log\int_{x_{i-1}}^{x_i}(x-\gamma)^{\alpha-1}exp(-\frac{x-\gamma}{\beta})dx \end{aligned}$$

for which the first-order conditions are

$$\begin{aligned} \frac{\partial logG}{\partial\gamma} &= \sum_{i=1}^{n+1}\frac{\int_{x_{i-1}}^{x_i}[-(\alpha-1)(x-\gamma)^{\alpha-2} + \frac{1}{\beta}(x-\gamma)^{\alpha-1}exp(-\frac{x-\gamma}{\beta})]dx}{(n+1)\int_{x_{i-1}}^{x_i}(x-\gamma)^{\alpha-1}exp(-\frac{x-\gamma}{\beta})dx} \\ &= 0 \end{aligned} \tag{3.21}$$

$$\begin{aligned} \frac{\partial logG}{\partial\beta} &= -\frac{\alpha}{\beta} + \frac{1}{n+1}\sum_{i=1}^{n+1}\frac{\int_{x_{i-1}}^{x_i}(x-\gamma)^{\alpha-1}\frac{x-\gamma}{\beta^2}exp(-\frac{x-\gamma}{\beta})dx}{\int_{x_{i-1}}^{x_i}(x-\gamma)^{\alpha-1}exp(-\frac{x-\gamma}{\beta})dx} \\ &= 0 \end{aligned} \tag{3.22}$$

$$\begin{aligned} \frac{\partial logG}{\partial\alpha} &= -log\beta - log\Psi(\alpha) + \sum_{i=1}^{n+1}\frac{\int_{x_{i-1}}^{x_i}(x-\gamma)^{\alpha-1}log(x-\gamma)exp(-\frac{x-\gamma}{\beta})dx}{(n+1)\int_{x_{i-1}}^{x_i}(x-\gamma)^{\alpha-1}exp(-\frac{x-\gamma}{\beta})dx} \\ &= 0. \end{aligned} \tag{3.23}$$

As compared with equations (5) to (7), it is easy to see that the MPS method is considerably more complicated than ML since it needs to solve such difficult equations, even though the plotting technique for γ to see if it has a (local) maximum may lead to the deletion of (21). The ML method is relatively more simple and efficient, particularly using the new approach and the improved general approach to estimate the parameters of the gamma distribution. It is known that complex equations may introduce iterative errors, may cause difficulties with convergence, and may yield inaccurate solutions. Even when the cumulative distribution functions have explicit forms, such as the Weibull distribution, the equations for estimation are still more complicated than for the ML method.

3.7.2 Comparison of the ML and MPS Methods for the Weibull Distribution

A simple comparison shows that the MPS method can be less accurate than the ML method. The assessment criterion is simply the relative maximised log-likelihood function values corresponding to the estimated parameters of each method, together with observing whether the ML estimate of the location parameter is equal to the minimum observed data point if the shape parameter is less than or equal to unity. The higher value of the log-likelihood function will indicate which method provides estimates more accurately. For illustrative purposes, the samples used are 20 observations on maximum flood levels (in millions of cubic feet per second) for the Susquehanna River of Harrisburg over twenty, 4-year periods given in Dumonceaux and Antle (1973) and 20 observations on beach pollution (measured in number of coliform per 100 ml) in South Wales on twenty days over a 5-week period, provided by Steen and Stickler (1976). Both data sets are employed for the MPS estimation in Cheng and Amin (1983).

Table 3.8 shows that the new ML method provides much higher values of the log-likelihood functions than the MPS method. It is useful to note that the estimates of the shape parameters are less than unity for both data sets using the new ML method, and for the pollution data set using the MPS method. In such situations, the ML estimate

of the location parameter is the minimum observed value of the data. For both samples, the log-likelihood values associated with the MPS estimates are much lower than those obtained by the new ML approach, indicating the inaccuracy of the estimates given by the MPS method. Moreover, for the pollution data set, the ML estimate of the location parameter is, in fact, the minimum observed data point, whereas the MPS estimate is not, providing further evidence of the superiority of the ML approach over the MPS method.

3.8 Concluding Remarks

In this chapter, a new approach to maximum likelihood (ML) estimation is developed to overcome the difficulties encountered in some continuous univariate distributions, such as the three-parameter gamma and Weibull distributions. This method can also essentially be used for other three-parameter univariate distributions, such as the log-normal distribution. It can provide consistent and efficient estimators without placing any restrictions on the range of parameters concerned. Through the Monte Carlo study and the empirical illustration, the new approach is shown to be capable of performing well over an extensive range of parameter values and sample sizes, and can easily be used in applications where the value of the shape parameter may be greater than, less than, or equal to unity. There are also important implications for theory: the new approach complements the standard ML approach and removes the difficulties of estimation for all values of the shape parameter.

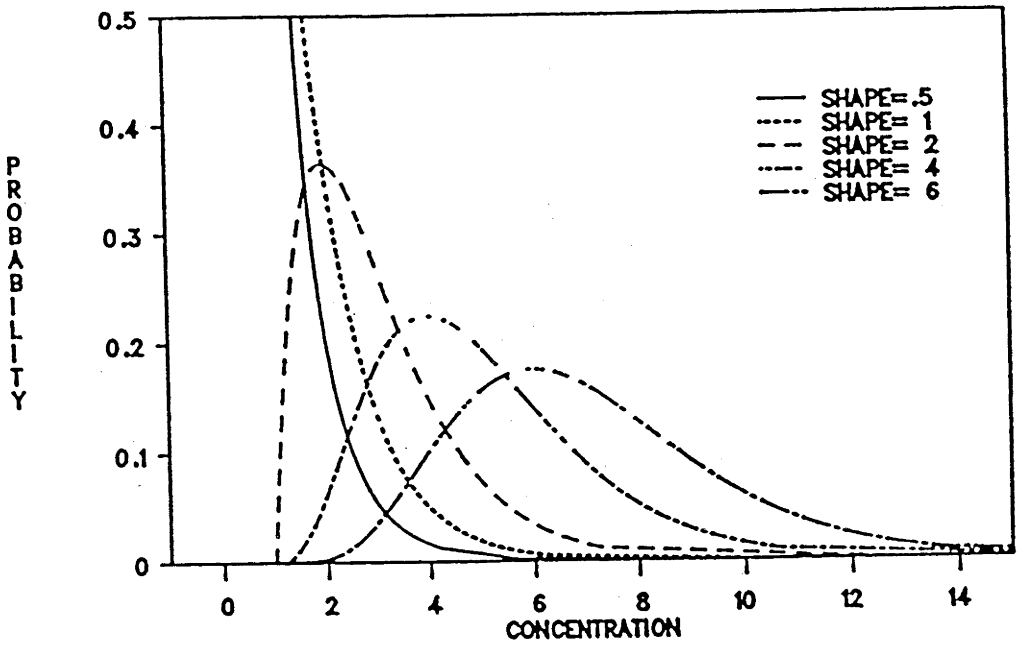

Figure 3.1: Profile of the gamma distribution for a range of shape parameters and unit scale and location parameters

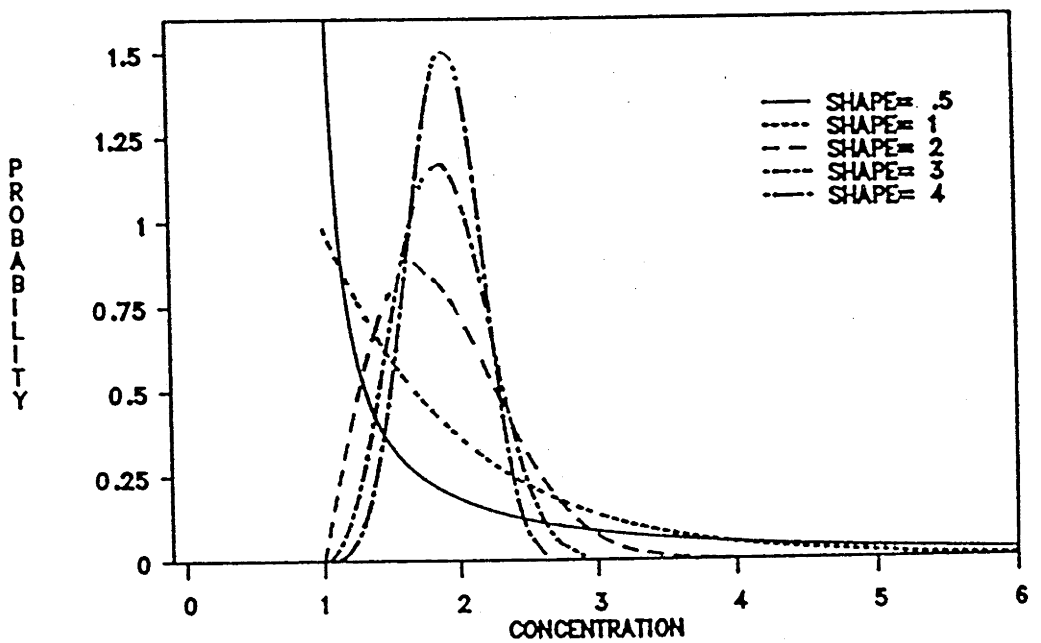

Figure 3.2: Profile of the Weibull distribution for a range of shape parameters and unit scale and location parameters

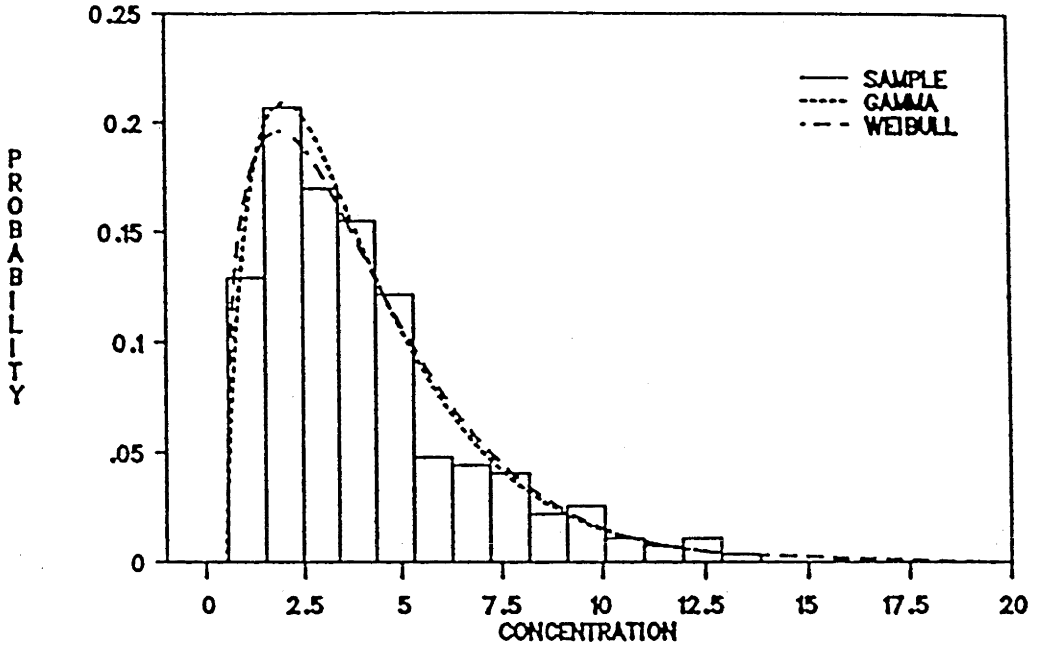

Figure 3.3: Fit of the gamma and Weibull distributions to an histogram of the nitrogen dioxide data at the Museum site

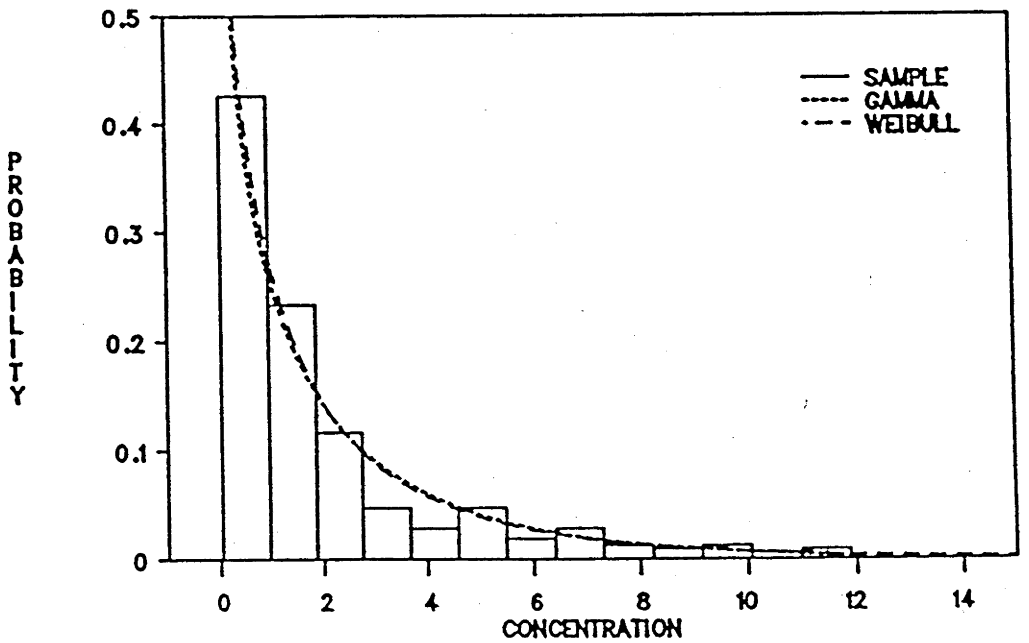

Figure 3.4: Fit of the gamma and Weibull distributions to an histogram of the nitrogen dioxide data at the Alphington site

TABLE 3.1

Estimates of parameters derived from four methods for the three-parameter gamma distribution based on 1000 Monte Carlo simulations with sample size n = 365 and $\beta = \gamma = 1$

Estimated Parameter	Performance Criteria	$\alpha = 1.5$ Method 1	2	3	4	$\alpha = 2.0$ Method 1	2	3	4
α	mean	1.458	1.460	1.460	1.454	1.953	1.947	1.948	1.947
	bias	-0.028	-0.027	-0.027	-0.031	-0.024	-0.026	0.026	0.027
	rrmse	0.089	0.089	0.089	0.091	0.103	0.104	0.105	0.106
β	mean	1.027	1.026	1.026	1.030	1.022	1.024	1.024	0.988
	bias	0.027	0.026	0.026	0.030	0.022	0.024	0.024	-0.012
	rrmse	0.097	0.098	0.098	0.098	0.100	0.100	0.101	0.097
γ	mean	1.011	1.011	1.011	1.011	1.019	1.021	1.021	1.022
	bias	0.011	0.011	0.011	0.011	0.019	0.021	0.021	0.022
	rrmse	0.021	0.020	0.020	0.021	0.048	0.049	0.050	0.052
Estimated Parameter	Performance Criteria	$\alpha = 4.0$ Method 1	2	3	4	$\alpha = 6.0$ Method 1	2	3	4
α	mean	3.947	3.947	3.947	3.947	6.034	6.038	6.034	6.026
	bias	-0.013	-0.013	-0.013	-0.013	0.006	0.006	0.006	0.0004
	rrmse	0.196	0.196	0.196	01961	0.268	0.270	0.268	0.284
β	mean	1.026	1.026	1.026	1.026	1.026	1.025	1.026	0.995
	bias	0.026	0.026	0.026	0.026	0.026	0.025	0.026	-0.005
	rrmse	0.134	0.134	0.134	0.134	0.160	0.159	0.160	0.168
γ	mean	1.045	1.044	1.045	1.045	1.047	1.045	1.047	1.044
	bias	0.045	0.044	0.045	0.045	0.047	0.045	0.047	0.044
	rrmse	0.291	0.291	0.291	0.291	0.652	0.657	0.652	0.701

TABLE 3.2

Estimates of parameters derived from new ML approach for the three-parameter gamma distribution based on 1000 Monte Carlo simulations with sample size n = 365 and $\beta = \gamma = 1$

Estimated Parameter	Performance Criteria	Shape Parameter α 0.3	0.5	0.8	1.0
α	mean	0.310	0.501	0.781	0.957
	bias	0.032	0.001	-0.023	-0.043
	rrmse	0.065	0.060	0.063	0.073
β	mean	0.973	1.003	1.029	1.046
	bias	-0.027	0.003	0.029	0.046
	rrmse	0.110	0.093	0.093	0.093
γ	mean	1.000	1.000	1.001	1.003
	bias	0.000	0.000	0.001	0.003
	rrmse	0.000	0.000	0.001	0.004

TABLE 3.3

Estimates of parameters derived from new ML approach for the three-parameter gamma distribution based on 1000 Monte Carlo simulations for various sample sizes

True Value	Estimated Parameter	Performance Criteria	Sample Size 50	100	200	300	400	500	600	700	800	900	1000
$\alpha = 2.0$	α	mean	1.862	1.889	1.934	1.948	1.960	1.964	1.974	1.981	1.983	1.983	1.985
		bias	-0.069	-0.055	-0.033	-0.026	-0.020	-0.018	-0.013	-0.010	-0.008	-0.009	-0.008
		rrmse	0.565	0.243	0.148	0.119	0.100	0.087	0.076	0.070	0.067	0.062	0.058
$\beta = 1.0$	β	mean	1.247	1.082	1.036	1.026	1.018	1.017	1.012	1.009	1.007	1.008	1.007
		bias	0.247	0.082	0.036	0.026	0.018	0.017	0.012	0.009	0.007	0.008	0.007
		rrmse	0.566	0.241	0.139	0.113	0.094	0.084	0.074	0.067	0.062	0.060	0.057
$\gamma = 1.0$	γ	mean	1.078	1.050	1.030	1.022	1.017	1.015	1.012	1.010	1.009	1.009	1.007
		bias	0.078	0.050	0.030	0.022	0.017	0.015	0.012	0.010	0.009	0.009	0.008
		rrmse	0.242	0.121	0.073	0.056	0.046	0.040	0.034	0.031	0.028	0.026	0.024
$\alpha = 0.8$	α	mean	0.662	0.728	0.766	0.778	0.783	0.787	0.789	0.791	0.792	0.794	0.794
		bias	-0.173	-0.090	-0.043	-0.028	-0.022	-0.017	-0.013	-0.012	-0.010	-0.008	-0.007
		rrmse	0.201	0.129	0.089	0.070	0.060	0.054	0.049	0.046	0.042	0.040	0.038
$\beta = 1.0$	β	mean	1.221	1.108	1.049	1.034	1.027	1.022	1.018	1.016	1.014	1.011	1.011
		bias	0.221	0.108	0.049	0.034	0.027	0.022	0.018	0.016	0.014	0.011	0.011
		rrmse	0.334	0.202	0.126	0.105	0.088	0.076	0.070	0.066	0.060	0.056	0.053
$\gamma = 1.0$	γ	mean	1.008	1.004	1.001	1.001	1.001	1.001	1.000	1.000	1.000	1.000	1.000
		bias	0.008	0.004	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000	0.000
		rrmse	0.013	0.006	0.002	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000

TABLE 3.4

Estimates of parameters derived from new ML approach for the three-parameter Weibull distribution based on 1000 Monte Carlo simulations with sample size n = 365 and $\beta = \gamma = 1$

Estimated Parameter	Performance Criteria	Shape Parameter α 1.5	2.0	3.0	4.0
α	mean	1.483	1.983	2.987	4.032
	bias	-0.012	-0.009	-0.004	0.008
	rrmse	0.047	0.058	0.092	0.137
β	mean	0.988	0.989	0.991	1.040
	bias	-0.012	-0.011	-0.009	0.040
	rrmse	0.043	0.045	0.073	0.577
γ	mean	1.008	1.010	1.008	0.985
	bias	0.008	0.009	0.008	-0.015
	rrmse	0.015	0.029	0.066	0.194

Note : The results from using Method 2 are virtually identical to those of Table 3.

TABLE 3.5

Estimates of parameters derived from new ML approach for the three-parameter Weibull distribution based on 1000 Monte Carlo simulations with sample size n = 365 and $\beta = \gamma = 1$

Estimated Parameter	Performance Criteria	Shape Parameter α 0.3	0.5	0.8	1.0
α	mean	0.305	0.501	0.794	0.983
	bias	0.016	0.003	-0.008	-0.017
	rrmse	0.041	0.039	0.039	0.042
β	mean	1.034	1.005	0.995	0.990
	bias	0.034	0.005	-0.005	-0.010
	rrmse	0.195	0.112	0.070	0.056
γ	mean	1.000	1.000	1.001	1.003
	bias	0.000	0.000	0.001	0.003
	rrmse	0.000	0.000	0.001	0.004

TABLE 3.6

Estimates of parameters derived from new ML approach for the three-parameter Weibull distribution based on 1000 Monte Carlo simulations for various sample sizes

True Value	Estimated Parameter	Performance Criteria	Sample Size 50	100	200	300	400	500	600	700	800	900	1000
$\alpha = 2.0$	α	mean	1.943	1.951	1.980	1.988	1.989	1.990	1.990	1.989	1.991	1.989	1.989
		bias	-.029	-.024	-.010	-.006	-.006	-.005	-.005	-.006	-.005	-.005	-.005
		rrmse	0.213	0.130	0.083	0.066	0.056	0.050	0.045	0.041	0.039	0.037	0.033
$\beta = 1.0$	β	mean	0.949	0.967	0.985	0.990	0.991	0.993	0.994	0.994	0.995	0.994	0.995
		bias	-.051	-.033	-.015	-.010	-.009	-.007	-.006	-.006	-.006	-.006	-.005
		rrmse	0.159	0.100	0.063	0.048	0.042	0.037	0.033	0.030	0.028	0.026	0.025
$\gamma = 1.0$	γ	mean	1.039	1.026	1.012	1.008	1.008	1.007	1.006	1.006	1.005	1.005	1.005
		bias	0.039	0.026	0.012	0.008	0.008	0.007	0.006	0.006	0.005	0.005	0.005
		rrmse	0.119	0.072	0.041	0.032	0.027	0.024	0.021	0.020	0.018	0.017	0.016
$\alpha = 0.8$	α	mean	0.735	0.769	0.786	0.792	0.795	0.796	0.797	0.797	0.798	0.798	0.798
		bias	-.081	-.039	-.017	-.010	-.007	-.005	-.004	-.004	-.003	-.003	-.003
		rrmse	0.119	0.078	0.054	0.044	0.038	0.034	0.031	0.029	0.028	0.026	0.024
$\beta = 1.0$	β	mean	0.942	0.972	0.987	0.992	0.995	0.998	0.998	0.998	0.998	0.999	0.999
		bias	-.058	-.028	-.013	-.008	-.005	-.002	-.002	-.002	-.002	-.001	-.001
		rrmse	0.183	0.131	0.094	0.075	0.066	0.060	0.054	0.050	0.046	0.043	0.041
$\gamma = 1.0$	γ	mean	1.008	1.004	1.001	1.001	1.001	1.001	1.000	1.000	1.000	1.000	1.000
		bias	0.008	0.004	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000
		rrmse	0.013	0.006	0.002	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000

TABLE 3.7

Estimates of parameters of two distributions and maximized log-likelihood values by various methods to fit nitrogen dioxide concentrations in the Museum and the Alphington stations, Melbourne, Australia

n	Data Station	Distribution	Method	α	β	γ	Max(logL)
271	Museum	Gamma	1,2,3,4	1.772	9.962	0.523	-586.21
		Weibull	1,2	1.336	3.749	0.565	-588.42
317	Alphington	Gamma	1	0.637	3.286	0.040	-525.33
		Weibull	1	0.773	1.840	0.040	-531.97

Note : For the data in the Alphington station, methods 2-4 failed to converge.

TABLE 3.8

Estimates of parameters and maximized log-likelihood values by new ML approach and MPS method to fit the three parameter Weibull distribution using the flood level data and the pollution data

n	Data Set	Method	α	β	γ	Max(logL)
20	Flood Level	ML	0.464	0.128	0.265	25.03
		MPS	1.310	0.202	0.244	16.36
20	Pollution	ML	0.430	4333	1364	-43.44
		MPS	0.950	6562	1085	-56.75

Chapter 4

Percentile Estimation: the Method of Moments versus Maximum Likelihood

4.1 Introduction

This chapter considers the problem of parametric fitting of probability distributions given a time series of observations by comparing the results of comprehensive Monte Carlo simulation experiments using the method of moments and maximum likelihood. From the analysis of air quality data sets collected at fixed sites over invariant averaging times, common distributional forms considered appropriate are the lognormal, gamma and Weibull distributions (e.g. Bencala and Seinfeld, 1976; Taylor et al. 1986), which have been discussed in previous chapters. The aim here is to summarise the data while obtaining accurate estimates of the percentiles required. For this application it must be assumed that air pollutant observations are independent, identically distributed random variables. Often this appears not too harsh an assumption in practice. Georgopoulos and Seinfeld (1982) noted in their review of the statistical distribution of air pollutant concentrations that the application of theoretical results derived for independent, identically distributed random variables produced satisfactory agreement with observations.

To fit air quality data, the upper percentiles of the distribution are the major concerns although fitting of the overall distribution is also very important. These

upper percentiles can generally be considered to consist of the 98-percentile and higher percentiles. This is the percentile range of concentrations to which most air quality standards refer. Particular emphasis is placed upon estimation of the maximum and second highest concentrations of a data set consisting of observations recorded over a fixed averaging time (sampling interval) of 24-hours and collected for a calendar year. However, the effect of other averaging times is investigated as well by varying the sample size in simulation experiments.

Little attention has previously been given directly to quantify performance for fitting high percentiles in the statistical literature. In general applications, the method of maximum likelihood and the method of moments are commonly used in estimation. It is well understood that the method of maximum likelihood provides more efficient estimates of the parameters than the method of moments at large sample sizes (Kendall and Stuart, 1979). However, the relevant performance of the fit will change according to the parent parameter values, the percentile of interest, the sample size and the selected performance criterion. This problem shall be examined here.

4.2 General Properties of the Methods of Moments and Maximum Likelihood related to Air Quality Application

As mentioned before, the methods of moments (MT) and maximum likelihood (ML) are the most popular statistical techniques for estimating parameters of a distribution. Even though numerous papers and text books in the statistical literature have discussed these two methods, most authors have focussed only on the theoretical aspects (Kendall and Stuart, 1979). Practical aspects have received little attention, especially investigation of the performance of these two methods by simulation experiments and fitting real data. Employing these methods to study air pollution data, it is necessary to re-examine their properties in relation to the requirements raised in applications. Particularly, there is concern with their behaviour in predicting upper percentiles of

distributions.

The method of moments is a traditional approach used for estimating the parameters of distributions. This method equates the sample moments with the corresponding population values, an equivalence which becomes true as the sample size tends to infinity. Generally, the moment generating function or characteristic function can be used to generate theoretical moments, and the sample moments can be easily obtained by simple calculations from the data. Compared to the method of maximum likelihood, the method of moments is simpler to use and its estimates are consistent. However, as a distribution departs considerably from normality, such as those of the Pearsonian type, the MT method becomes very inefficient in estimating parameters of the distribution (Cramer, 1946). In many cases, the estimates of the method of moments have large variance, and they are not unbiased (Kendall and Stuart, 1979).

The method of maximum likelihood is a more sophisticated technique in estimating parameters of distributions. Since it maximizes the likelihood function of the distributions, the ML estimates have many optimal properties. Asymptotically, the estimates of ML follow the normal distribution with the true value as mean, and a variance achieving the Cramer-Rao bound (Kendall and Stuart, 1979). Also such estimates are asymptotically efficient, sufficient and consistent. Hence, they play a very important role in statistical inference, such as in hypothesis testing. On the other hand, because the ML method usually involves a numerical procedure to estimate the parameters iteratively and simultaneously, it can provide convergence difficulties. Personal experience suggests that such difficulties occur particularly when the number of parameters of the distribution is more than 3 or the sample size is below 30. On the contrary, the method of moments does not suffer from this convergence problem.

In practice, it is desirable to use a simple method as long as its error in estimation can be tolerated. There is a lack of knowledge in the literature, indicating the exact quantitative difference between the MT and ML methods in estimating parameters of probability distributions. This may be obtained via extensive investigations and

Monte Carlo computer simulations. The aim here is to provide such information to the practitioner, so that the decision of which method should be used can be made based upon the results given.

Even less attention has been given to upper percentile estimation by MT and ML methods. In recent years, it has become very important since many environmental regulations, such as air quality standards, are stated specifically in terms of upper percentiles. It is important to predict these percentiles as accurately as possible because the results have serious practical consequences.

It is well understood that the behaviour of the sampling distribution differs over the entire range of percentiles. Corresponding to the lower, middle, and upper percentiles, the standard errors of the samples are not identical. For illustrative purposes, consider the following simple analysis. Suppose $x_1, \ldots, x_n$ is a random sample of n observations in ascending order of magnitude, i.e., it can be expressed as

$$x_1 < x_2 < \ldots < x_n$$

and $f(x)$ is the parent probability density function of the distribution. Based upon the theory of quantiles (Kendall and Stuart, 1979), the variance of x_i can be calculated as

$$var(x_i) = \frac{pq}{nf_i^2} \tag{4.1}$$

where $p = F(x_p)$. For any p, $0 < p < 1$, and $p + q = 1$. The quantity f_i is the ordinate of the parent density distribution (Kendall and Stuart, (1979, p. 237)). Considering the entire range of percentiles, the relationship between $var(x_i)$ and p, in terms of (4.1), can be shown in Fig 4.1 for the three-parameter gamma distribution. Generally, $var(x_i)$ becomes larger as p is increased. A similar analysis can be used to obtain the relationship between $var(x_i)$ and f_i shown in Fig 4.2. $var(x_i)$ is high for low value of f_i at the upper percentiles. Hence, these plots clearly illustrate that at upper percentiles, where f_i has the lowest value, the largest variance in sampling is focussed. Obviously the upper percentiles yield the largest sampling errors for positively skewed distributions with long tails.

It is easy to see the effects of sample upper percentiles on the ML and MT estimation methods. Estimation by maximum likelihood depends on the values of f_i to maximise the likelihood function

$$L = L(x_1, \ldots, x_n; \theta) = max \prod_{i=1}^{n} f(x_i; \theta) \tag{4.2}$$

in order to obtain estimates of θ where the maximized value of the function occurs. The middle range of percentiles with high values of f_i has greatest influence on the estimation results, while changes in the upper percentiles have less influence on the estimation. In other words, this method is not especially sensitive to upper percentile variation because f_i is very low there. On the other hand, the method of moments generally employs the first two or three moments of the sampling distribution, which heavily depend on the values of x_i rather than f_i. High values of x_i at the upper percentiles have strong influence on the estimation results, particularly when using high moments. Hence, the method of moments is more sensitive to sampling errors in the upper percentiles, and it will simply fit the data better there (but not necessarily the parent distribution) than the method of maximum likelihood.

4.3 Estimation by the Method of Moments

Estimation by maximum likelihood has been discussed in the previous chapter. Here the principle of the method of moments is outlined. Using the characteristic function, formulas for parameter estimation of the two- and three-parameter gamma, Weibull and lognormal distributions are obtained.

4.3.1 The Characteristic Function and Moments

In general, the distribution function is closely related to the characteristic function. The latter has many useful and important properties and plays a central role in statistical theory. Like moment generating functions, it can be used to derive properties of theoretical moments which widely apply to statistical inference and sampling analysis.

Another important aspect of the characteristic function is that it uniquely determines a distribution function since the reciprocal relationship between distribution density function and characteristic function exists. From the limiting properties of distribution and characteristic functions which are demonstrated by the uniqueness theorem and the continuity theorem (Mann, 1974), the characteristic function of the joint distribution of a number of independent variables has provided a fundamentally important result in the theory of sampling (Kendall and Stuart, 1979).

Let $F(x)$ denote a one-dimensioned distribution function and t a real number, the characteristic function corresponding to $F(x)$ is defined as

$$\phi(t) = \int_{-\infty}^{\infty} e^{itx} dF(x) \tag{4.3}$$

where $\phi(t)$ is generally a complex-valued function of t and always has $\phi(0) = 1$ and $e^{itx} = cos\, tx + i\, sin\, tx$. Differentiating (4.3) k times with respect to t,

$$\phi^{\nu}(t) = i^{\nu} \int_{-\infty}^{\infty} x^{\nu} e^{itx} dF(x) \tag{4.4}$$

Hence, the characteristic function can be expressed as a MacLaurin's series in the neighbourhood of $t = 0$ by

$$\phi(t) = 1 + \sum_{1}^{k} \frac{a_{\nu}}{\nu!}(it)^{\nu} + 0(t^{k}) \tag{4.5}$$

where the error term tends to zero as $t \rightarrow 0$. (Cramer, 1946). Here a_{ν} is just the moment of order ν of the distribution, having the form

$$a_{\nu} = \int_{-\infty}^{\infty} x^{\nu} dF(x) \tag{4.6}$$

The first moment a_1 is the mean of the distribution, denoted also by m. Based on the mean of the distribution, the central moments can be defined by

$$\mu_{\nu} = \int_{-\infty}^{\infty} (x - m)^{\nu} dF(x) \tag{4.7}$$

Using (4.7), the relation between moments and central moments are easily found as the following (Cramer, 1946)

$$\mu_0 = 1, \tag{4.8}$$

$$\mu_1 = 0, \tag{4.9}$$

$$\mu_2 = a_2 - m^2, \tag{4.10}$$

$$\mu_3 = a_3 - 3ma_2 + 2m^3, \tag{4.11}$$

$$\mu_4 = a_4 - 4ma_3 + 6m^2a_2 - 3m^4, \tag{4.12}$$

.....

where μ_2 is the variance of the distribution, denoted by σ^2 (σ is the standard deviation). In general applications, the first moment, and second and third central moments are used quite often. To distinguish between symmetric and skewed distributions, the coefficient of skewness is commonly introduced as

$$g = \frac{\mu_3}{\sigma^3} \tag{4.13}$$

The skewness is used to measure the departure of a skewed distribution from a symmetric distribution. Positive skewness shows that the frequency curve forms a long tail in the positive direction; similarly negative skewness will lead to extension in the negative direction. Normally, the positive-skewed distribution is the usual case in air pollution applications and will be the major focus in this chapter.

In the next section, the three-parameter gamma distribution is used for illustrative purposes to show how to obtain moments from the characteristic function. Subsequently, the moment estimators for the two- and three-parameter gamma, Weibull and lognormal distributions will be given, as they are the most commonly used forms in the air pollution and much of the statistical literature.

4.3.2 The Moments Estimators for Two- and Three-parameter Gamma,Weibull and Lognormal Distributions

As discussed above, the characteristic function can be used to derive the theoretical moments for a particular distribution function of interest. Based on the central limit theorem, the sampling moments will tend to the theoretical or parent moments as sample size tends to infinity. For finite samples the assumption of equality is made to derive the distribution parameters. As an illustrative example, the three-parameter gamma distribution is now employed to demonstrate the derivation of moment estimators. In the previous chapter, the density function of the three-parameter gamma distribution was given in (3.1). Inserting this in equation (4.3), the characteristic function for the three-parameter gamma becomes

$$\phi(t) = \frac{1}{\beta\Gamma(\alpha)} \int_0^\infty e^{itx} (\frac{x-\gamma}{\beta})^{\alpha-1} exp[-(\frac{x-\gamma}{\beta})] dx \tag{4.14}$$

where the lower bound of the interval of integration is zero according to the definition of the gamma distribution. Using the substitution $u = x - \gamma$ and re-arranging, the equation becomes

$$\phi(t) = \frac{e^{it\gamma}}{\beta^\alpha\Gamma(\alpha)} \int_0^\infty u^{\alpha-1} e^{-(\frac{1}{\beta}-it)u} du \tag{4.15}$$

Substituting again with $v = (\frac{1}{\beta} - it)u$, the form of the characteristic function becomes

$$\begin{aligned} \phi(t) &= \frac{1}{\beta^\alpha\Gamma(\alpha)} (\frac{1}{\beta} - it)^{-\alpha} \int_0^\infty v^{\alpha-1} e^{-v} dv \\ &= e^{it\gamma}(1 - it\beta)^{-\alpha} \end{aligned} \tag{4.16}$$

After expansion of (4.16) in a MacLaurin's series in the neighbourhood of $t = o$ and comparing it with (4.5), the ν-th moments for the three-parameter gamma distribution can be obtained as

$$a_1 = \gamma + \beta\alpha \tag{4.17}$$

$$a_2 = \gamma^2 + 2\gamma\beta\alpha + \beta^2\alpha(\alpha+1) \tag{4.18}$$

$$a_3 = \gamma^3 + 3\gamma^2\beta\alpha + 3\gamma\beta^2\alpha(\alpha+1) + \beta^3(\alpha+1)(\alpha+2) \tag{4.19}$$

.....

For three-parameter distributions, the skewness is often used in deriving the moment estimators. Inserting (4.17), (4.18) and (4.19) into (4.10) and (4.11) and rearranging, the second and third central moments become

$$\mu_2 = \alpha\beta^2 \tag{4.20}$$

$$\mu_3 = 2\alpha\beta^3 \tag{4.21}$$

From the definition of skewness in (4.13),

$$g = \frac{2\alpha\beta^3}{(\alpha\beta^2)^{\frac{3}{2}}} = \frac{2}{\sqrt{\alpha}} \tag{4.22}$$

Therefore, the shape parameter α can be obtained by

$$\alpha = \frac{4}{g^2} \tag{4.23}$$

By using the variance μ_2, the scale parameter can be determined from

$$\beta = \sqrt{\frac{\mu_2}{\alpha}} \tag{4.24}$$

When the first two parameters are calculated, the remaining location parameter γ can be directly obtained from the definition of the mean as

$$\gamma = a_1 - \beta\alpha \tag{4.25}$$

Following a similar procedure to the above, the moment estimators for the two-parameter gamma, two- and three-parameter Weibull and lognormal distributions can also be derived. The sample moments are equated to their theoretical counterparts. For a sample $x_1, x_2, \ldots, x_n$ of n independently and identically distributed random observations, the sample moments can be expressed as

$$a_\nu = \frac{1}{n}\sum_{i=1}^{n} x_i^\nu \tag{4.26}$$

and the central moments as

$$\mu_\nu = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^\nu \tag{4.27}$$

where $\bar{x}$ is the sample mean. Note that normal practice is to use $s = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$ instead of the sample variance because of its unbiased property. Using the sample moments, method of moment estimators of the parameters of the six distributions considered in this thesis can be given as follows (Johnson and Kotz (1970), Dubey (1966) and Aitchison and Brown (1975)):

three-parameter Gamma:

$$\hat{g} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2]^{\frac{3}{2}}} \tag{4.28}$$

$$\hat{\alpha} = \frac{4}{\hat{g}^2} \tag{4.29}$$

$$\hat{\beta} = \sqrt{\frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\hat{\alpha}}} \tag{4.30}$$

$$\hat{\gamma} = \bar{x} - \hat{\beta}\hat{\alpha} \tag{4.31}$$

two-parameter Gamma:

$$\hat{\alpha} = \frac{(n-1)\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{4.32}$$

$$\hat{\beta} = \frac{\bar{x}}{\hat{\alpha}} \tag{4.33}$$

three-parameter Weibull:

$$\hat{g} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2]^{\frac{3}{2}}} = \frac{\Gamma(\frac{3}{\hat{\alpha}} + 1) - 3\Gamma(\frac{2}{\hat{\alpha}} + 1)\Gamma(\frac{1}{\hat{\alpha}} + 1) + 2\Gamma^3(\frac{1}{\hat{\alpha}} + 1)}{[\Gamma(\frac{2}{\hat{\alpha}} + 1) - \Gamma^2(\frac{1}{\hat{\alpha}} + 1)]^{\frac{3}{2}}} \tag{4.34}$$

$$\hat{\beta} = \sqrt{\frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\Gamma(\frac{2}{\hat{\alpha}} + 1) - \Gamma^2(\frac{1}{\hat{\alpha}} + 1)}} \tag{4.35}$$

$$\hat{\gamma} = \bar{x} - \hat{\beta}\Gamma(\frac{1}{\hat{\alpha}} + 1) \tag{4.36}$$

two-parameter Weibull:

$$\frac{(n-1)\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{\Gamma^2(\frac{1}{\hat{\alpha}} + 1)}{\Gamma(\frac{1}{\hat{\alpha}} + 2) - \Gamma^2(\frac{1}{\hat{\alpha}} + 1)} \tag{4.37}$$

$$\hat{\beta} = \frac{\bar{x}}{\Gamma(\frac{1}{\hat{\alpha}} + 1)} \tag{4.38}$$

three-parameter Lognormal:

$$\hat{g}^2 = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{[\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2]^{\frac{3}{2}}} = (e^{\hat{\alpha}^2} - 1)(e^{\hat{\alpha}^2} + 2) \tag{4.39}$$

$$\hat{\alpha} = \{\log\{[1 + \frac{1}{2}\hat{g}^2 + \sqrt{(1 + \frac{1}{2}\hat{g}^2)^2 - 1}]^{\frac{1}{3}} + [1 + \frac{1}{2}\hat{g}^2 - \sqrt{(1 + \frac{1}{2}\hat{g}^2)^2 - 1}]^{\frac{1}{3}} - 1\}\}^{\frac{1}{2}} \tag{4.40}$$

$$\hat{\beta} = \frac{1}{2}\log[\frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{e^{\hat{\alpha}^2}(e^{\hat{\alpha}^2} - 1)}] \tag{4.41}$$

$$\hat{\gamma} = \bar{x} - exp(\hat{\beta} + \frac{1}{2}\hat{\alpha}^2) \tag{4.42}$$

two-parameter Lognormal:

$$\hat{\alpha} = \sqrt{\log[\frac{1}{n-1}\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\bar{x}^2} + 1]} \tag{4.43}$$

$$\hat{\beta} = \log \bar{x} - \frac{1}{2}\hat{\alpha}^2 \tag{4.44}$$

Note that (4.34) and (4.38) define $\hat{\alpha}$ implicitly.

The two- and three-parameter gamma, and two-parameter lognormal distributions require only simple calculations. For the two- and three-parameter Weibull and three-parameter lognormal distributions, the method of moment estimators are obtained by iterative numerical methods.

It should be appreciated that there will be different solutions for parameter estimates if different moments are employed, such as higher moments. In other words, the method of moment estimators are not unique. Thus, using different moments can generate quite different performance. However for the purpose of comparing performance with the maximum likelihood estimator, common practice is followed here in using the first two moments for two-parameter distributions and the first three moments for 3 parameter distributions.

4.4 Loss Functions

In order to assess the performance of the methods of moments and maximum likelihood in predicting the percentiles of a distribution, loss functions recommended for assessing air quality models are used (see Fox (1981)). These functions are given in equations (3.18) and (3.19), and are reproduced below for convenience

$$BIAS(q) = \frac{1}{N}\sum_{i=1}^{N}(\frac{\hat{q}_i - q}{q}) \tag{4.45}$$

$$RRMSE(q) = [\frac{1}{N}\sum_{i=1}^{N}(\frac{\hat{q}_i - q}{q})^2]^{0.5} \tag{4.46}$$

where N is the number of replications in the Monte Carlo experiments. For present purposes, q denotes the parent parameter values and upper percentile quantities of the underlying distributions. The upper percentile quantities may be observed samples

or underlying values of the parent distribution as errors of departure from both are calculated.

4.5 Monte Carlo Experiments

To undertake this assessment of fitting percentiles of the gamma, Weibull and lognormal distributions for both the sample and parent cases, simulation over an extensive range of possible cases was undertaken. Results are reported in tabular form for shape parameters of value 2, 4 and 6 for the gamma distribution, of value 2, 3 and 4 for the Weibull distribution and of value 0.5, 0.7 and 0.9 for the lognormal distribution. These values span a large range of shapes which may arise in the analysis of air pollutant concentrations. In order to assess the sensitivity of results to the location parameter, the values of 1, 3 and 5 were selected for each distribution. The same values of 1, 3 and 5 are also set for the scale parameter. For all parameter sets in the tables and figures, one thousand simulations are used. The major sample size used is $n = 365$, since it represents a common case, namely a full year of 24-hourly average observations. However, a range of sample sizes from 100 to 1000 is also considered to show that the qualitative conclusions derived for sample size $n = 365$ are applicable for other sample sizes.

Through the Monte Carlo simulations, estimates of both the parameters and percentiles are obtained. In the percentile estimation, the emphasis is placed on the following upper percentiles: the 98-percentile (98%), second maximum (MAX2) and maximum (MAX1) percentiles. The predicted values are compared with both parent and sampling distributions. In the case of the parameter estimation, only deviations from the parent (true) value can be calculated. Note that the new methodology of the method of maximum likelihood proposed in Chapter 3 is used which can provide satisfactory ML estimates in a computationally efficient manner for both the 3-parameter gamma and Weibull distributions. When the sampling distribution is quite skewed, it was found that the classical maximum likelihood method for estimation of the 3-

parameter lognormal distribution, which involves solving the first derivative equations, has difficulty in converging. Using the new approach in Chapter 3, this problem can also be avoided. Therefore, in this chapter, the new ML method is used for the estimation of the parameters of the three-parameter gamma, Weibull and lognormal distributions.

As is always the case in this thesis, the random sample generators used for the Monte Carlo experiments are DRNGAM, DRNWIB and DRNLNL for the gamma, Weibull and lognormal distributions, respectively. These are available as subroutines in the International Mathematical and Statistical Library (IMSL) in version 1.0 (April 1987). The same seed number (1234) is used to obtain the first random sample of the first of the 1000 replications. Varying the initial seed produces similar results to those reported in the tables. For the maximum likelihood estimation, a golden section search algorithm is used to obtain the roots of equations and the final estimate of the root was accepted when the relative error between two successive approximations was less than 10^{-6}. A VAX 8700 mainframe computer was used.

4.6 Monte Carlo Results

4.6.1 Estimation for the Three-parameter Gamma Distribution

Table 4.1 lists the MEAN, BIAS and RRMSE for parameter estimates of the gamma distribution derived using the method of maximum likelihood and method of moments. As would be expected from theoretical considerations (Kendall and Stuart, 1979), Table 4.1 demonstrates that the method of maximum likelihood yields improved estimates of the parameters of the distribution, in terms of BIAS and RRMSE, over the method of moments. Note that parameter errors for shape and scale parameters are almost identical for different scale and location values of the parent distribution. This is not the situation for percentile errors reported in Tables 4.2 and 4.3. Generally, the relative performance of the two methods converges as the skewness decreases (shape increases).

For the three-parameter gamma distribution, Table 4.2 lists estimates of the BIAS

and RRMSE of predictions of the maximum value (MAX1), the second highest value (MAX2) and the 98-percentile (98%). These error criteria are based upon deviations from the true parent distributional percentiles. Again the method of maximum likelihood yields consistently better RRMSE values than the method of moments. The estimates of BIAS show that for the gamma distribution the method of moments produces estimates of the percentiles with slightly smaller BIAS than the maximum likelihood method. The important point is that for both methods the BIAS is much lower than the corresponding RRMSE, often around 10 per cent of the RRMSE value. A comparison of the results of Tables 4.2 and 4.1 is interesting. The method of moments is seen to produce much higher RRMSE values for the parameter estimates than the maximum likelihood method. In Table 4.2 the improved accuracy of the method of maximum likelihood over the method of moments is relatively lower. Thus, the method of moments provides a much better fit to the upper percentiles of the distribution than the estimates of parameters, considered as an isolated factor, would indicate.

Table 4.3 is based upon similar calculations to those for Table 4.2 except that the performance criteria are based upon deviations from the relevant sample percentile in each of the 1000 Monte Carlo simulations. The results of Table 4.3 show that the method of moments produces lower BIAS and RRMSE values for the three percentiles. Hence the method of moments provides a closer description of sample upper percentiles than the method of maximum likelihood.

The results of Tables 4.2 and 4.3 also yield a consistent pattern in the percentile error sensitivity to location, scale and shape parameters. Consider the effect of changes in location parameter first. For either method a larger location parameter yields a general decrease in BIAS accompanied always by a decrease in RRMSE from the parent and sample percentiles. This pattern holds for all shape parameterisations and for the three percentiles investigated. The reverse effect can be observed for changes in the scale parameter. BIAS and RRMSE will increase when the value of the scale parameter increases. For the effect of shape parameter changes, an equally consistent pattern can be seen whichever the method, percentile, scale and location parameter values. In this

case the BIAS and RRMSE generally drop as skewness decreases.

In order to investigate the performance of the method of moments and method of maximum likelihood over a range of percentiles, the RRMSE for the full range of percentiles was evaluated. Again $N = 1000$ and $n = 365$ were chosen for the Monte Carlo study. Results are plotted for gamma parameter values of $(\alpha, \beta, \gamma) = (2,1,1)$. The other parameter sets yield analogous behaviour. Figure 4.3 presents these results for RRMSE deviations from the parent percentiles. Figure 4.4 presents the corresponding results for RRMSE, that is for deviations from the sample percentiles. Figure 4.3 shows that the method of maximum likelihood produces the lower RRMSE (from the parent percentile) values over all percentiles. However, Figure 4.4 indicates that the fit to the observed percentiles by the method of moments is superior for the upper percentiles. This is not the case for the lower percentiles. Thus there is a cross-over percentile above which the method of moments becomes superior.

The effect of sample size was also investigated by evaluating RRMSE at the 98-percentile (98%) over the range of sample sizes $n = 50, 100(100)1000$. Figure 4.5 presents the indicative results for deviations from the true value. The parent parameter values used to illustrate are $(\alpha, \beta, \gamma) = (2,1,1)$ for the gamma distribution. Figure 4.6 shows the corresponding results for deviations from the sample 98-percentile. Figure 4.5 demonstrates the expected result that the method of maximum likelihood yields the lowest values of RRMSE relative to the parent percentile for the whole range of sample sizes considered here. By contrast, Figure 4.6 illustrates that for the gamma distribution the method of moments has lower deviations for the estimates of sample percentiles. However the method of maximum likelihood appears to be approaching the method of moments in accuracy as the sample size increases.

4.6.2 Estimation for the Three-parameter Weibull Distribution

Comparison of estimates of both parameters and percentiles by the methods of moments and maximum likelihood for the three-parameter Weibull distribution yields

qualitatively similar patterns to the results above for the gamma distribution. These are shown in Tables 4.4 to 4.6. In Table 4.4, the MEAN, BIAS and RRMSE for parameter estimates of the Weibull distribution are derived by using the method of maximum likelihood and method of moments. Again the method of maximum likelihood yields improved estimates of the parameters of the distribution, in terms of BIAS and RRMSE, over the method of moments. Parameter errors for shape and scale parameters remain almost identical for different scale and location values of the parent distribution.

Table 4.5 reports estimates of the BIAS and RRMSE of predictions of the three upper percentiles based upon deviations from the true parent percentiles. The method of maximum likelihood yields consistently better RRMSE values than the method of moments. Again, as been seen in the three-parameter gamma case, the estimates of BIAS show that for the Weibull distribution the method of moments produces estimates of the percentiles with slightly smaller BIAS than the maximum likelihood method.

The results of estimation Based upon deviations from the relevant sample percentile, are obtained which are reported in Table 4.6. The method of moments has lower BIAS and RRMSE values for the three percentiles investigated. It demonstrates that the method of moments provides a better fit of sample upper percentiles than the method of maximum likelihood.

For the three-parameter Weibull distribution, the percentile errors are also sensitive to location, scale and shape parameters. When the location parameter changes, BIAS and RRMSE generally decrease for both the parent and sample percentiles, but an increase in the scale parameter value leads to an increase in BIAS and RRMSE. Similar to the three-parameter gamma case, the effects of shape parameter changes are significant whatever the method, percentile, scale and location parameter values.

As with the gamma distribution, an investigation was undertaken for examining the performance of the method of moments and method of maximum likelihood over a range of percentiles for the three-parameter Weibull distribution. Still using $N = 1000$

and $n = 365$ for the Monte Carlo study, the RRMSE for the full range of percentiles was evaluated. In illustration, three parameter values of $(\alpha, \beta, \gamma) = (2, 1, 1)$ are used. In Figure 4.7, RRMSE is plotted for deviation from the parent percentiles. The method of maximum likelihood produces the lower RRMSE values over percentiles in terms of the parent distribution. Figure 4.8 is the corresponding plot for deviations from the sample percentiles. It shows that the fit to the observed percentiles by the method of moments is superior to that obtained using the method of maximum likelihood. It should be noted that the method of moments performs better here than the method of maximum likelihood over a much wider range of percentiles than was found for the three-parameter gamma distribution in Figure 4.6.

The effects of sample size are shown in Figures 4.9 and 4.10, which investigate RRMSE at the 98-percentile (98%) over the range of sample sizes $n = 50, 100(100)1000$. For indicative results, the parent parameter values used are $(\alpha, \beta, \gamma) = (2,1,1)$ for the Weibull distribution. Figure 4.9 provides the expected result that the method of maximum likelihood yields the lowest values of RRMSE relative to the parent percentile for the whole range of sample sizes considered here. However, the corresponding results for deviations from the sample 98-percentile in Figure 4.10 reverses this situation.

4.6.3 Estimation for the Three-parameter Lognormal Distribution

The same investigation procedure was applied for estimating both parameters and percentiles for the three-parameter lognormal distribution by the methods of moments and maximum likelihood. The results are shown in Tables 4.7 to 4.9 which indicate a qualitatively similar pattern of the results as with the previous two distributions. Note that the lognormal distribution has an apparent opposite behaviour to the other two as the shape parameter changes. This is because skewness increases for the lognormal distribution as the value of shape parameter increases. The results of applying the method of maximum likelihood and method of moments to estimate the parameters of the lognormal distribution are given in Table 4.7. In terms of BIAS and RRMSE,

the method of maximum likelihood yields improved estimates of the parameters of the distribution over the method of moments. It should be noted that the relative performance in terms of RRMSE differences between the method of moments and method of maximum likelihood is much worse than for the gamma and Weibull cases. Again, parameter errors for shape and scale parameters remain almost identical for different scale and location values of the parent distribution.

In Table 4.8, estimates are given of the BIAS and RRMSE for predictions of the three upper percentiles, based upon deviations from the true parent distributional percentiles. The method of maximum likelihood yields consistently better RRMSE values than the method of moments. Unlike the two previous distributions, the method of moments produces estimates of the percentiles with both larger BIAS and RRMSE than the maximum likelihood method.

Calculations of the deviations from the relevant sample percentiles (Table 4.9) show that the method of moments has lower BIAS and RRMSE values for the three percentiles. It provides further evidence that the method of moments yields a better fit of sample upper percentiles than the method of maximum likelihood.

The percentile errors for the three-parameter lognormal distribution also vary with the change of location, scale and shape parameters. BIAS and RRMSE generally decrease (for deviations from both the parent and sample percentiles) as the location parameter is increased. On the other hand, increases in the value of the scale parameter causes increases in BIAS and RRMSE, and it seems that the proportional increase is much larger than for the other two distributions. Changes in shape parameter affect both BIAS and RRMSE whatever the method, percentile, scale and location parameter values.

The RRMSE were calculated for the full range of percentiles of the three-parameter lognormal distribution. To illustrate the results, three parameter parent values of $(\alpha, \beta, \gamma) = (0.9,1,1)$ are used. Figure 4.11 presents deviations from the parent percentiles which shows that the method of maximum likelihood produces the lower

RRMSE (from the parent percentile) values over all percentiles. In Figure 4.12, it is seen that the fit to the observed percentiles by the method of moments is superior to those predicted by the method of maximum likelihood for the upper percentiles, if errors is based upon deviations from the samples. Similar to the three-parameter gamma distribution, there is a cross-over percentile above which the method of moments becomes superior.

The same parent parameter sets and the 98 percentile was used to illustrate results of comparing the method of maximum likelihood and method of moments. The results are demonstrated in Figure 4.13. In contrast, Figure 4.14 shows the corresponding results for deviations from the sample 98-percentile when the method of moments produces lower deviations when estimating sample percentiles.

4.6.4 Estimation for the Two-parameter Gamma, Weibull and Lognormal Distributions

Similar investigations were also undertaken for estimating of both parameters and percentiles by the methods of moments and maximum likelihood for the two-parameter gamma, Weibull and lognormal distributions. In principle, the results are qualitatively very close to those of their three-parameter versions. For the three two-parameter models, the method of maximum likelihood always yields improved estimates of the parameters of the distribution over the method of moments in terms of BIAS and RRMSE. Also, being similar to the three-parameter models, the method of maximum likelihood yields consistently better RRMSE values of parent percentiles than the method of moments. For deviations from the sample percentiles, the results demonstrate that the method of moments provides a better fit of sample upper percentiles than the method of maximum likelihood.

Considering the effects of changes in parameter values, the shape parameter becomes the only factor to influence RRMSE since the location for two-parameter version is always zero, and RRMSE values remain almost identical as the scale parameter changes. These results are not given in this chapter. The interested readers should refer

to the paper by Bai and Taylor (1986) for the two-parameter gamma distribution and a forthcoming paper by Bai for the two-parameter Weibull and lognormal distributions.

4.7 Concluding Remarks

Probability distributions can be employed in the study of air quality data to overcome many of the problems of managing large data bases with possible information deficiencies. A data set can be reduced to just a few parameters if an appropriate distributional function can be identified. Probability distributions can also be employed to produce estimates of properties of the probability density function when the sample contains errors and missing data.

The results presented in this section demonstrate that, where gamma, Weibull or lognormal description of the raw data set is required, the method of moments provides more accurate estimates of the highest and second highest sample concentrations than maximum likelihood for the parameter space investigated. For the gamma and lognormal distributions and sample size $n = 365$, this result holds from MAX1 to the sample 98-percentile, but this is at the expense of poorer estimates at the lower sample percentiles. Table 4.3 shows for the parameter space investigated that the improvement obtained by using the method of moments is well worth having for the sample maximum and second highest value and this is doubly important if the distribution is highly skewed. For example, the worst root mean square error is 33.3 per cent for the maximum likelihood method but reduces to 14.4 per cent for the method of moments. The method of moments also offers the advantage of yielding numerically simpler solutions than that of maximum likelihood.

However, the method of maximum likelihood should be employed when the most likely (true) estimate is desired. This method provides reasonable estimates over the entire range of percentiles of the parent distribution. The method of moments is much more sensitive to the sample data and tends to produce estimates which weight the upper percentiles of the sample in favour of the entire data set. Table 4.2, for example,

shows that the improvement obtained by using maximum likelihood to reduce random error is well worth the effort in all cases for the gamma distribution. The worst root mean square error (for high skewness) in estimating the underlying maximum value is just 1 per cent compared to 23.9 per cent for the method of moments.

Finally, as the method of maximum likelihood does not weight the largest observations as significantly as the method of moments, the method of maximum likelihood will be far less sensitive to outliers produced through systematic, experimental, data handling or other errors which may arise in the course of extensive routine and experimental monitoring programs. This is of particular practical importance in the management of air quality, where, for example, the prediction of extreme pollutant concentrations plays an important role in the consideration and implementation of pollution control strategies.

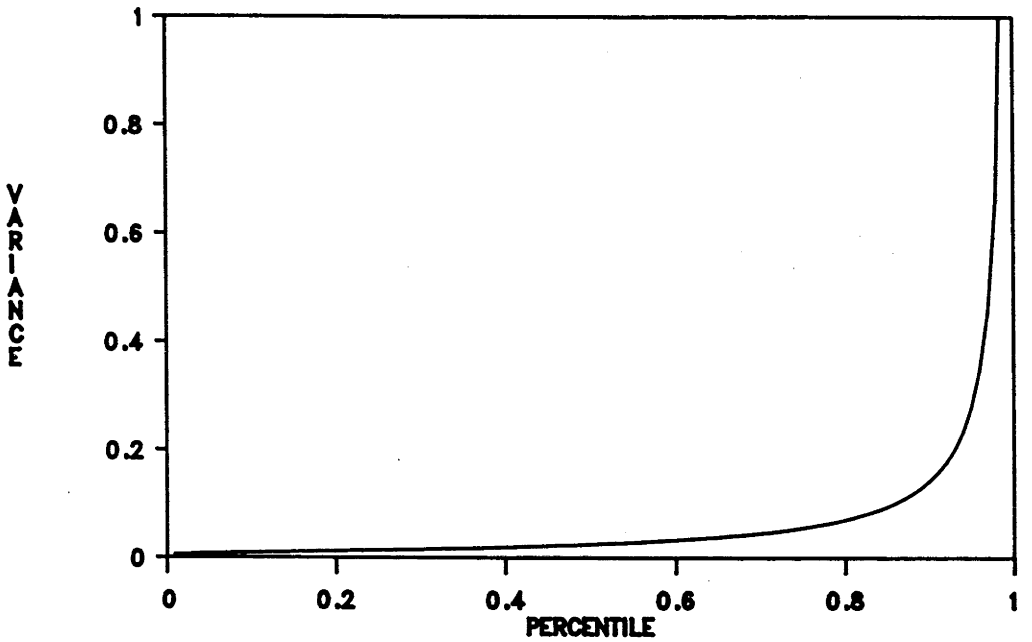

Figure 4.1: The variance with respect to each percentile for the 3-parameter gamma distribution $(\alpha, \beta, \gamma) = (2, 1, 1)$ and sample size $n = 100$

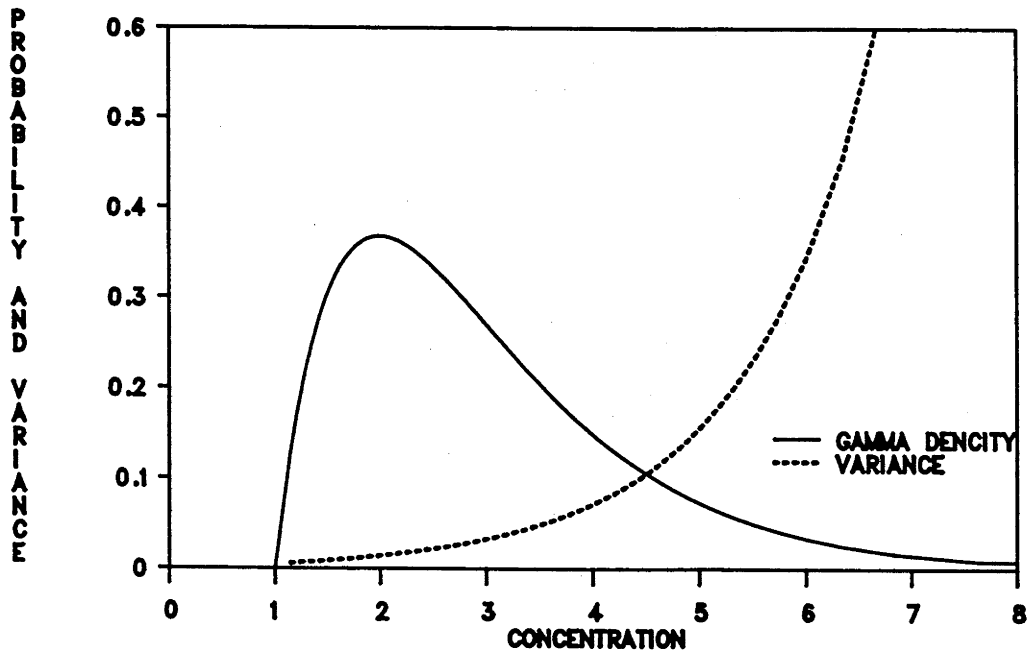

Figure 4.2: The variance with respect to each value f_i of the density distribution for the 3-parameter gamma distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$ and sample size $n = 100$

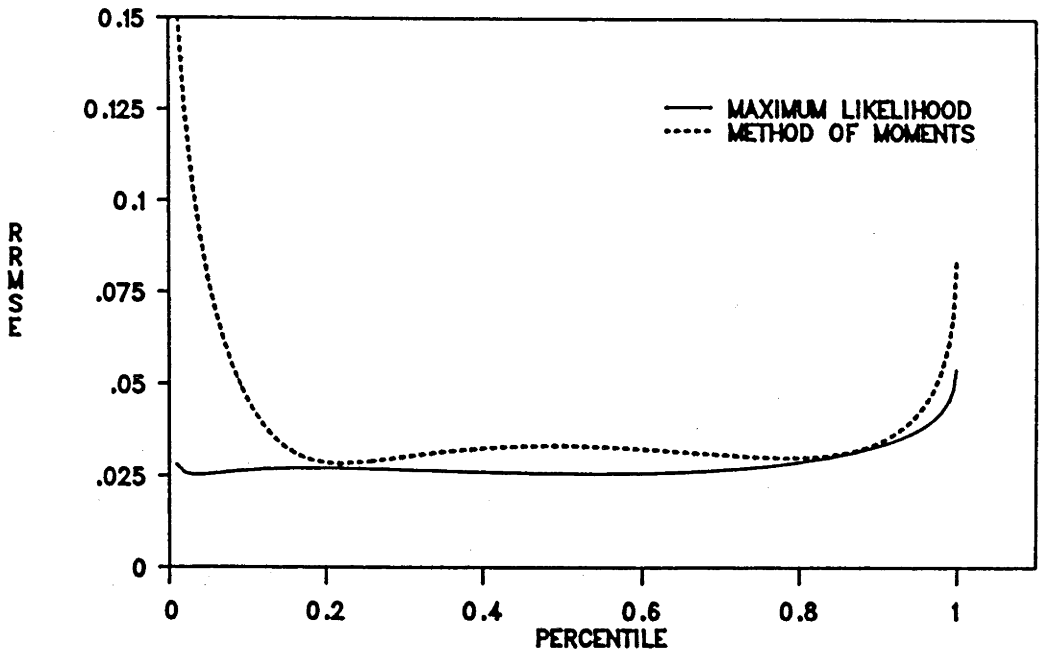

Figure 4.3: The RRMSE values with respect to each true percentile for the 3-parameter gamma distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$ and $n = 365$

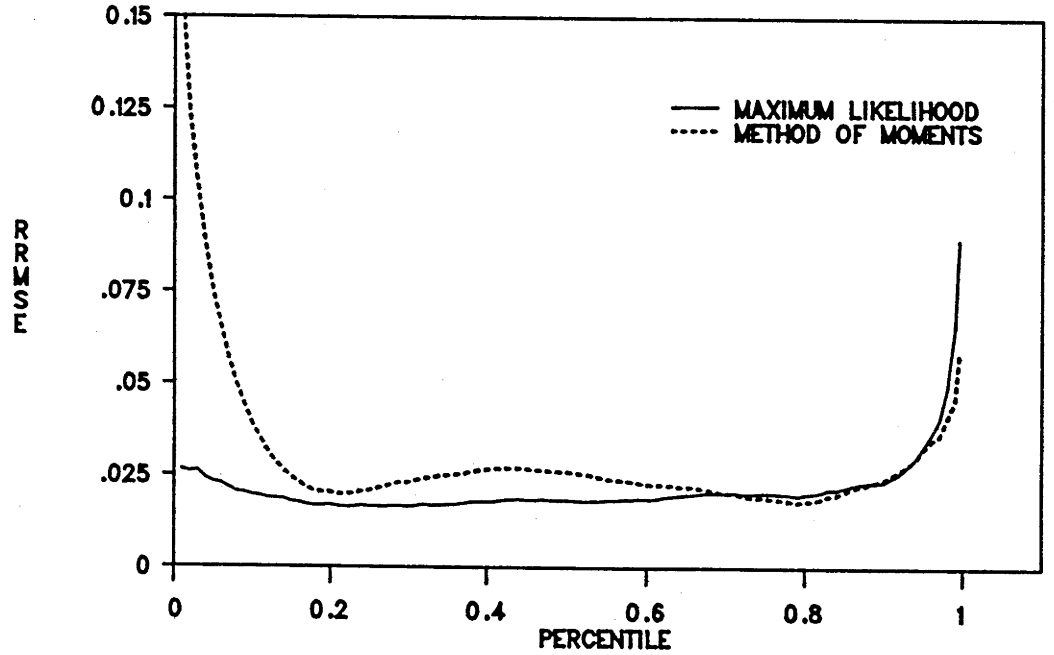

Figure 4.4: The RRMSE values with respect to each sample percentile for the 3-parameter gamma distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$ and $n = 365$

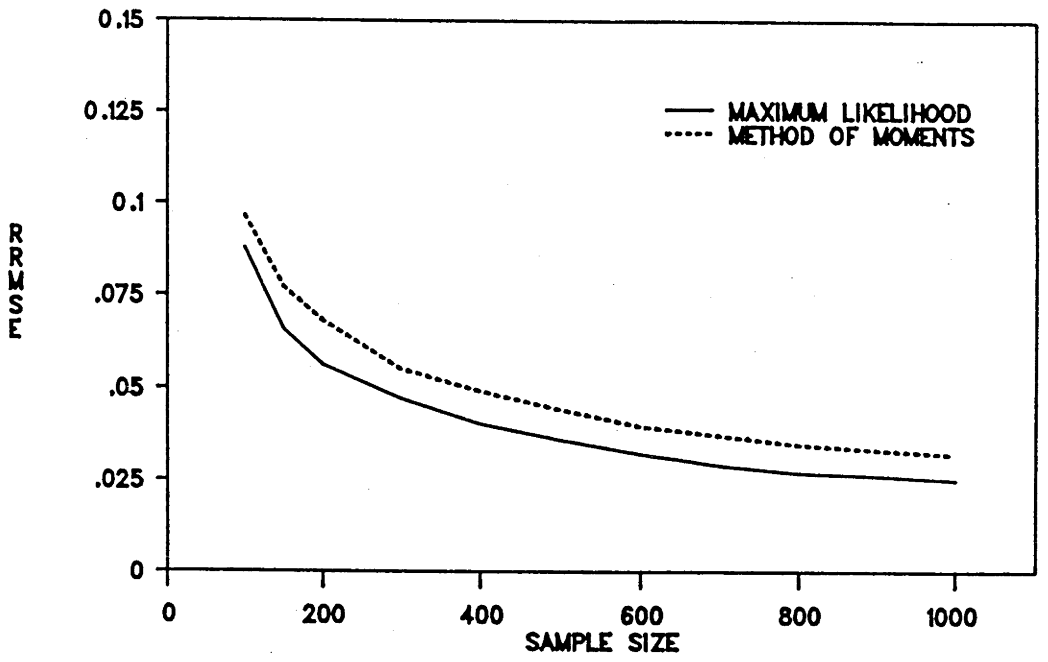

Figure 4.5: The RRMSE values with respect to true percentile (98%) for the 3-parameter gamma distribution, versus n

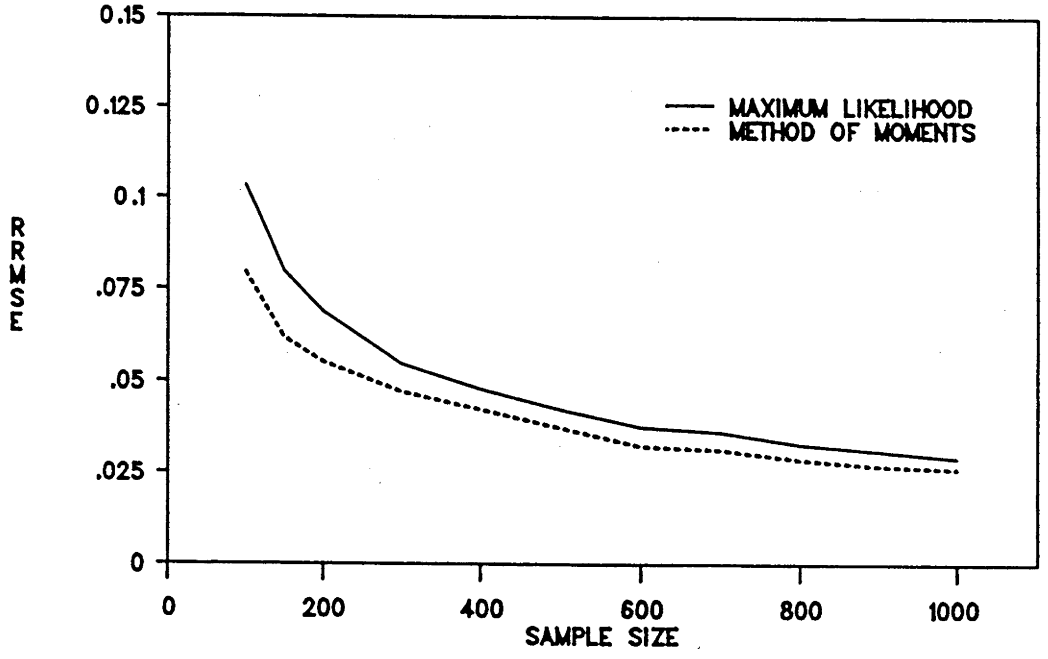

Figure 4.6: The RRMSE values with respect to sample percentile (98%) for the 3-parameter gamma distribution, versus n

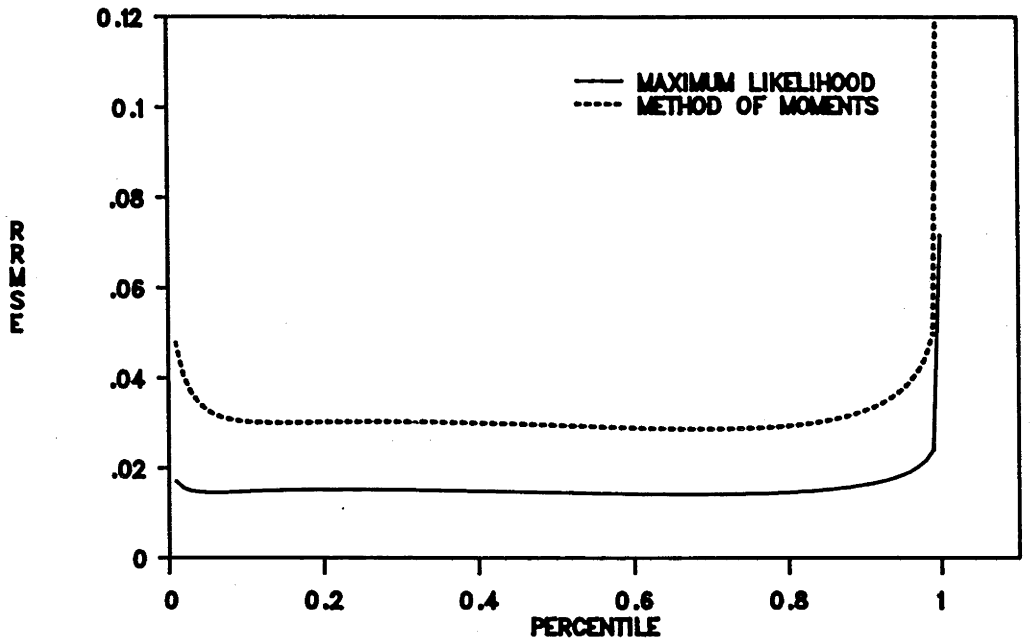

Figure 4.7: The RRMSE values with respect to each true percentile for the 3-parameter Weibull distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$ and $n = 365$

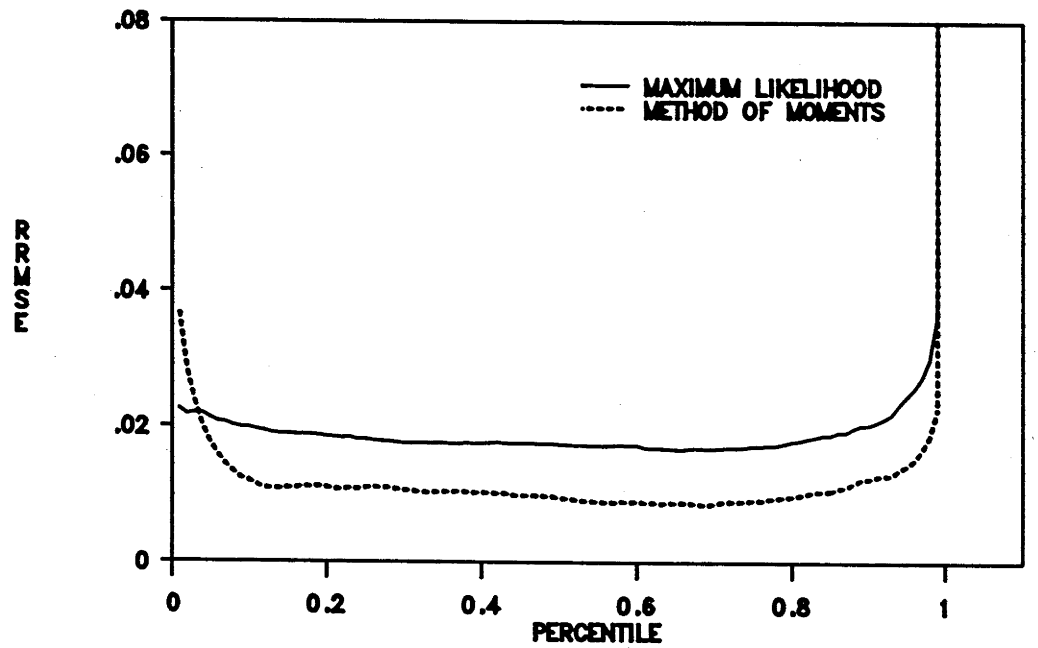

Figure 4.8: The RRMSE values with respect to each sample percentile for the 3-parameter Weibull distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$ and $n = 365$

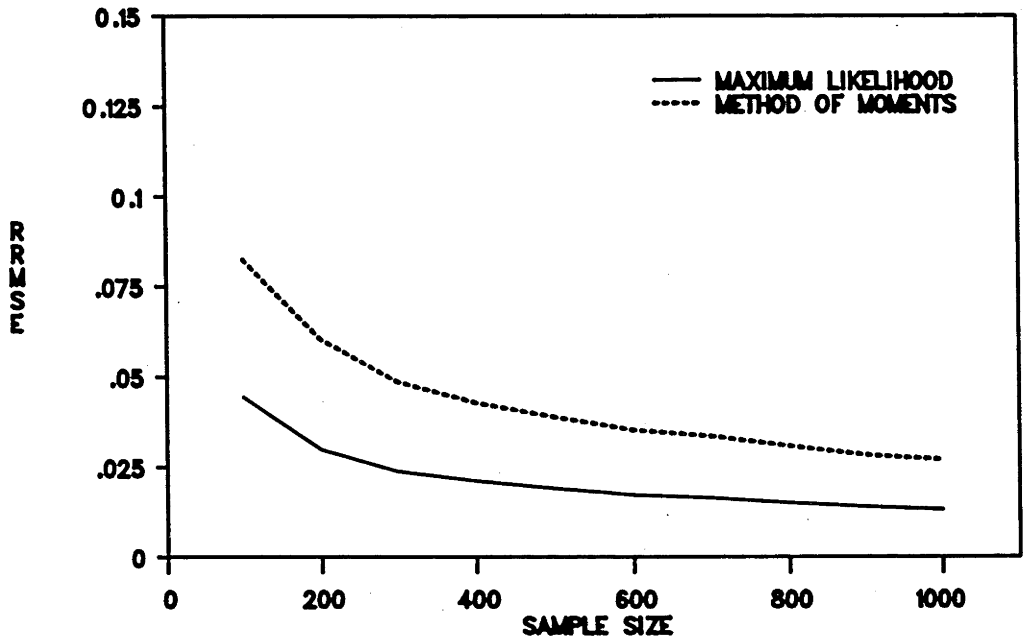

Figure 4.9: The RRMSE values with respect to true percentile (98%) for the 3-parameter Weibull distribution, versus n

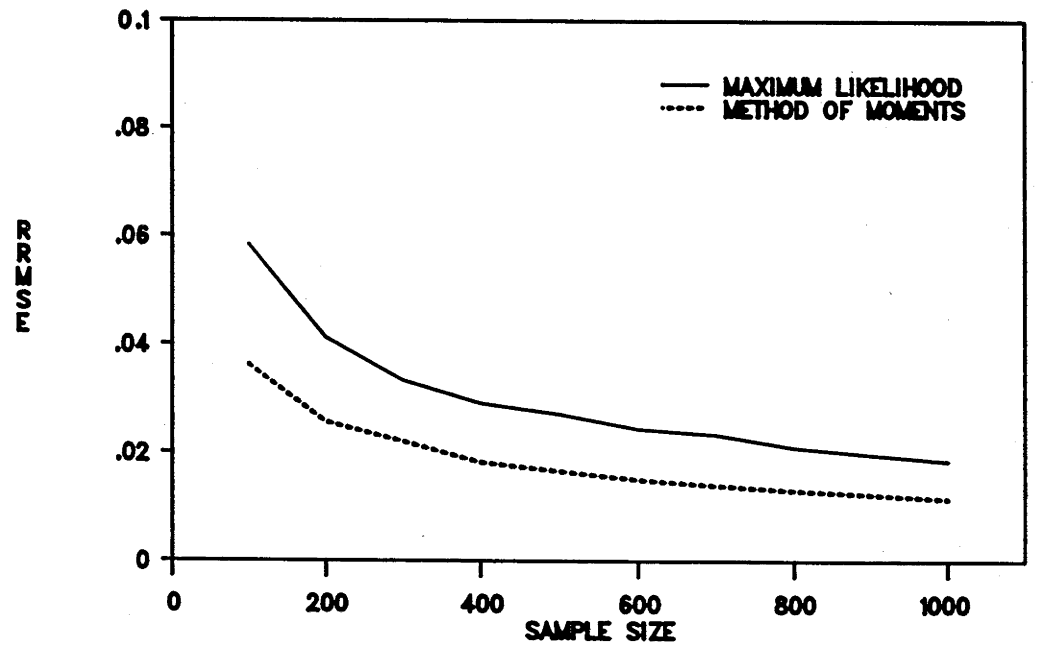

Figure 4.10: The RRMSE values with respect to sample percentile (98%) for the 3-parameter Weibull distribution, versus n

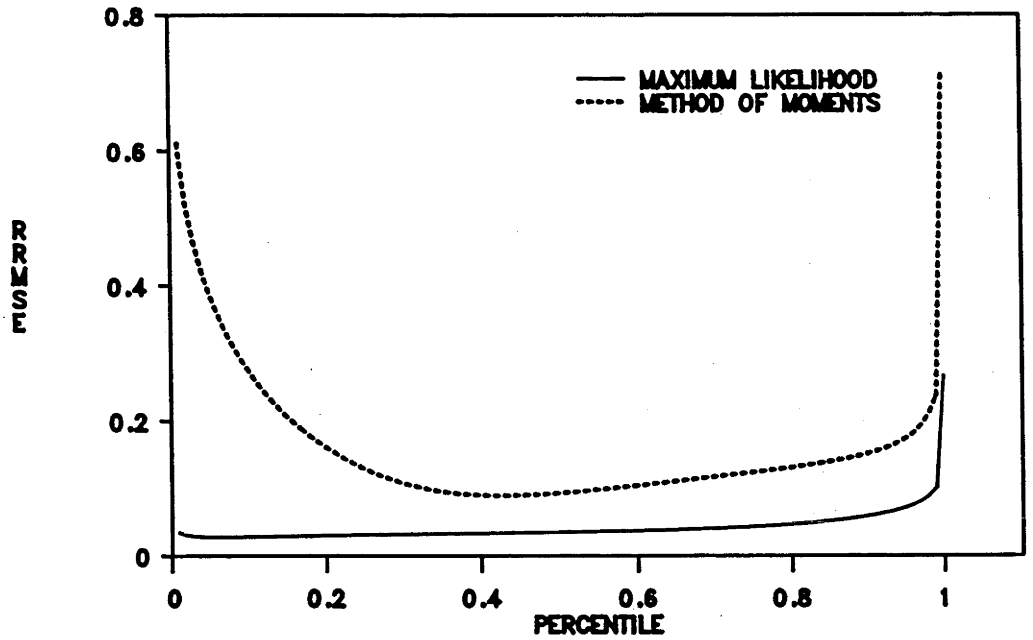

Figure 4.11: The RRMSE values with respect to each true percentile for the 3-parameter lognormal distribution with $(\alpha, \beta, \gamma) = (0.9, 1, 1)$ and $n = 365$

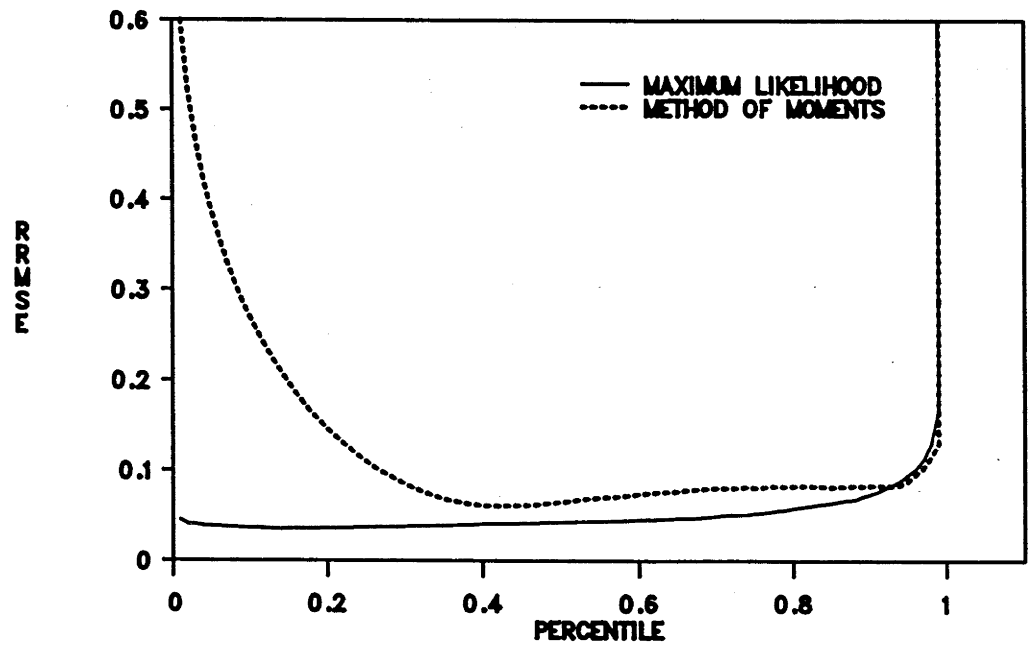

Figure 4.12: The RRMSE values with respect to each sample percentile for the 3-parameter lognormal distribution with $(\alpha, \beta, \gamma) = (0.9, 1, 1)$ and $n = 365$

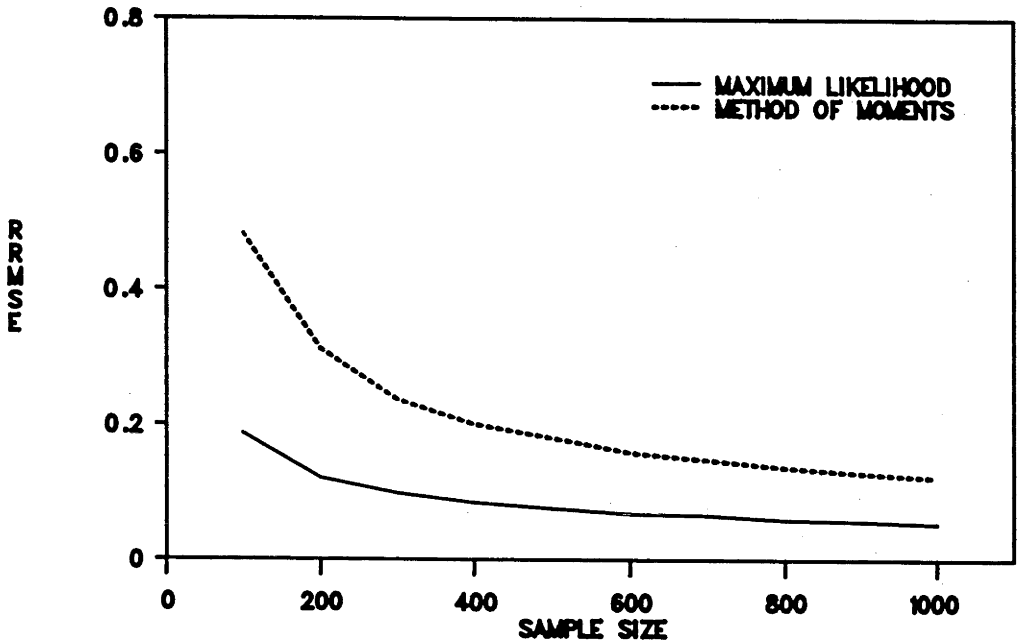

Figure 4.13: The RRMSE values with respect to true percentile (98%) for the 3-parameter lognormal distribution, versus n

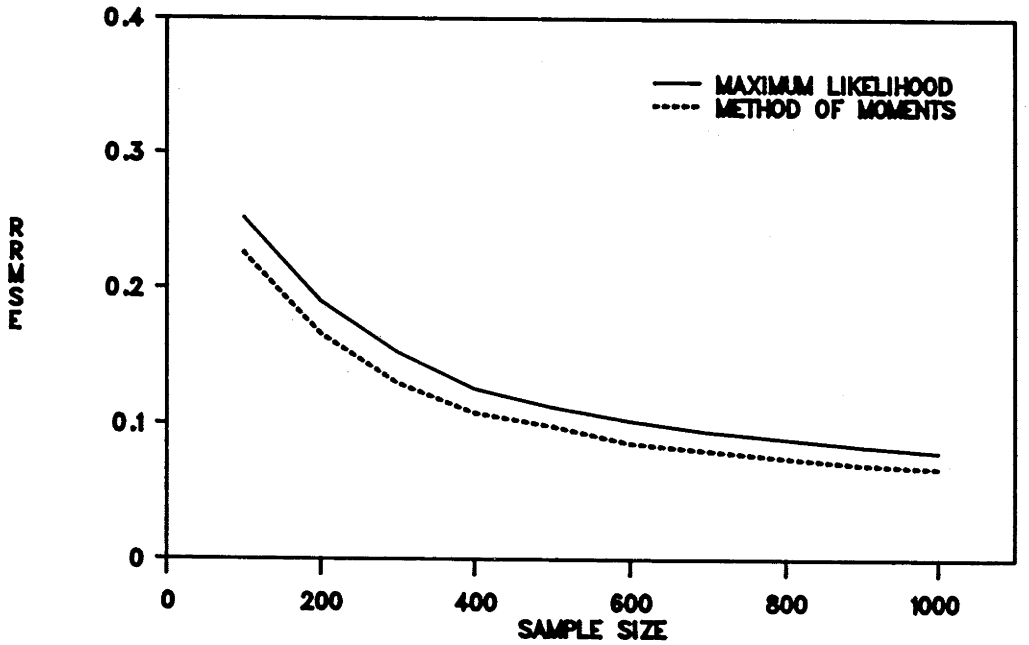

Figure 4.14: The RRMSE values with respect to sample percentile (98%) for the 3-parameter lognormal distribution, versus n

Table 4.1

Estimates of parameters derived from ML and MT approaches for the three-parameter gamma distribution based on 1000 Monte Carlo simulations for sample size n=365

	Estimated Parameter	Performance Criteria	ML α=2.0	MT β=1.0	ML α=6.0	MT β=1.0	ML α=2.0	MT β=3.0	ML α=6.0	MT β=3.0	ML α=2.0	MT β=5.0	ML α=6.0	MT β=5.0
γ=1.0	α	MEAN	1.95	2.26	6.03	7.24	1.95	2.25	6.03	7.30	1.95	2.25	6.03	7.30
		BIAS	-0.02	0.13	0.01	0.21	-0.03	0.12	0.01	0.22	-0.03	0.12	0.01	0.22
		RRMSE	0.10	0.46	0.27	0.76	0.11	0.46	0.27	0.78	0.11	0.46	0.27	0.78
	β	MEAN	1.02	1.01	1.03	1.02	3.07	3.03	3.08	3.05	5.12	5.05	5.13	5.09
		BIAS	0.02	0.01	0.03	0.02	0.02	0.01	0.03	0.02	0.02	0.01	0.03	0.02
		RRMSE	0.10	0.25	0.16	0.30	0.10	0.25	0.16	0.30	0.10	0.25	0.16	0.30
	γ	MEAN	1.02	0.91	1.05	0.67	1.06	0.75	1.14	-0.01	1.10	0.59	1.23	-0.69
		BIAS	0.02	-0.09	0.05	-0.33	0.06	-0.25	0.14	-1.01	0.10	-0.41	0.23	-1.69
		RRMSE	0.05	0.34	0.65	1.58	0.15	1.02	1.95	4.86	0.25	1.70	3.26	8.10
γ=3.0	α	MEAN	1.95	2.25	6.03	7.30	1.95	2.25	6.03	7.30	1.95	2.25	6.03	7.30
		BIAS	-0.03	0.12	0.01	0.22	-0.03	0.12	0.01	0.22	-0.03	0.12	0.01	0.22
		RRMSE	0.11	0.46	0.27	0.78	0.11	0.46	0.27	0.78	0.11	0.46	0.27	0.78
	β	MEAN	1.02	1.01	1.03	1.02	3.07	3.03	3.08	3.05	5.12	5.05	5.13	5.09
		BIAS	0.02	0.01	0.03	0.02	0.02	0.01	0.03	0.02	0.02	0.01	0.03	0.02
		RRMSE	0.10	0.25	0.16	0.30	0.10	0.25	0.16	0.30	0.10	0.25	0.16	0.30
	γ	MEAN	3.02	2.92	3.05	2.66	3.06	2.75	3.14	1.99	3.10	2.59	3.23	1.31
		BIAS	0.01	-0.03	0.02	-0.11	0.02	-0.08	0.05	-0.34	0.03	-0.14	0.08	-0.56
		RRMSE	0.02	0.11	0.22	0.54	0.05	0.34	0.65	1.62	0.08	0.57	1.09	2.70
γ=5.0	α	MEAN	1.95	2.25	6.03	7.30	1.95	2.25	6.03	7.30	1.95	2.25	6.03	7.30
		BIAS	-0.03	0.12	0.01	0.22	-0.03	0.12	0.01	0.22	-0.03	0.12	0.01	0.22
		RRMSE	0.11	0.46	0.27	0.78	0.11	0.46	0.27	0.78	0.11	0.46	0.27	0.78
	β	MEAN	1.02	1.01	1.03	1.02	3.07	3.03	3.08	3.05	5.12	5.05	5.13	5.09
		BIAS	0.02	0.01	0.03	0.02	0.02	0.01	0.03	0.02	0.02	0.01	0.03	0.02
		RRMSE	0.10	0.25	0.16	0.30	0.10	0.25	0.16	0.30	0.10	0.25	0.16	0.30
	γ	MEAN	5.02	4.92	5.05	4.66	5.06	4.75	5.14	3.99	5.10	4.59	5.23	3.31
		BIAS	0.00	-0.02	0.01	-0.07	0.01	-0.05	0.03	-0.20	0.02	-0.08	0.05	-0.34
		RRMSE	0.01	0.07	0.13	0.32	0.03	0.20	0.39	0.97	0.05	0.34	0.65	1.62

Table 4.2

Estimates of percentile deviation from parent values derived from ML and MT approaches for the three-parameter gamma distribution based on 1000 Monte Carlo simulations for sample size n=365

	Estimated Parameter	Performance Criteria	ML α=2.0	MT β=1.0	ML α=6.0	MT β=1.0	ML α=2.0	MT β=3.0	ML α=6.0	MT β=3.0	ML α=2.0	MT β=5.0	ML α=6.0	MT β=5.0
γ=1.0	MAX1	BIAS	0.007	0.003	0.004	0.003	0.009	0.004	0.004	0.003	0.009	0.004	0.004	0.003
		RRMSE	0.053	0.097	0.044	0.074	0.057	0.102	0.046	0.078	0.058	0.104	0.046	0.078
	MAX2	BIAS	0.006	0.002	0.003	0.002	0.007	0.003	0.003	0.003	0.007	0.003	0.004	0.003
		RRMSE	0.049	0.086	0.039	0.065	0.053	0.092	0.041	0.068	0.054	0.093	0.041	0.069
	98	BIAS	0.004	0.001	0.002	0.001	0.005	0.002	0.002	0.001	0.005	0.002	0.002	0.001
		RRMSE	0.042	0.067	0.031	0.051	0.047	0.073	0.033	0.054	0.048	0.075	0.034	0.054
γ=3.0	MAX1	BIAS	0.007	0.003	0.004	0.003	0.008	0.004	0.004	0.003	0.008	0.004	0.004	0.003
		RRMSE	0.044	0.079	0.039	0.067	0.053	0.095	0.044	0.074	0.055	0.099	0.045	0.076
	MAX2	BIAS	0.005	0.003	0.003	0.002	0.007	0.003	0.003	0.002	0.007	0.003	0.003	0.002
		RRMSE	0.040	0.069	0.035	0.058	0.049	0.085	0.039	0.066	0.052	0.089	0.040	0.067
	98	BIAS	0.003	0.001	0.002	0.001	0.004	0.002	0.002	0.001	0.005	0.002	0.002	0.001
		RRMSE	0.033	0.051	0.027	0.044	0.042	0.066	0.031	0.051	0.045	0.070	0.032	0.053
γ=5.0	MAX1	BIAS	0.006	0.003	0.003	0.003	0.008	0.004	0.004	0.003	0.008	0.004	0.004	0.003
		RRMSE	0.038	0.068	0.036	0.060	0.050	0.089	0.042	0.072	0.053	0.095	0.044	0.074
	MAX2	BIAS	0.005	0.002	0.003	0.002	0.006	0.003	0.003	0.002	0.007	0.003	0.003	0.002
		RRMSE	0.034	0.058	0.031	0.052	0.046	0.079	0.038	0.063	0.049	0.085	0.039	0.066
	98	BIAS	0.003	0.001	0.001	0.001	0.004	0.002	0.002	0.001	0.004	0.002	0.002	0.001
		RRMSE	0.027	0.042	0.024	0.039	0.039	0.060	0.030	0.049	0.042	0.066	0.031	0.051

Table 4.3

Estimates of percentile deviation from sample values derived from ML and MT approaches for the three-parameter gamma distribution based on 1000 Monte Carlo simulations for sample size n=365

	Estimated Parameter	Performance Criteria	ML α=2.0	MT β=1.0	ML α=6.0	MT β=1.0	ML α=2.0	MT β=3.0	ML α=6.0	MT β=3.0	ML α=2.0	MT β=5.0	ML α=6.0	MT β=5.0
γ=1.0	MAX1	BIAS	0.032	0.019	0.019	0.014	0.035	0.020	0.020	0.014	0.036	0.021	0.020	0.014
		RRMSE	0.140	0.086	0.104	0.073	0.151	0.094	0.108	0.075	0.153	0.095	0.109	0.076
	MAX2	BIAS	0.012	0.003	0.008	0.004	0.012	0.002	0.008	0.003	0.013	0.002	0.008	0.004
		RRMSE	0.104	0.064	0.076	0.050	0.111	0.070	0.078	0.051	0.113	0.071	0.079	0.052
	98	BIAS	0.013	0.008	0.010	0.007	0.017	0.011	0.010	0.008	0.017	0.011	0.010	0.008
		RRMSE	0.063	0.041	0.044	0.028	0.070	0.047	0.047	0.030	0.071	0.048	0.047	0.030
γ=3.0	MAX1	BIAS	0.023	0.014	0.016	0.011	0.031	0.018	0.019	0.013	0.033	0.020	0.020	0.014
		RRMSE	0.115	0.072	0.092	0.064	0.140	0.087	0.103	0.072	0.146	0.091	0.106	0.074
	MAX2	BIAS	0.007	0.000	0.006	0.003	0.010	0.001	0.007	0.003	0.011	0.002	0.007	0.003
		RRMSE	0.083	0.052	0.066	0.043	0.102	0.064	0.075	0.049	0.108	0.067	0.077	0.050
	98	BIAS	0.011	0.007	0.008	0.006	0.015	0.010	0.009	0.007	0.016	0.011	0.010	0.008
		RRMSE	0.048	0.032	0.038	0.025	0.063	0.042	0.044	0.028	0.067	0.045	0.046	0.029
γ=5.0	MAX1	BIAS	0.018	0.011	0.014	0.010	0.028	0.017	0.018	0.013	0.031	0.018	0.019	0.013
		RRMSE	0.098	0.062	0.083	0.058	0.130	0.082	0.099	0.069	0.140	0.087	0.103	0.072
	MAX2	BIAS	0.005	0.000	0.005	0.002	0.009	0.001	0.007	0.003	0.010	0.001	0.007	0.003
		RRMSE	0.070	0.044	0.059	0.039	0.095	0.060	0.072	0.047	0.102	0.064	0.075	0.049
	98	BIAS	0.008	0.006	0.007	0.005	0.013	0.009	0.009	0.007	0.015	0.010	0.009	0.007
		RRMSE	0.039	0.026	0.034	0.022	0.057	0.038	0.042	0.027	0.063	0.042	0.044	0.028

Table 4.4
Estimates of parameters derived from ML and MT approaches for the three-parameter Weibull distribution based on 1000 Monte Carlo simulations for sample size n=365

	Estimated Parameter	Performance Criteria	ML α=2.0	MT β=1.0	ML α=4.0	MT β=1.0	ML α=2.0	MT β=3.0	ML α=4.0	MT β=3.0	ML α=2.0	MT β=5.0	ML α=4.0	MT β=5.0
γ=1.0	α	MEAN	1.99	2.05	4.02	4.27	1.98	2.03	4.03	4.27	1.98	2.04	4.03	4.28
		BIAS	-0.01	0.02	0.01	0.07	-0.01	0.02	0.01	0.07	-0.01	0.02	0.01	0.07
		RRMSE	0.06	0.15	0.13	0.34	0.06	0.14	0.13	0.33	0.06	0.15	0.13	0.34
	β	MEAN	0.99	1.01	1.00	1.05	2.97	3.02	3.01	3.15	4.94	5.04	5.01	5.27
		BIAS	-0.01	0.01	0.00	0.05	-0.01	0.01	0.00	0.05	-0.01	0.01	0.00	0.05
		RRMSE	0.04	0.12	0.11	0.29	0.04	0.11	0.11	0.28	0.04	0.12	0.11	0.29
	γ	MEAN	1.01	0.99	1.00	0.95	1.03	0.98	0.99	0.85	1.05	0.96	0.99	0.73
		BIAS	0.01	-0.01	0.00	-0.05	0.03	-0.02	-0.01	-0.15	0.05	-0.04	-0.01	-0.27
		RRMSE	0.03	0.09	0.11	0.28	0.09	0.25	0.32	0.81	0.15	0.45	0.54	1.39
γ=3.0	α	MEAN	1.98	2.04	4.03	4.28	1.98	2.04	4.03	4.28	1.98	2.04	4.03	4.28
		BIAS	-0.01	0.02	0.01	0.07	-0.01	0.02	0.01	0.07	-0.01	0.02	0.01	0.07
		RRMSE	0.06	0.15	0.13	0.34	0.06	0.15	0.13	0.34	0.06	0.15	0.13	0.34
	β	MEAN	0.99	1.01	1.00	1.05	2.97	3.03	3.01	3.16	4.94	5.04	5.01	5.27
		BIAS	-0.01	0.01	0.00	0.05	-0.01	0.01	0.00	0.05	-0.01	0.01	0.00	0.05
		RRMSE	0.04	0.12	0.11	0.29	0.04	0.12	0.11	0.29	0.04	0.12	0.11	0.29
	γ	MEAN	3.01	2.99	3.00	2.95	3.03	2.98	2.99	2.84	3.05	2.96	2.99	2.73
		BIAS	0.00	0.00	0.00	-0.02	0.01	-0.01	0.00	-0.05	0.02	-0.01	0.00	-0.09
		RRMSE	0.01	0.03	0.04	0.09	0.03	0.09	0.11	0.28	0.05	0.15	0.18	0.46
γ=5.0	α	MEAN	1.98	2.03	4.03	4.28	1.98	2.04	4.03	4.28	1.98	2.04	4.03	4.27
		BIAS	-0.01	0.01	0.01	0.07	-0.01	0.02	0.01	0.07	-0.01	0.02	0.01	0.07
		RRMSE	0.06	0.12	0.13	0.34	0.06	0.15	0.13	0.34	0.06	0.15	0.13	0.33
	β	MEAN	0.99	1.01	1.00	1.05	2.97	3.03	3.01	3.16	4.94	5.04	5.01	5.25
		BIAS	-0.01	0.01	0.00	0.05	-0.01	0.01	0.00	0.05	-0.01	0.01	0.00	0.05
		RRMSE	0.04	0.10	0.11	0.29	0.04	0.12	0.11	0.29	0.04	0.12	0.11	0.28
	γ	MEAN	5.01	4.99	5.00	4.95	5.03	4.98	4.99	4.84	5.05	4.96	4.98	4.74
		BIAS	0.00	0.00	0.00	-0.01	0.01	0.00	0.00	-0.03	0.01	-0.01	0.00	-0.05
		RRMSE	0.01	0.02	0.02	0.06	0.02	0.05	0.06	0.17	0.03	0.09	0.11	0.27

Table 4.5

Estimates of percentile deviation from parent values derived from ML and MT approaches for the three-parameter Weibull distribution based on 1000 Monte Carlo simulations for sample size n=365

	Estimated Parameter	Performance Criteria	ML α=2.0	MT β=1.0	ML α=4.0	MT β=1.0	ML α=2.0	MT β=3.0	ML α=4.0	MT β=3.0	ML α=2.0	MT β=5.0	ML α=4.0	MT β=5.0
γ=1.0	MAX1	BIAS	0.002	0.001	0.000	0.001	0.003	0.002	0.000	0.001	0.004	0.003	0.000	0.001
		RRMSE	0.030	0.064	0.014	0.028	0.037	0.072	0.019	0.037	0.039	0.082	0.020	0.040
	MAX2	BIAS	0.001	0.001	0.000	0.000	0.003	0.001	0.000	0.000	0.003	0.002	0.000	0.000
		RRMSE	0.027	0.056	0.012	0.025	0.034	0.064	0.017	0.033	0.036	0.074	0.018	0.036
	98	BIAS	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.001	0.000	0.000
		RRMSE	0.022	0.045	0.010	0.020	0.028	0.052	0.014	0.027	0.030	0.061	0.015	0.029
γ=3.0	MAX1	BIAS	0.002	0.001	0.000	0.000	0.003	0.002	0.000	0.001	0.003	0.002	0.000	0.001
		RRMSE	0.019	0.041	0.008	0.016	0.030	0.064	0.014	0.028	0.034	0.072	0.017	0.033
	MAX2	BIAS	0.001	0.001	0.000	0.000	0.002	0.001	0.000	0.000	0.002	0.002	0.000	0.000
		RRMSE	0.017	0.035	0.007	0.014	0.027	0.056	0.012	0.025	0.031	0.064	0.015	0.029
	98	BIAS	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.001	0.000	0.000
		RRMSE	0.013	0.027	0.005	0.011	0.022	0.045	0.010	0.020	0.025	0.051	0.012	0.024
γ=5.0	MAX1	BIAS	0.001	0.000	0.000	0.000	0.002	0.002	0.000	0.000	0.003	0.002	0.000	0.000
		RRMSE	0.014	0.024	0.006	0.011	0.026	0.054	0.011	0.022	0.030	0.064	0.014	0.027
	MAX2	BIAS	0.001	0.000	0.000	0.000	0.002	0.001	0.000	0.000	0.002	0.001	0.000	0.000
		RRMSE	0.012	0.020	0.005	0.010	0.023	0.047	0.010	0.020	0.027	0.056	0.012	0.024
	98	BIAS	0.000	0.000	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.000	0.000
		RRMSE	0.009	0.015	0.004	0.007	0.018	0.036	0.008	0.015	0.022	0.045	0.010	0.019

Table 4.6

Estimates of percentile deviation from sample values derived from ML and MT approaches for the three-parameter Weibull distribution based on 1000 Monte Carlo simulations for sample size n=365

	Estimated Parameter	Performance Criteria	ML α=2.0	MT β=1.0	ML α=4.0	MT β=1.0	ML α=2.0	MT β=3.0	ML α=4.0	MT β=3.0	ML α=2.0	MT β=5.0	ML α=4.0	MT β=5.0
γ=1.0	MAX1	BIAS	0.013	0.010	0.005	0.005	0.018	0.013	0.007	0.007	0.019	0.014	0.008	0.008
		RRMSE	0.070	0.053	0.030	0.026	0.087	0.065	0.041	0.036	0.091	0.068	0.043	0.038
	MAX2	BIAS	0.005	0.003	0.002	0.002	0.008	0.004	0.003	0.003	0.008	0.004	0.003	0.003
		RRMSE	0.051	0.034	0.022	0.017	0.064	0.042	0.030	0.024	0.067	0.044	0.032	0.025
	98	BIAS	0.007	0.006	0.003	0.003	0.010	0.007	0.004	0.004	0.010	0.008	0.005	0.005
		RRMSE	0.030	0.019	0.013	0.009	0.039	0.024	0.018	0.012	0.042	0.026	0.020	0.014
γ=3.0	MAX1	BIAS	0.007	0.005	0.002	0.003	0.013	0.010	0.005	0.005	0.016	0.011	0.006	0.006
		RRMSE	0.045	0.034	0.017	0.015	0.071	0.053	0.030	0.026	0.080	0.060	0.036	0.031
	MAX2	BIAS	0.003	0.001	0.001	0.001	0.005	0.003	0.002	0.002	0.007	0.003	0.002	0.002
		RRMSE	0.032	0.021	0.012	0.010	0.051	0.034	0.022	0.017	0.058	0.038	0.026	0.021
	98	BIAS	0.004	0.003	0.002	0.002	0.007	0.006	0.003	0.003	0.008	0.006	0.004	0.004
		RRMSE	0.018	0.011	0.007	0.005	0.030	0.019	0.013	0.009	0.035	0.022	0.016	0.011
γ=5.0	MAX1	BIAS	0.006	0.004	0.002	0.002	0.010	0.008	0.004	0.004	0.013	0.010	0.005	0.005
		RRMSE	0.033	0.025	0.012	0.010	0.059	0.045	0.024	0.021	0.071	0.053	0.030	0.026
	MAX2	BIAS	0.002	0.001	0.001	0.001	0.004	0.002	0.001	0.002	0.005	0.003	0.002	0.002
		RRMSE	0.023	0.015	0.009	0.007	0.042	0.028	0.017	0.014	0.051	0.034	0.022	0.017
	98	BIAS	0.003	0.002	0.001	0.001	0.006	0.005	0.002	0.002	0.007	0.006	0.003	0.003
		RRMSE	0.013	0.008	0.005	0.003	0.024	0.015	0.010	0.007	0.030	0.019	0.013	0.009

Table 4.7

Estimates of parameters derived from ML and MT approaches for the three-parameter lognormal distribution based on 1000 Monte Carlo simulations for sample size n=365

	Estimated Parameter	Performance Criteria	ML α=0.5	MT β=1.0	ML α=0.9	MT β=1.0	ML α=0.5	MT β=3.0	ML α=0.9	MT β=3.0	ML α=0.5	MT β=5.0	ML α=0.9	MT β=5.0
γ=1.0	α	MEAN	0.50	0.47	0.91	0.76	0.50	0.47	0.91	0.76	0.50	0.47	0.91	0.76
		BIAS	0.01	-0.06	0.01	-0.15	0.01	-0.06	0.01	-0.15	0.01	-0.06	0.01	-0.15
		RRMSE	0.09	0.24	0.06	0.22	0.09	0.24	0.06	0.22	0.09	0.24	0.06	0.22
	β	MEAN	0.99	1.10	0.99	1.33	2.99	3.10	2.99	3.33	4.99	5.10	4.99	5.33
		BIAS	-0.01	0.10	-0.01	0.33	0.00	0.03	0.00	0.11	0.00	0.02	0.00	0.07
		RRMSE	0.09	0.28	0.06	0.41	0.03	0.09	0.02	0.14	0.02	0.06	0.01	0.08
	γ	MEAN	1.01	0.63	1.01	-0.08	1.04	-1.74	1.11	-6.95	1.29	-19.27	1.80	-57.72
		BIAS	0.01	-0.37	0.01	-1.08	0.04	-2.74	0.11	-7.95	0.29	-20.27	0.80	-58.72
		RRMSE	0.20	0.82	0.08	1.29	1.49	6.04	0.58	9.53	11.03	44.61	4.30	70.43
γ=3.0	β	MEAN	0.50	0.47	0.91	0.76	0.50	0.47	0.91	0.76	0.50	0.47	0.91	0.76
		BIAS	0.01	-0.06	0.01	-0.15	0.01	-0.06	0.01	-0.15	0.01	-0.06	0.01	-0.15
		RRMSE	0.09	0.24	0.06	0.22	0.09	0.24	0.06	0.22	0.09	0.24	0.06	0.22
	γ	MEAN	0.99	1.10	0.99	1.33	2.99	3.10	2.99	3.33	4.99	5.10	4.99	5.33
		BIAS	-0.01	0.10	-0.01	0.33	0.00	0.03	0.00	0.11	0.00	0.02	0.00	0.07
		RRMSE	0.09	0.28	0.06	0.41	0.03	0.09	0.02	0.14	0.02	0.06	0.01	0.08
γ=5.0	β	MEAN	3.01	2.63	3.01	1.92	3.04	0.26	3.11	-4.95	3.29	-17.27	3.80	-55.72
		BIAS	0.00	-0.12	0.00	-0.36	0.01	-0.91	0.04	-2.65	0.10	-6.76	0.27	-19.57
		RRMSE	0.07	0.27	0.03	0.43	0.50	2.01	0.19	3.18	3.68	14.87	1.43	23.48
γ=1.0	γ	MEAN	0.50	0.47	0.91	0.76	0.50	0.47	0.91	0.76	0.50	0.47	0.91	0.76
		BIAS	0.01	-0.06	0.01	-0.15	0.01	-0.06	0.01	-0.15	0.01	-0.06	0.01	-0.15
		RRMSE	0.09	0.24	0.06	0.22	0.09	0.24	0.06	0.22	0.09	0.24	0.06	0.22
	β	MEAN	0.99	1.10	0.99	1.33	2.99	3.10	2.99	3.33	4.99	5.10	4.99	5.33
		BIAS	-0.01	0.10	-0.01	0.33	0.00	0.03	0.00	0.11	0.00	0.02	0.00	0.07
		RRMSE	0.09	0.28	0.06	0.41	0.03	0.09	0.02	0.14	0.02	0.06	0.01	0.08
	γ	MEAN	5.01	4.63	5.01	3.92	5.04	2.26	5.11	-2.95	5.29	-15.27	5.80	-53.72
		BIAS	0.00	-0.07	0.00	-0.22	0.01	-0.55	0.02	-1.59	0.06	-4.05	0.16	-11.74
		RRMSE	0.04	0.16	0.02	0.26	0.30	1.21	0.12	1.91	2.21	8.92	0.86	14.09

Table 4.8

Estimates of percentile deviation from parent values derived from ML and MT approaches for the three-parameter lognormal distribution based on 1000 Monte Carlo simulations for sample size n=365

	Estimated Parameter	Performance Criteria	ML α=0.5	MT β=1.0	ML α=0.9	MT β=1.0	ML α=0.5	MT β=3.0	ML α=0.9	MT β=3.0	ML α=0.5	MT β=5.0	ML α=9.9	MT β=5.0
γ=1.0	MAX1	BIAS	0.005	-0.008	0.017	-0.066	0.006	-0.008	0.017	-0.067	0.006	-0.008	0.017	-0.068
		RRMSE	0.078	0.173	0.134	0.322	0.084	0.186	0.137	0.328	0.085	0.187	0.138	0.329
	MAX2	BIAS	0.003	-0.007	0.013	-0.040	0.004	-0.007	0.013	-0.041	0.004	-0.007	0.013	-0.041
		RRMSE	0.066	0.142	0.117	0.275	0.072	0.154	0.120	0.284	0.073	0.156	0.121	0.285
	98	BIAS	0.001	-0.003	0.007	0.003	0.001	-0.004	0.007	0.003	0.001	-0.004	0.007	0.003
		RRMSE	0.048	0.099	0.089	0.210	0.054	0.110	0.093	0.220	0.054	0.112	0.094	0.221
γ=3.0	MAX1	BIAS	0.005	-0.007	0.016	-0.063	0.005	-0.008	0.017	-0.067	0.006	-0.008	0.017	-0.067
		RRMSE	0.068	0.150	0.128	0.307	0.082	0.182	0.136	0.326	0.085	0.187	0.138	0.329
	MAX2	BIAS	0.003	-0.006	0.012	-0.037	0.004	-0.007	0.013	-0.041	0.004	-0.007	0.013	-0.041
		RRMSE	0.056	0.120	0.109	0.258	0.070	0.150	0.119	0.281	0.072	0.155	0.120	0.284
	98	BIAS	0.001	-0.003	0.006	0.003	0.001	-0.003	0.007	0.003	0.001	-0.004	0.007	0.003
		RRMSE	0.039	0.080	0.080	0.189	0.052	0.107	0.092	0.217	0.054	0.111	0.094	0.221
γ=5.0	MAX1	BIAS	0.004	-0.006	0.015	-0.060	0.005	-0.008	0.017	-0.066	0.006	-0.008	0.017	-0.067
		RRMSE	0.060	0.133	0.123	0.293	0.080	0.178	0.135	0.324	0.084	0.186	0.137	0.329
	MAX2	BIAS	0.003	-0.005	0.011	-0.035	0.004	-0.007	0.013	-0.040	0.004	-0.007	0.013	-0.041
		RRMSE	0.049	0.105	0.103	0.243	0.068	0.146	0.118	0.278	0.072	0.154	0.120	0.284
	98	BIAS	0.001	-0.002	0.005	0.003	0.001	-0.003	0.007	0.003	0.001	-0.004	0.007	0.003
		RRMSE	0.033	0.068	0.073	0.172	0.050	0.103	0.091	0.213	0.054	0.111	0.094	0.221

Table 4.9

Estimates of percentile deviation from sample values derived from ML and MT approaches for the three-parameter lognormal distribution based on 1000 Monte Carlo simulations for sample size n=365

	Estimated Parameter	Performance Criteria	ML α=0.5	MT β=1.0	ML α=0.9	MT β=1.0	ML α=0.5	MT β=3.0	ML α=0.9	MT β=3.0	ML α=9.5	MT β=5.0	ML α=0.9	MT β=5.0
γ=1.0	MAX1	BIAS	0.040	0.004	0.109	-0.059	0.046	0.006	0.114	-0.060	0.047	0.006	0.115	-0.060
		RRMSE	0.176	0.091	0.351	0.145	0.190	0.097	0.362	0.148	0.192	0.098	0.363	0.148
	MAX2	BIAS	0.010	-0.012	0.034	-0.061	0.012	-0.013	0.037	-0.063	0.013	-0.013	0.037	-0.063
		RRMSE	0.116	0.070	0.225	0.147	0.126	0.076	0.233	0.151	0.128	0.076	0.234	0.152
	98	BIAS	0.012	0.002	0.027	0.006	0.014	0.003	0.030	0.006	0.014	0.003	0.030	0.006
		RRMSE	0.066	0.047	0.129	0.115	0.073	0.052	0.136	0.121	0.074	0.053	0.137	0.121
γ=3.0	MAX1	BIAS	0.031	0.002	0.098	-0.058	0.044	0.005	0.113	-0.060	0.047	0.006	0.115	-0.060
		RRMSE	0.152	0.079	0.329	0.140	0.186	0.095	0.358	0.147	0.191	0.098	0.363	0.148
	MAX2	BIAS	0.006	-0.011	0.028	-0.059	0.011	-0.013	0.036	-0.062	0.012	-0.013	0.037	-0.063
		RRMSE	0.099	0.060	0.209	0.139	0.123	0.074	0.230	0.150	0.127	0.076	0.233	0.151
	98	BIAS	0.008	0.002	0.023	0.005	0.013	0.003	0.029	0.006	0.014	0.003	0.030	0.006
		RRMSE	0.053	0.038	0.115	0.104	0.071	0.050	0.134	0.119	0.074	0.053	0.137	0.121
γ=5.0	MAX1	BIAS	0.024	0.001	0.088	-0.057	0.042	0.005	0.111	-0.059	0.046	0.006	0.115	-0.060
		RRMSE	0.134	0.070	0.310	0.135	0.181	0.093	0.355	0.146	0.191	0.097	0.362	0.148
	MAX2	BIAS	0.004	-0.010	0.023	-0.056	0.011	-0.013	0.035	-0.062	0.012	-0.013	0.037	-0.063
		RRMSE	0.086	0.053	0.195	0.132	0.120	0.072	0.228	0.149	0.127	0.076	0.233	0.151
	98	BIAS	0.006	0.001	0.019	0.004	0.012	0.003	0.028	0.006	0.014	0.003	0.030	0.006
		RRMSE	0.044	0.032	0.104	0.094	0.068	0.049	0.132	0.117	0.074	0.052	0.137	0.121

Chapter 5

Empirical Models of Fitting Errors

5.1 Introduction

The objective of this chapter is to show that empirical models can be constructed to describe the dependence of the error in fitting parametric models of probability distributions on type of distribution, sample size, parent parameter values and percentile property of interest. The procedure and results are described for three-parameter gamma, lognormal and Weibull distributions. Monte Carlo simulations are used to infer the true errors used as dependent variables to calibrate or infer the parameters of the empirical model. The simulations reflect errors obtained under ideal conditions where the samples generated are independent and identically distributed according to a known parameterisation, i.e. the three-parameter gamma, Weibull or lognormal distributions. The errors obtained in this way are therefore the minimum ones that could be expected in practice where the samples are not likely to be as strongly independent and identically distributed, nor will the true distribution be known or even exist.

The work was undertaken so that practitioners who wish to apply identification and estimation methods to random samples from some probability density function will be able to appeal to a simple formula to obtain an appreciation of the minimum error associated with fitting percentile values of interest. A similar exercise was undertaken by Jakeman, Taylor and Simpson (1986) for two-parameter distributions.

The approach used to construct the empirical models can be placed within the

framework of Response Surface Methodology (RSM). Therefore, we begin with a discussion of the history and steps in this general framework. This is followed by a description of some of the different techniques used. We then go on to show its use for our application to fitting probability density functions for air quality management uses.

5.2 Literature Review

RSM is a set of techniques for predicting empirical model structure from experiments. It employs some well-known statistical methodologies in solving model estimation and identification problems. Experiments can be carefully designed in order to optimally explore the relationship between dependent and independent variables. Initial developments can be traced back to contributions from J. Wishart, C.P. Winsor, E.A. Mitscherlich, F. Yates and others in the early 1930's. However, the major success was achieved in applying methods to explore the relationships between the yield of a chemical process and a set of input variables presumed to influence the yield. In 1951, G.E.P. Box and K.B. Wilson with their colleagues formally established RSM. Many followers, such as R.H Myers (1971), provided new developments and a full description of RSM. Since then, RSM has been successfully used and applied in many diverse fields such as environmental, agricultural, biological, computer and social sciences. However, the techniques for the RSM vary in different fields, depending upon the motivation for the experiments and the specific application. It is the task of different experts and specialists to develop RSM in their own areas.

5.3 Empirical Model-building and the Similarity to Time Series Analysis

Using RSM to build an empirical model generally encompasses steps in the following procedure:

1. Designing carefully a series of experiments within the region of most interest and

obtaining measurements of the response-surface variables;

2. Proceeding with an estimation method such as regression analysis to determine a mathematical model appropriate for fitting the data; scaling and other transformations can be used if necessary;

3. Testing hypotheses of model parameter significance and goodness of fit;

4. Employing optimization techniques to discover the best predicted value of the response and find the best operating conditions.

From the procedure above, it is easy to see that apart from the first and last steps, the techniques used are quite similar to others such as those used in time series analysis. Time series analysis is used predominently to describe a stochastic process (e.g. Box and Jenkins, 1976) while RSM presents the approximation to an underlying mechanism in the presence of noise. Both of them use very similar statistical techniques in dealing with the data. In both cases, a set of data collected from experiments can be summarised by fitting some form of mathematical model through an estimation method. Consequently, some well-established statistical techniques, such as least squares procedures, maximum likelihood estimation and hypotheses testing, can be easily adopted. Also the results of analysis for the empirical model can be presented in similar forms to those for time series models with an indication of performance using a statistic, such as variance. If the empirical model is an optimal approximation of a true mechanistic model, the variance between predicted and observed values will tend to be minimized.

5.4 The Response Function

RSM takes its name from the study of the relationship between the dependent variable y and a number of predictor variables $\boldsymbol{x} = (x_1, \ldots, x_k)'$ where the variable y is called the response. By using graphical techniques, when there is only a single independent variable, the relationship between input and output of a model can be easily plotted, hence the terminology, a response curve. If there are two independent variables as input which will determine a surface region of interest, the three dimensional curve is

known as a response surface. Generally, input variables can be k dimensional and the region of interest is still known as a response surface being defined in $k+1$ dimensional space.

The response function can be expressed as

$$y = f(\boldsymbol{x}) + \varepsilon \tag{5.1}$$

where ε is noise from measurement sampling and model errors, and y is the actual observed response (which may be a vector). If the approximation of the response surface model is optimal, the expected value of y is equal to the hypothetical response η, that is $E(y) = \eta$, while the discrepancy of $y - \eta$ is the error ε often assumed to have a normal distribution. The response function $f(x)$ is usually assumed to be a continuous function.

Basically, there are two ways of approximating a hypothetical response function. One is to use a series approximation which most commonly employs a Taylor series expansion; another is to assume a basic model structure by physical considerations and fit the data in order to yield relevant values of parameters for the response function.

5.4.1 Series Approximation

As an illustrative example, consider the response function with a single factor. If $f(x)$ is a continuous function with K derivatives, the function can be expressed as a Taylor series expansion about an arbitrary point x_0, as

$$\eta = f(x_0) + f'(x_0)(x - x_0) + \frac{1}{2}f''(x_0)(x - x_0)^2 + \ldots \tag{5.2}$$

where $f'(x_0)$ and $f''(x_0)$ are the first and second derivatives respectively of $f(x)$ with respect to x evaluated at x_0. Here x_0 can be any arbitrary value but it is usually defined at the center of the region of interest. Most commonly, a polynominal approximation is used as a response function. Using the coefficients $\lambda_0, \lambda_1, \lambda_2, \ldots$ as parameters which depend on x and the derivatives of $f(x)$ at x_0, the expansion of (5.2) can be transformed

to a polynominal form of degree n as

$$\eta = \lambda_0 + \lambda_1 x + \lambda_2 x^2 \ldots + \lambda_n x^n \tag{5.3}$$

In general, the higher the degree of the approximating function, the more closely the Taylor series can approximate the true function. However, higher degrees of approximation substantially increase computational complexity. In practice, a polynomial of first or second degree is often chosen to adequately represent the true function by limiting its application to an appropriately random region of the factor space.

Normally the computational procedure is first to assume a form with a lower degree of polynomial model. Then, fit observations to estimate the parameters of the model to obtain an estimate of the experimental error variance. Hypothesis tests may be used subsequently to evaluate the performance of the model. When performance is not satisfactory, a higher degree of the polynomial model may be selected and the procedure can be repeated until satisfactory prediction is achieved.

5.4.2 Deterministic Approximation

Series approximation may provide a satisfactory prediction of the true model but the form of such a model can be totally different from the underlying hypothetical form. Such empirical models cannot be used to explain the underlying process mechanisms and to explore the relationship between the variables in order to improve our understanding of natural phenomena. Normally, deterministic approximations to the underlying model are used in this case. The approximation involves similar factors to the series approximation, but the model structure is based on certain physical considerations. A major step is the quantitative analysis used to establish the relationship among these factors. By employing some well-known estimation procedure, the parameters of each factor will be obtained. Similarly, the goodness of fit test is usually to test the model performance. The final form of the model can be obtained by carefully designing and proceeding with this model-building procedure.

5.5 Empirical Model-building in Predicting RRMSE for Air Quality Management

The aim here is to construct an empirical model to predict the minimum relative root mean square errors (RRMSE) when using probability density function to represent frequency distribution of sampled environmental phenomena such as air pollutant concentrations. We demonstrate with the theoretical RRMSE for upper percentiles of the gamma, Weibull and lognormal distribution, which are of strong interest in air quality monitoring and modelling studies.

Recall from Chapter 2 that a statistical distribution can be used as a summary of a set of data by incorporating the data into an appropriate member of a general class of distributions. When a particular data set can be summarized by a certain form of distribution with few parameters, it is useful to know how well the distribution fits the data. This can be achieved partly by using a goodness of fit test which may give some indication of variability. The RSM approach here yields this variability in estimates of the upper percentiles for any sample size of interest. The quantity analysed is the RRMSE in upper percentiles.

The RRMSE is normally obtained by fitting the probability distribution to the data. Such errors depend on the type of the distribution, the values of its parameters, the estimation method used and the sample size of the data being fitted. These are the major factors in determining the response surface.

Studying the features of individual plots between each factor and response variable, where other factors are fixed, helps understanding of the effect between each factor and response. For illustrative purposes, the procedure of empirical model-building for the three-parameter gamma distribution will be shown in the following. A similar procedure can also be used to develop analogous models for most distributions, such as the two- and three-parameter Weibull and lognormal distributions.

5.5.1 Hypothetical Response Function of RRMSE

The first step in deriving an empirical model is to examine the effect of each factor on the response surface. These factors include: sample size n, the shape parameter α, scale parameter β, location parameter γ, and percentile ρ. Hence, the response function can be expressed as

$$y = f(n, \alpha, \beta, \gamma, \rho) + \varepsilon \tag{5.4}$$

where y is the RRMSE for fitting the three-parameter gamma distribution. Based on the analysis of the effects of each factor, a more specific function than this notional representation can be derived. These effects will now be discussed.

The Effect of Sample Size n

According to asymptotic theory, as sample size increases, the deviation between the sampling and theoretical (or parent) distribution becomes smaller. That is, RRMSE is reduced as n increases as shown in Figure 5.1. Therefore, the relationship between y and n can be expressed as

$$y \propto \frac{1}{n^{a_1}} \tag{5.5}$$

where a_1 is a constant positive parameter of the empirical model that needs to be estimated.

The Effect of Shape, Scale and Location Factors

Similar analysis can be used for studying the effect of the shape, scale and location factors to the response variable. The shape parameter is very sensitive to RRMSE as shown in Figure 5.2, and the proportional relation is

$$y \propto \frac{1}{\alpha^{a_2}} \tag{5.6}$$

where a_2 is also a constant positive parameter of the empirical model to be determined.

The scale factor β has the opposite effect as shown in Figure 5.3. It can be shown that its response is of the form of

$$y \propto \beta^{a_3} \tag{5.7}$$

where a_3 is a constant positive parameter.

For the location factor γ, a similar pattern can be seen in that RRMSE decreases as γ increases, which is shown in Figure 5.4. Then the relationship between γ and y can be written as

$$y \propto \frac{1}{\gamma^{a_4}} \tag{5.8}$$

where a_4 is a constant positive parameter.

The Effect of Percentiles

The percentile ρ is quite different in affecting the response variable. From Figure 5.5, it can be seen that RRMSE is reasonably consistent over the range of most percentiles except the lower and upper extremes. This is because the method of maximum log-likelihood used yields poor performance on the two extreme cases as discussed in the previous chapter. Here we are only concerned with the upper percentiles which have an exponential relationship with the response variable as

$$y \propto \rho^{a_5} \tag{5.9}$$

where a_5 is a constant positive parameter for the percentile factor.

Based on the analysis above and ignoring the error term, the essential structure of the empirical model is proposed as

$$y = \frac{C\beta^{a_3}\rho^{a_5}}{n^{a_1}\alpha^{a_2}\gamma^{a_4}} \tag{5.10}$$

where C is a constant which contains composite effects from all factors.

5.5.2 Transformation, Estimation and Identification

In the next stage, the major task is to determine the appropriate parameters for each factor and ensure that these parameters lead to an empirical model with adequate performance. To achieve this goal, there are three procedures to be typically followed. These are: transformation, estimation and identification.

Transformation

In order to simplify the estimation procedure, linear regression is used instead of non-linear regression. Nonlinear transformations are therefore necessary. A natural transformation is a logarithmic one of the form

$$\begin{aligned} logy &= logC - a_1 logn - a_2 log\alpha + a_3 log\beta - a_4 log\gamma + a_5 log\rho \\ &= a_0 + a_1 u_1 + a_2 u_2 + a_3 u_3 + a_4 u_4 + a_5 u_5 \end{aligned} \tag{5.11}$$

where $a_0 = logC$ and $u_1 = logn, u_2 = log\alpha$, and so on. These parameters can be estimated by standard linear least-squares methods.

Estimation

A standard ordinary least-squares method (regression analysis) can be easily applied to obtain parameters when the function is linear of the form (5.11). Adding experimental error, and considering n observations of the response surface, the standard form is

$$g_j = a_0 + a_1 u_{j1} + a_2 u_{j2} + a_3 u_{j3} + a_4 u_{j4} + a_5 u_{j5} + \varepsilon_j \tag{5.12}$$

$$j = 1, 2, \ldots n$$

where g_j denotes the observed response for the j'th trial, u_{ji} represents the level of factor i at the jth trial, and ε_j are the random errors in g_j. In general, the error assumptions are

1. Random errors ε_j are normally distributed with zero mean and variance σ^2;
2. They are mutually independent in the statistical sense.

The method of ordinary least-squares selects values of the parameters which minimise the quantity

$$R(a_0, a_2, \ldots, a_5) = \sum_{j=1}^{n} (g_j - a_0 - a_1 u_{j1} - \ldots - a_5 u_{j5})^2 \tag{5.13}$$

In matrix notation this can be written as

$$\mathbf{y} = \theta \mathbf{u} + \varepsilon \tag{5.14}$$

where

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

$$\mathbf{u} = \begin{bmatrix} 1 & u_{11} & u_{12} & \ldots & u_{15} \\ 1 & u_{21} & u_{22} & \ldots & u_{25} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & u_{n1} & u_{n2} & \ldots & u_{n5} \end{bmatrix}$$

$$\mathbf{a} = \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_5 \end{bmatrix}$$

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

The equation for the minimiser of (5.13) is then

$$\mathbf{u'ub} = \mathbf{u'y} \tag{5.15}$$

where $\boldsymbol{b} = (b_0, b_1, \ldots, b_k)$ is the parameter estimates. The solution for the least-squares estimates is

$$\mathbf{b} = (\mathbf{u'u})^{-1}\mathbf{u'y} \tag{5.16}$$

where $(\boldsymbol{u'u})^{-1}$ is the inverse of $\boldsymbol{u'u}$ which is symmetric.

Identification

Several statistics are commonly used to examine the performance of a fitted regression model. The F-statistic is normally used for testing the significance of the fitted regression equation. Under the assumption of normality for the errors, the test of the null hypothesis is H_0: all values of b_i (excluding b_0) are zero, against the alternative hypothesis H_a: at least one value of b_i (excluding b_i) is not zero. The F-statistic can be written as (Khuri and Cornell, 1987)

$$F = \frac{MeanSquareRegression}{MeanSquareResidual} = \frac{SSR/(k-1)}{SSE/(n-k)} \tag{5.17}$$

where SSR and SSE are defined as

$$SSR = \sum_{j=1}^{n}(\hat{y}_j - \bar{y})^2 \tag{5.18}$$

$$SSE = \sum_{j=1}^{n}(y_j - \hat{y}_j)^2 \tag{5.19}$$

respectively. The average value equals $\bar{y} = (y_1 + y_2 + \ldots + y_n)/n$ while $\hat{y}_j$ denotes the empirical estimate of y_j. If the null hypothesis is true, the F-statistic will follows an F-distribution with $k-1$ in the numerator and and $n-k$ degrees of freedom in the denominator. However, if the value of F exceeds the upper critical value $F_{a,k-1,n-k}$, then the null hypothesis is rejected at the a-level of significance.

For testing an hypothesis concerning the individual parameters in the proposed model, the T statistic is used which compares the parameter estimates in the fitted model to their respective estimated standard errors. A test of the null hypothesis H_0: $a_i = 0$ is performed against the alternative H_1: $a_i \neq 0$. The form of the T test statistic is (Khuri and Cornell, 1987)

$$T = \frac{a_i}{SD_i} \tag{5.20}$$

where SD_i is the estimated standard error corresponding to a_i; and the value of T is compared with the critical value from the t distribution.

In addition, the coefficient of determination is also very useful as an accompanying statistic to the F-statistic, which is the form of

$$R^2 = \frac{SSR}{SST} \tag{5.21}$$

where SSR is defined as

$$SST = \sum_{j=1}^{n} (y_j - \bar{y})^2. \tag{5.22}$$

The value R_2 is a measure of the proportion of total variation of the values of y_j about the mean $\bar{y}$ which is explained by the fitted model.

When the fitted model is rejected, a new search must be initiated for a more adequate model. It can be started from the analysis of residuals, and remedial measures, such as other transformations of y. In general, the true model is difficult to find, and attention should be focused on searching for a model with reasonable accuracy, which is useful for the intended application.

5.5.3 Simulation and Experimental Design

In general, the smaller the specific region of interest over which the approximation needs to be made, the better is the approximation achieved. This region is normally

known as the factor space in which the experiments can actually be performed. For some cases, the experiments require exploration of the whole region in order to derive the generalised response surface, but this is sometimes quite difficult in practice, perhaps involving large computational time and cost. Instead, the investigation can be restricted to certain limited regions of interest, which are particularly important in application.

From prior knowledge it is known that the density functions of air pollution data are positively skewed. Hence, the possible range of parameter values for a given distribution can be delineated. Since the shape parameter is the most sensitive parameter, it is chosen over the widest range of possible values: the shape parameter is assumed to take values within the range [0.5, 6] for the gamma distribution; [0.5, 4] for the Weibull distribution; and [0.4, 1.2] for the lognormal distribution. In the simulation experiments, the scale parameter and location parameter take on values from the range [1, 5] for these three distributions, respectively. Note that the lognormal distribution has opposite behaviour to the gamma and Weibull distributions as the shape parameter increases. For each entry in the tables, N=1000 replications of the experiments are processed. The sample size considered covers the ranges, from 50 to 1000.

Random sample generators are used for the Monte Carlo experiments. As in previous chapters these are DRNGAM, DRNWIB and DRNLNL for the gamma, Weibull and lognormal distributions, respectively. These subroutines are available in the International Mathematical and Statistical Library (IMSL) in version 1.0 (April 1987). The same seed number (1234) is used to obtain the first random sample of the first of the 1000 replications. Varying the initial seed produces similar results to those reported in the tables. The new maximum likelihood approach discussed in the Chapter 3 is adopted here for estimating the parameters of the distributions. For the regression analysis, the 'Shazam' package is used to obtain the parameters of the linear equation (5.14) and to calculate goodness-of-fit of the empirical model to the data.

5.5.4 The Experimental Result

The Gamma Distribution

With careful design of experiments in the factor space and through extensive simulation experiments, a large data set was obtained. For the three-parameter gamma distribution, 684 data points were used for the regression analysis of equation (5.12). The resulting model for the gamma distribution is

$$y = \frac{1.482\beta^{0.118}\rho^{14.871}}{n^{0.544}\alpha^{0.211}\gamma^{0.100}} \tag{5.23}$$

The details of fitting the data and the results of using significant tests are listed in Table 5.1 and 5.2.

In Table 5.1 and 5.2, all of the statistics used show that the empirical model fits the data well. By using the F statistic, the null hypothesis is rejected at the 0 level of significance, indicating strongly that at least one of the five parameters a_i in the equation (5.11) (excluding a_0) is not zero. The results of T tests show that each of these six parameters is significant. The R^2 value between observed and predicted quantities is 93.87 per cent which implies that 93.87 per cent of the total variation in the values of data is explained by the fitted model. The standard error of the estimated $\hat{\sigma}$ is .075. The sum of residuals is -0.756^{-12} with the variance of residuals 0.006, which indicate that this model is well-fitted to the data, and the mean of the residuals is close to zero and the variance of residuals is constant.

Figures 5.6 to 5.8 illustrate the fit of the empirical model of RRMSE to the data. The values of the shape, scale and location factors are $(\alpha, \beta, \gamma) = (2, 1, 1)$. Each plot uses a range of sample sizes between 100 to 1000. Figure 5.6 shows the fit of the model to the data at the 98 percentile, Figure 5.7 at the 99 percentile and Figure 5.8 at the 99.8 percentile which is equivalent to the maximum percentile obtained from Larsen's (Larsen, 1971) calculation when the sample size is 365. These figures clearly show that the empirical model provides a good fit to the data with reasonable accuracy.

The model (5.23) is a generalised form for the upper percentile estimation of RRMSE. To emphasize estimation of the maximum percentile, for example, the percentile factor in the model (5.12) can be omitted and then the regression analysis is performed by using the remaining factors against the estimated RRMSE related to the maximum percentile. Thus the data points are reduced from 684 to 114. The resulting model is

$$y = \frac{1.481\beta^{0.105}}{n^{0.555}\alpha^{0.176}\gamma^{0.091}}. \tag{5.24}$$

Comparing model (5.24) to (5.23), the parameters of this simplified model, a_0, a_1, a_2, a_3 and a_4, are slightly changed. The performances of (5.23) and (5.24) are almost identical which can be seen from the fit of both models to the data for the maximum percentile in Figure 5.8 and Figure 5.9. Hence, the simpler model of (5.24) can be used in predicting RRMSE at maximum percentile which is particularly important in air quality management.

The Weibull and Lognormal Distributions

Following a similar procedure to the model construction for the three-parameter gamma distribution, empirical models of the three-parameter Weibull and lognormal distributions for predicting RRMSE can be obtained. For the regression analysis, 510 data points were used for the three-parameter Weibull and 594 data points for the lognormal. The resulting empirical models for these two distributions are

Weibull:

$$y = \frac{0.977\beta^{0.387}\rho^{19.317}}{n^{0.470}\alpha^{1.146}\gamma^{0.363}} \tag{5.25}$$

Lognormal:

$$y = \frac{2.687\alpha^{0.974}\beta^{0.112}\rho^{23.757}}{n^{0.499}\gamma^{0.054}} \tag{5.26}$$

Note that the shape parameter α has a positive exponent for the three-parameter lognormal distribution, which is different to the other two distributions.

The details of fitting the data and the results of using significant tests are listed in Table 5.3 and 5.4. It can be seen from these tables that the empirical models fit the data well. The F statistic shows that the null hypothesis is rejected at the 0 level of significance for both models, which strongly indicates at least one value of the five parameters a_i (excluding a_0) is not zero. The T tests indicate that each of these five parameters (excluding a_0) is significant for both the Weibull and lognormal distribution. The R^2 value between observed and predicted quantities is 95.29 per cent for the three-parameter Weibull and 95.59 per cent for three-parameter lognormal distribution. The standard error of the estimated $\hat{\sigma}$ is 0.127 for three-parameter Weibull and .077 for the three-parameter lognormal distribution. The sum of residuals is -0.112^{-11} with the variance of residuals 0.016 for the three-parameter Weibull, and is -0.170^{-11} with the variance of residuals 0.006 for the three-parameter lognormal distribution. The properties of the residuals indicate that these models fit the data well, and the means of the residuals are close zero and the variances of the residuals appear constant.

Figures 5.10 to 5.15 illustrate the fit of the empirical model to the data for these two distributions. The values of the shape, scale and location factors are $(\alpha, \beta, \gamma) = (2, 1, 1)$ for the Weibull, and $(\alpha, \beta, \gamma) = (0.9, 1, 1)$ for the lognormal distribution. Each plot uses a range of sample sizes between 100 to 1000. Figures 5.10 and 5.13 shows the fit of model to the data at the 98 percentile for 3-parameter Weibull and lognormal distributions, respectively. Figures 5.8 and 5.11 are the fit at the 99 percentile and Figure 5.9 and 5.12 at the 99.8 percentile which is equivalent to the maximum percentile by using Larsen's calculation when the sample size is 365. The figures for the three-parameter Weibull clearly show that the empirical model provides a good fit to the data with reasonable accuracy. Although the fit of the model for the three-parameter lognormal is not as good as the two other distributions, the accuracy of the fit still appears to be useful.

When estimation is restricted to the maximum percentile, that is the percentile factor in the estimated models is omitted and the regression analysis performed on data associated with the maximum percentile only, these models become

$$y = \frac{0.932\beta^{0.368}}{n^{0.471}\alpha^{1.124}\gamma^{0.344}} \tag{5.27}$$

$$y = \frac{1.243\alpha^{0.911}\beta^{0.086}}{n^{0.389}\gamma^{0.044}} \tag{5.28}$$

The comparison of model performances between (5.25) and (5.27), and (5.26) and (5.28), shows that the models (5.25) and (5.27) perform almost identically but (5.28) has a slightly improved fit over (5.26). These can be seen from the fit of the model to the data for the maximum percentile as indicated in Figures 5.12 and 5.16, and Figures 5.15 and 5.17.

5.6 Concluding Remarks

In this chapter, empirical models of the three-parameter gamma, Weibull and lognormal distributions for predicting RRMSE are developed by employing RSM. Such models are important in evaluating the goodness of fit of some distributional forms to air pollution data and, once calibrated, require only simple calculations. To obtain a reasonable causal linkage between the factors and the response surface, graphical techniques were used to show the major relation between each factor and response variable. Based on prior knowledge for air quality assessment, the factors space is carefully designed so that the simulation experiments proceed efficiently. By using the extensive data from the simulation experiments, regression analysis derives the expected model structure by the ordinary least squares method of estimation. A variety of model selection criteria are used to examine the model performance. From the results, it is seen that the models developed here have reasonable accuracy in predicting the RRMSE. Obviously, using such models can improve the efficiency of assessment procedures in air quality management. Also, such empirical models are important for the study of probability distributions in extreme theory applications.

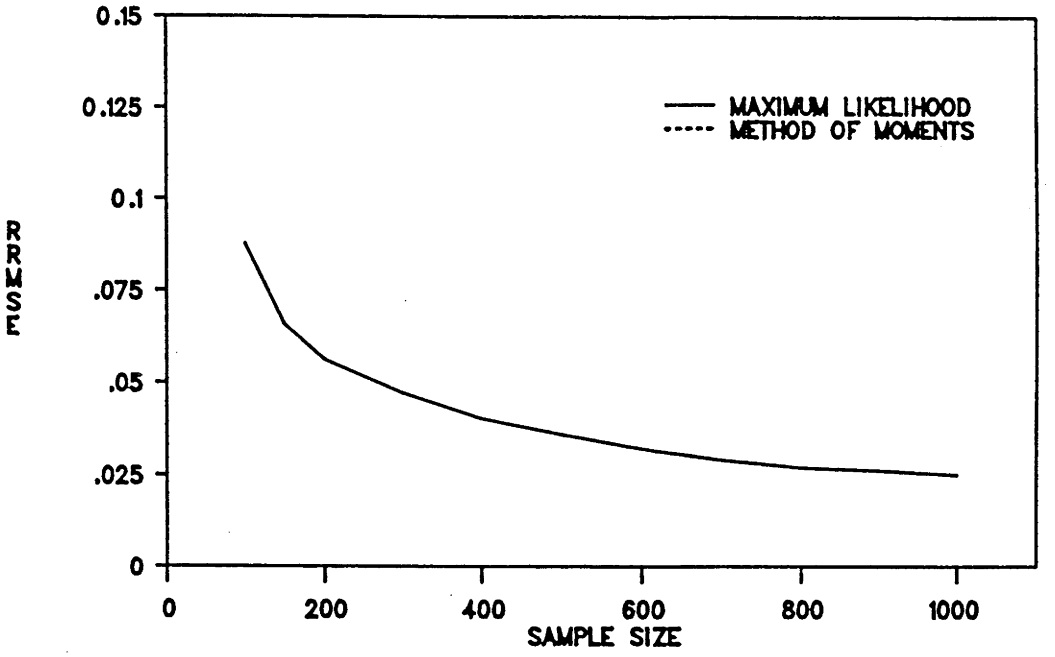

Figure 5.1: The RRMSE values at 98 percentile versus sample size for the 3-parameter gamma distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$

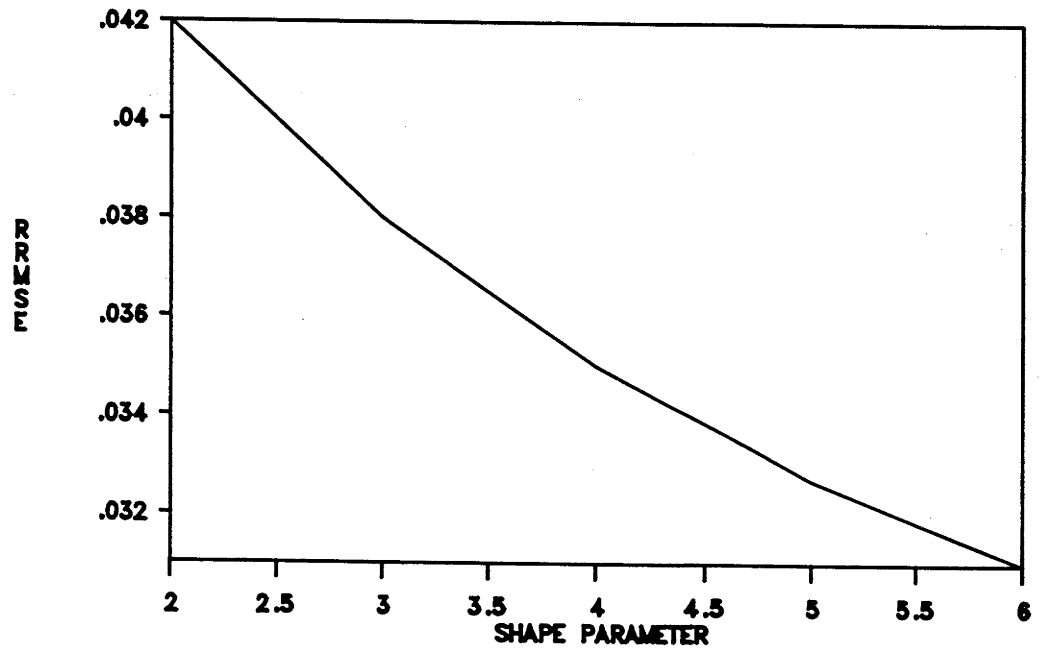

Figure 5.2: The RRMSE values at 98 percentile versus the shape parameter for the 3-parameter gamma distribution with $(\beta, \gamma) = (1, 1)$ and $n = 365$

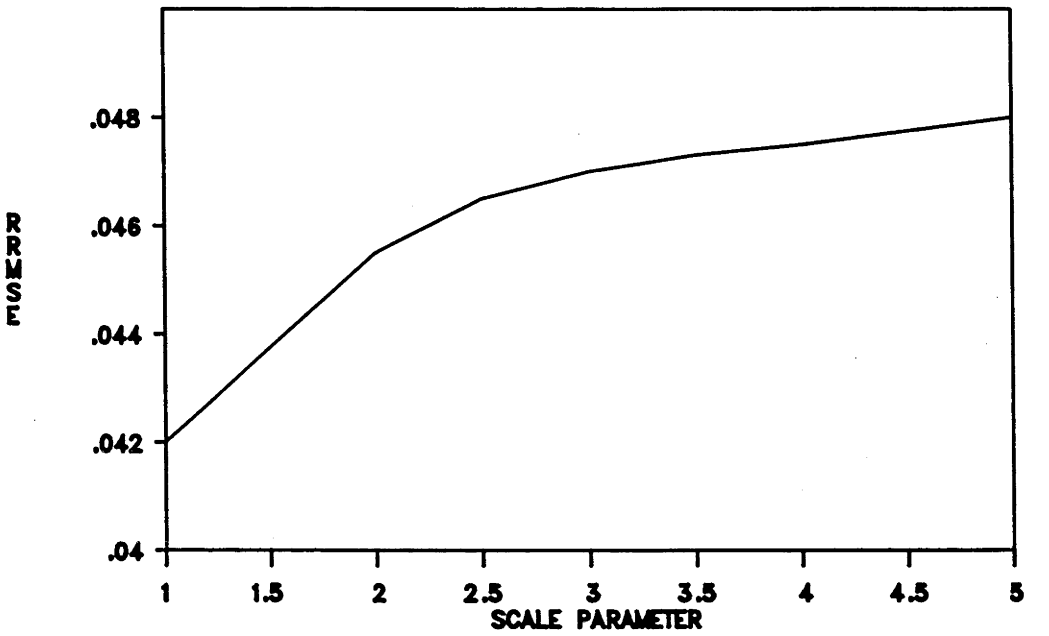

Figure 5.3: The RRMSE values at 98 percentile versus the scale parameter for the 3-parameter gamma distribution with $(\alpha, \gamma) = (2, 1)$ and $n = 365$

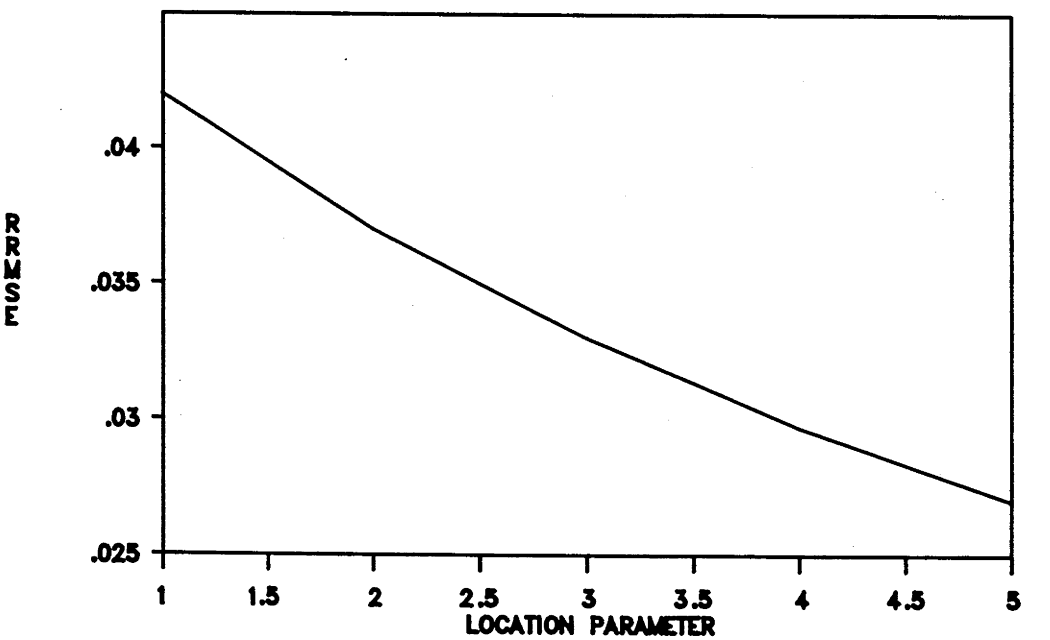

Figure 5.4: The RRMSE values at 98 percentile versus the location parameter for the 3-parameter gamma distribution with $(\alpha, \beta) = (2, 1)$ and $n = 365$

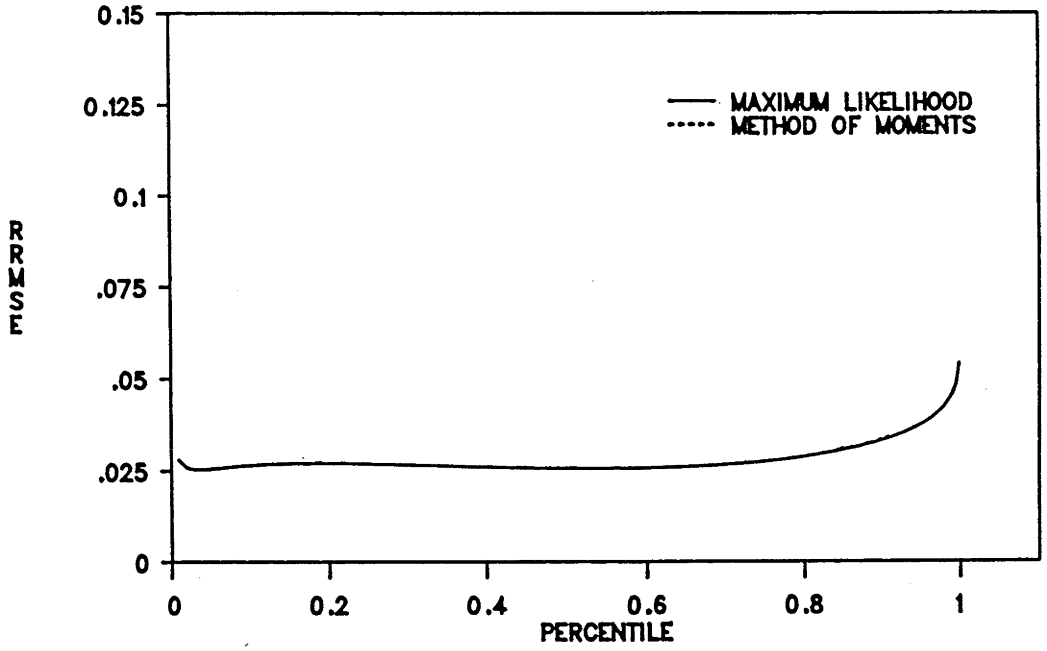

Figure 5.5: The RRMSE values versus each value of percentiles for the 3-parameter gamma distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$ and $n = 365$

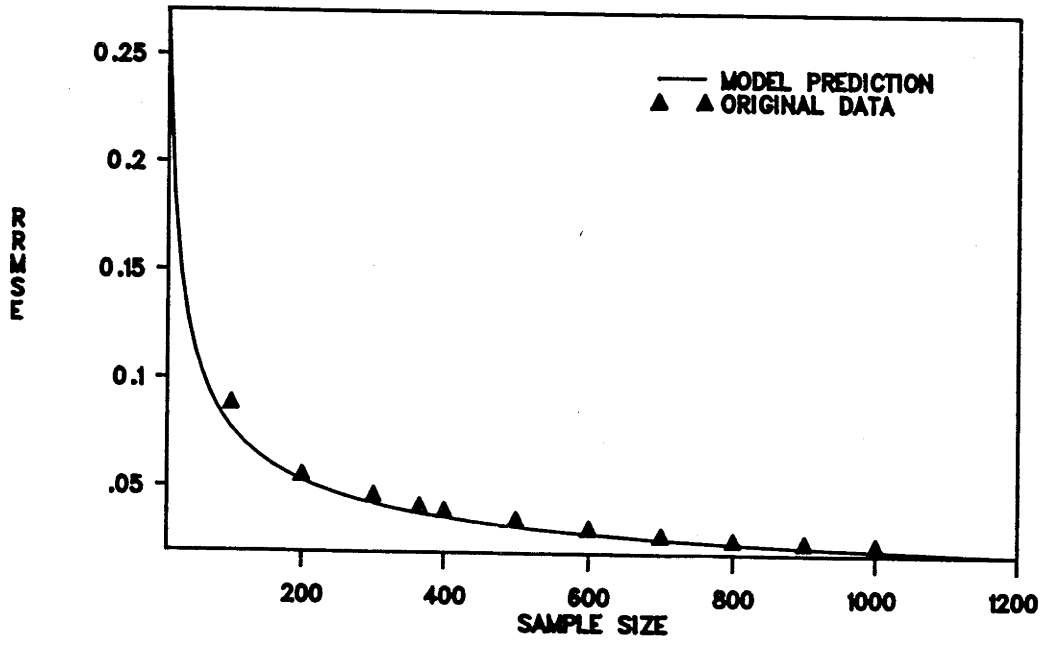

Figure 5.6: Fit of error model (5.23) to the RRMSE values at 98 percentile versus sample size for the 3-parameter gamma distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$

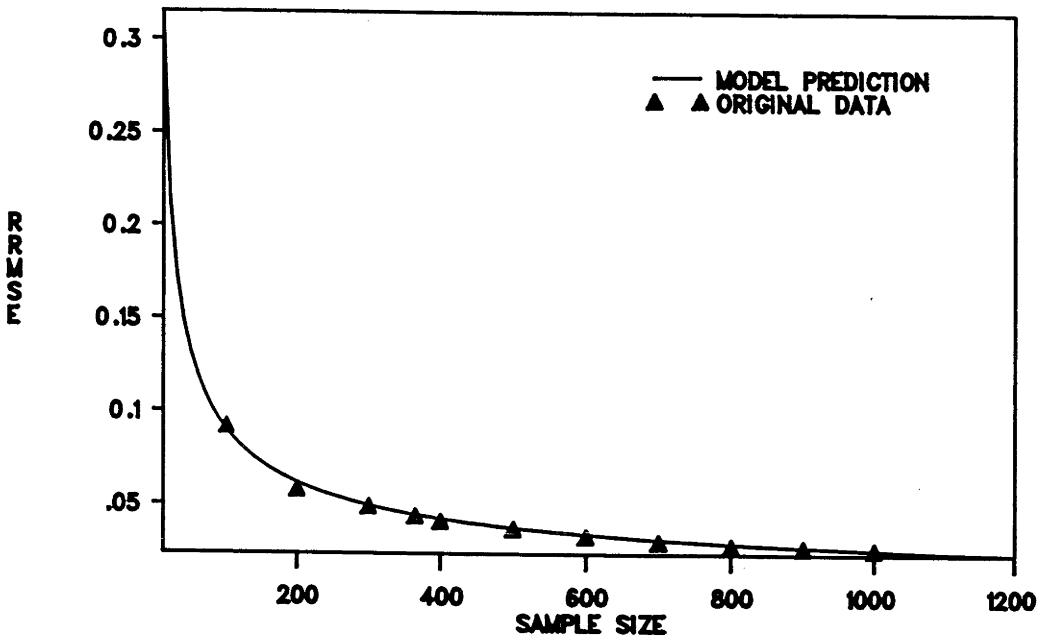

Figure 5.7: Fit of error model (5.23) to the RRMSE values at 99 percentile versus sample size for the 3-parameter gamma distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$

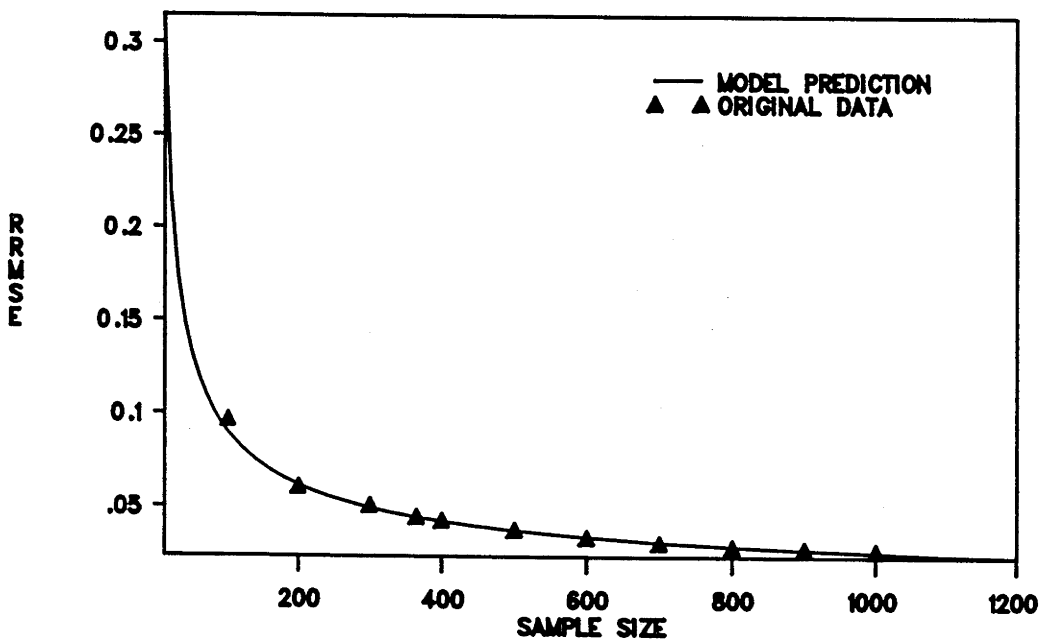

Figure 5.8: Fit of error model (5.23) to the RRMSE values at maximum percentile versus sample size for the 3-parameter gamma distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$

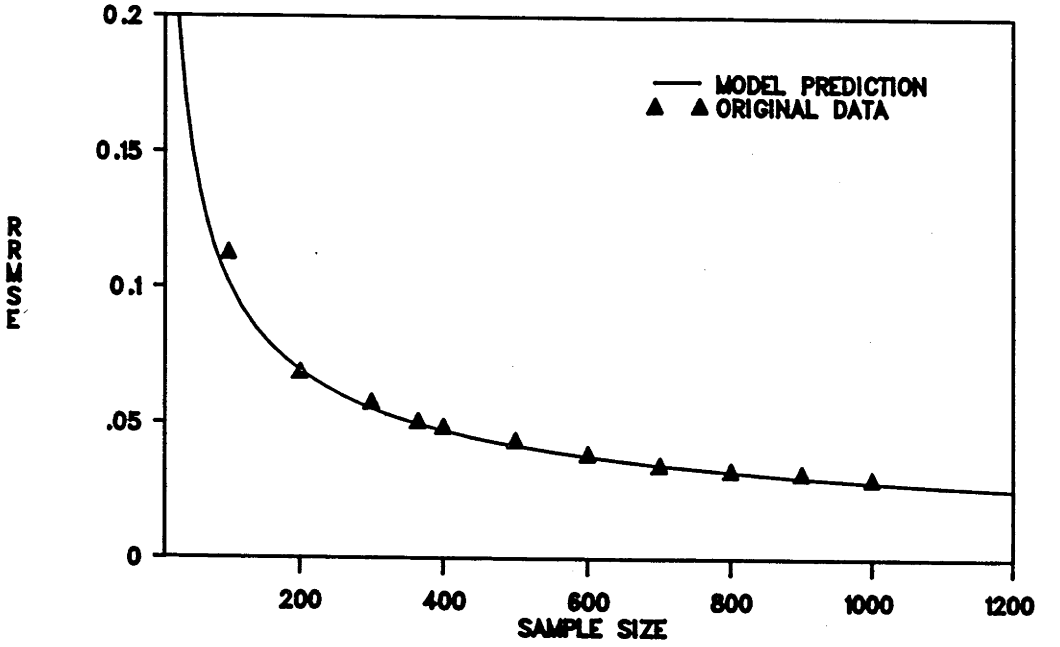

Figure 5.9: Fit of error model (5.24) to the RRMSE values at maximum percentile versus sample size for the 3-parameter gamma distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$

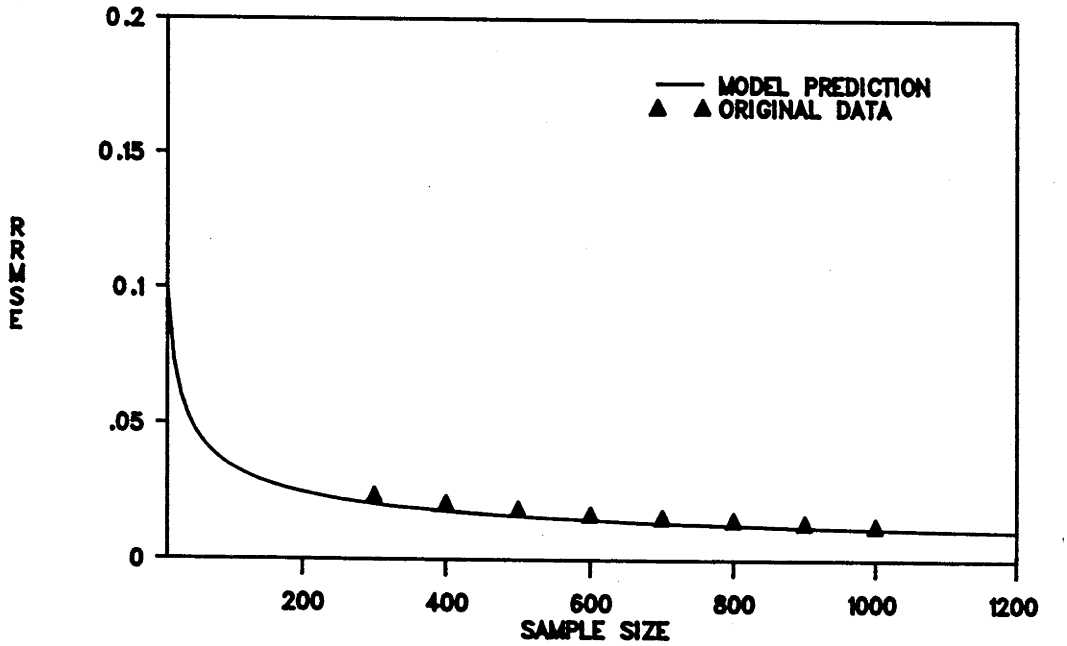

Figure 5.10: Fit of error model (5.25) to the RRMSE values at 98 percentile versus sample size for the 3-parameter Weibull distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$

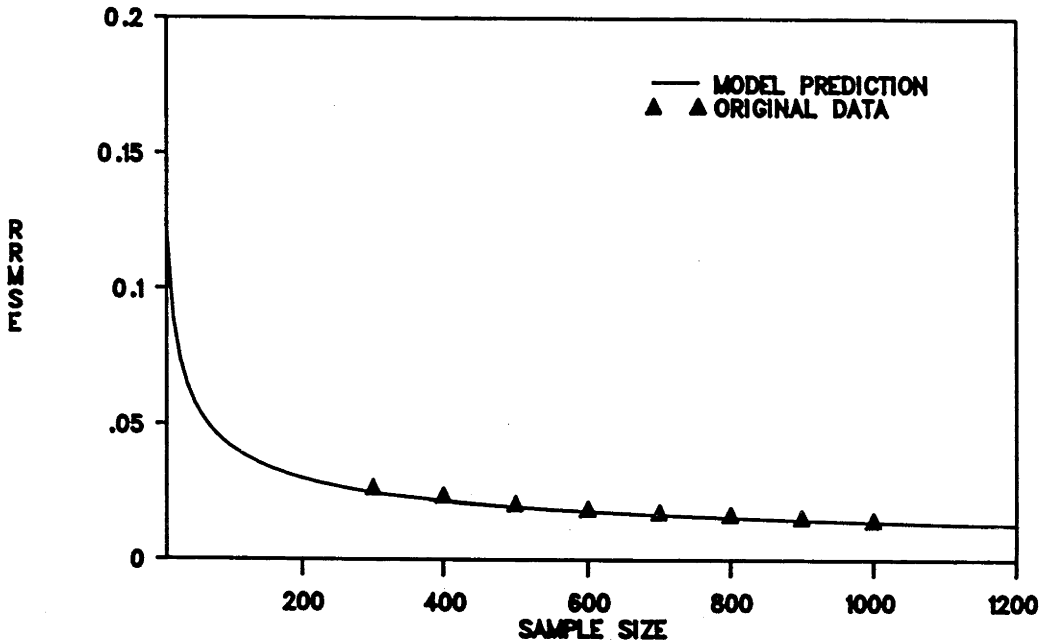

Figure 5.11: Fit of error model (5.25) to the RRMSE values at 99 percentile versus sample size for the 3-parameter Weibull distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$

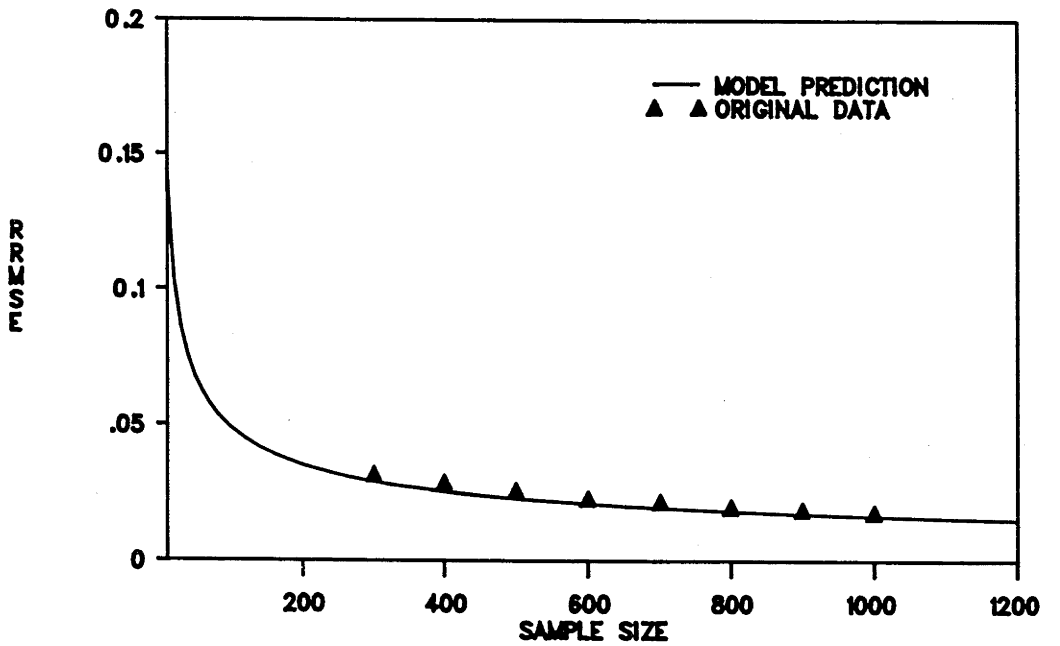

Figure 5.12: Fit of error model (5.25) to the RRMSE values at maximum percentile versus sample size for the 3-parameter Weibull distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$

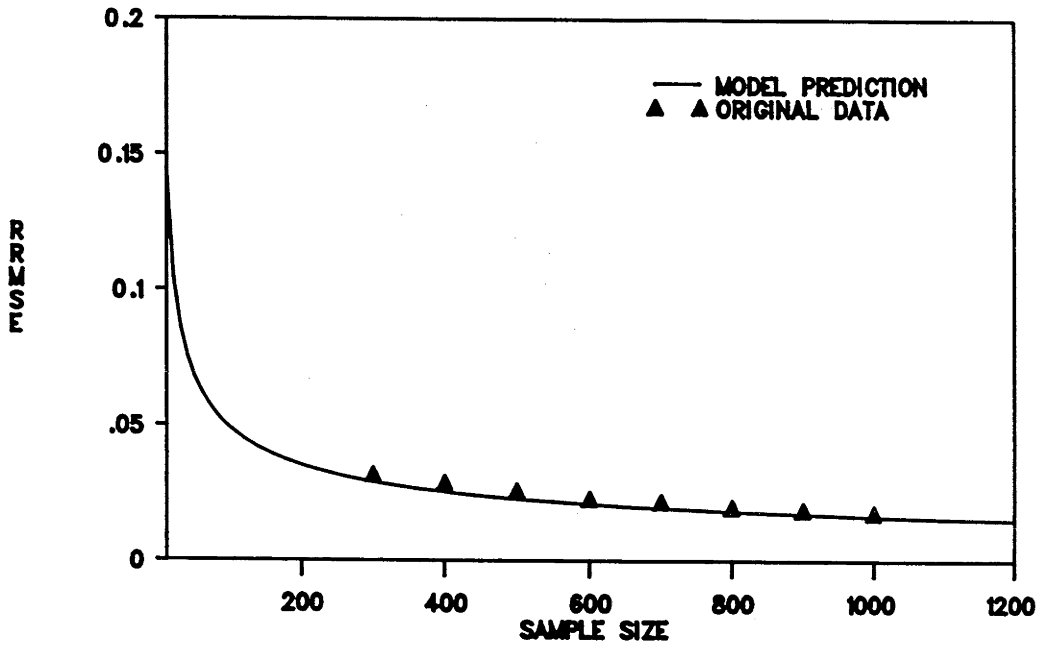

Figure 5.13: Fit of error model (5.27) to the RRMSE values at maximum percentile versus sample size for the 3-parameter Weibull distribution with $(\alpha, \beta, \gamma) = (2, 1, 1)$

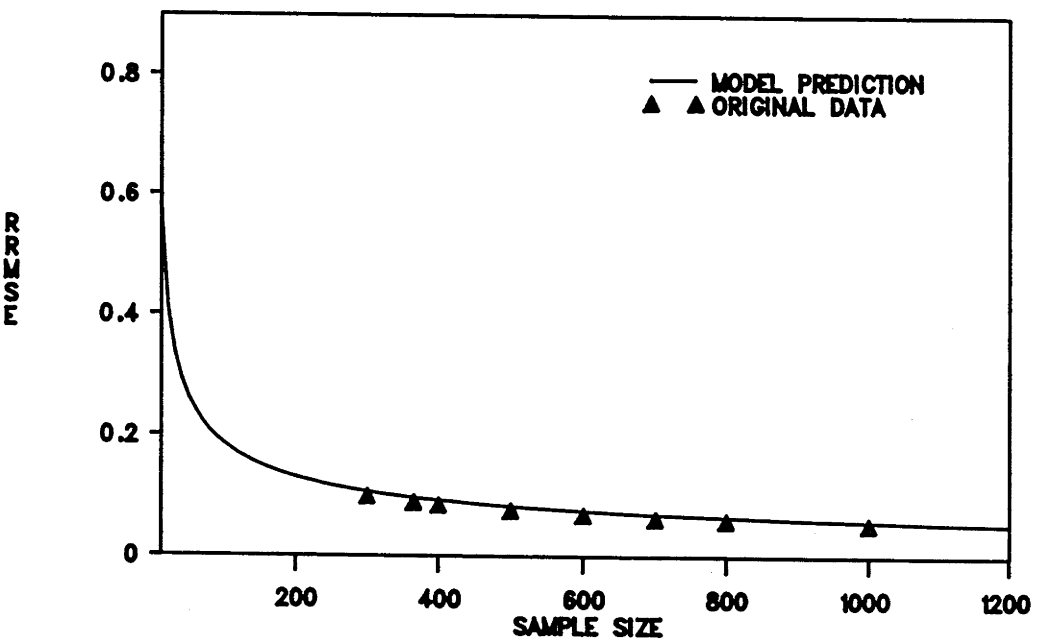

Figure 5.14: Fit of error model (5.26) to the RRMSE values at 98 percentile versus sample size for the 3-parameter lognormal distribution with $(\alpha, \beta, \gamma) = (0.9, 1, 1)$

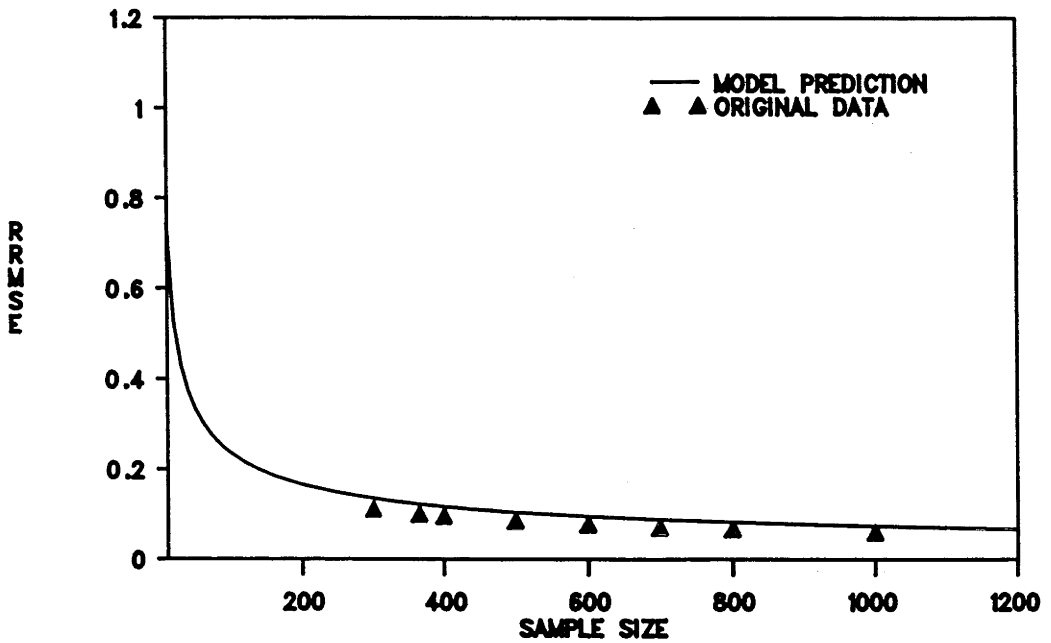

Figure 5.15: Fit of error model (5.26) to the RRMSE values at 99 percentile versus sample size for the 3-parameter lognormal distribution with $(\alpha, \beta, \gamma) = (0.9, 1, 1)$

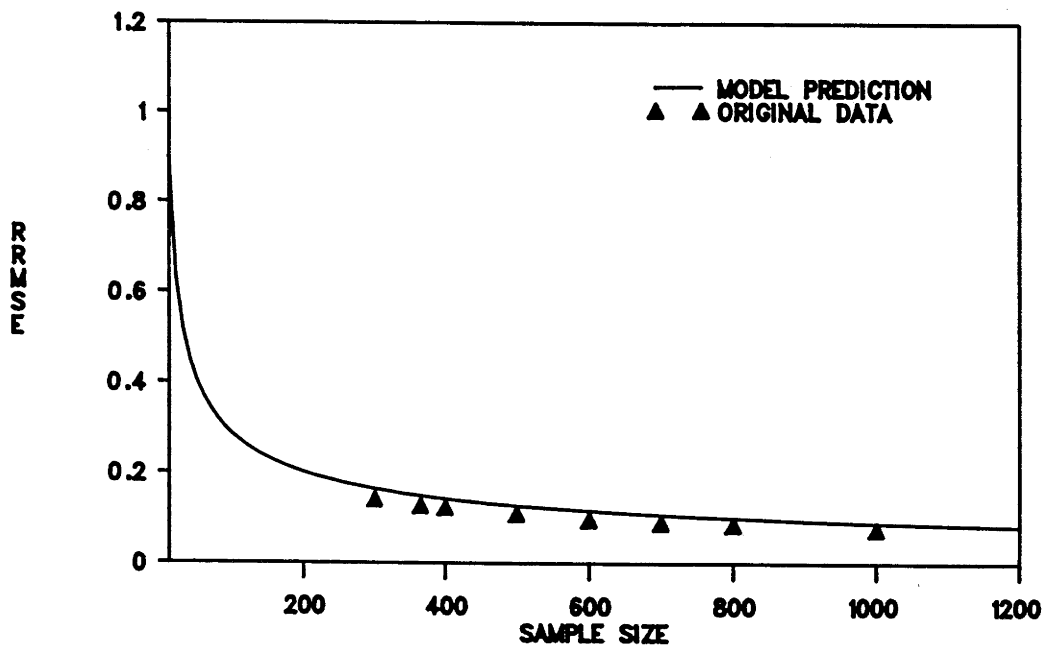

Figure 5.16: Fit of error model (5.26) to the RRMSE values at maximum percentile versus sample size for the 3-parameter lognormal distribution with $(\alpha, \beta, \gamma) = (0.9, 1, 1)$

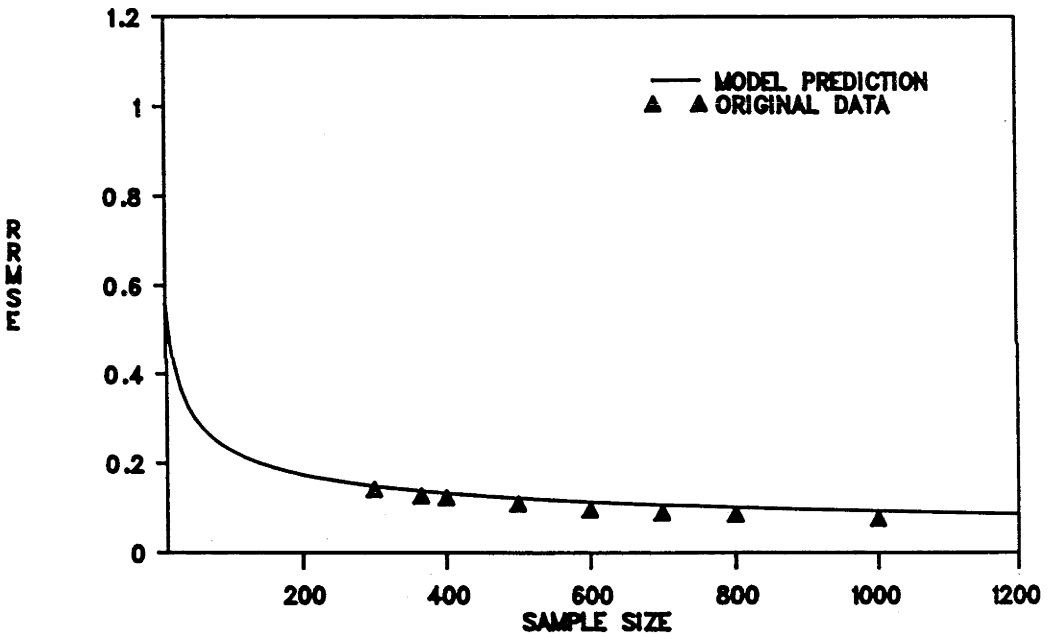

Figure 5.17: Fit of error model (5.28) to the RRMSE values at maximum percentile versus sample size for the 3-parameter lognormal distribution with $(\alpha, \beta, \gamma) = (0.9, 1, 1)$

TABLE 5.1
Descriptive statistics for estimated parameters of error model for the three-parameter gamma distribution

Parameter Name	Estimated Coefficient	Standard Error	T-Ratio (678 DF)	Standardized Coefficient
a_1	0.544	0.007	79.153	0.756
a_2	0.211	0.006	33.634	0.325
a_3	0.118	0.004	27.984	0.270
a_4	0.100	0.004	23.718	0.227
a_5	14.871	0.412	36.054	0.343
a_0	0.394	0.044	8.858	0.000

Note: DF denotes the degrees of freedom and $a_0 = logC$.

TABLE 5.2
Descriptive statistics for fit of error model to data for the three-parameter gamma distribution

F Test							
F-value	DF1	DF2	R^2	$\hat{\sigma}$	SR	VR	SAE
2076.394	5	678	93.87	0.075	-0.756^{-12}	0.006	38.078

Note: DF1 and DF2 denote the numeritor and denominator degrees of freedom, respectively. SR denotes the sum of residuals, VR the variance of residuals and SAE the sum of absolute errors.

TABLE 5.3
Descriptive statistics for estimated parameters of error model for the three-parameter Weibull distribution

Parameter Name	Estimated Coefficient	Standard Error	T-Ratio (504 DF)	Standardized Coefficient
a_1	0.470	0.012	39.040	0.381
a_2	1.146	0.020	57.295	0.562
a_3	0.387	0.085	45.300	0.444
a_4	0.363	0.082	44.284	0.432
a_5	19.317	0.802	24.081	0.233
a_0	-0.023	0.078	-0.296	0.000

Note: DF denotes the degrees of freedom and $a_0 = logC$.

TABLE 5.4
Descriptive statistics for fit of error model to data for the three-parameter Weibull distribution

F Test							
F-value	DF1	DF2	R^2	$\hat{\sigma}$	SR	VR	SAE
2037.314	5	504	95.29	0.127	-0.112^{-11}	0.016	52.584

Note: DF1 and DF2 denote the numeritor and denominator degrees of freedom, respectively. SR denotes the sum of residuals, VR the variance of residuals and SAE the sum of absolute errors.

TABLE 5.5
Descriptive statistics for estimated parameters of error model for the three-parameter lognormal distribution

Parameter Name	Estimated Coefficient	Standard Error	T-Ratio (588 DF)	Standardized Coefficient
a_1	0.499	0.009	56.798	0.496
a_2	0.974	0.012	78.612	0.683
a_3	0.112	0.005	23.789	0.208
a_4	0.054	0.005	11.503	0.100
a_5	23.757	0.451	52.635	0.456
a_0	0.988	0.057	17.358	0.000

Note: DF denotes the degrees of freedom and $a_0 = logC$.

TABLE 5.6
Descriptive statistics for fit of error model to data for the three-parameter lognormal distribution

F Test							
F-value	DF1	DF2	R^2	$\hat{\sigma}$	SR	VR	SAE
2550.303	5	588	95.59	0.077	-0.170^{-11}	0.006	35.187

Note: DF1 and DF2 denote the numeritor and denominator degrees of freedom, respectively. SR denotes the sum of residuals, VR the variance of residuals and SAE the sum of absolute errors.

Part III

DISCRIMINATION AND MISSPECIFICATION

Chapter 6

Discrimination Between Nested Two- and Three-parameter Distributions

6.1 Introduction

Several statistical criteria have been developed to discriminate among alternative parametric probability distributions. This chapter deals with discrimination between two- and three-parameter nested alternatives for three common shape-scale-location parametric distributions, namely the gamma, Weibull and lognormal distributions. These two- and three-parameter distributions have frequently been used to model air pollution and environmental quality data; for example, see Jakeman and Taylor (1989) and the references cited therein. In the Monte Carlo experiments, we evaluate the well-known likelihood ratio (LR) test, Akaike's (1974) Information Criterion (AIC), Schwarz's (1978) Information Criterion (SIC), the Chi-square test, and the Kolmogorov-Smirnov test. Using extensive Monte Carlo simulations from two- and three-parameter parent distributions, we investigate the performance of these tests and information criteria. The performance of the tests and criteria depends to some extent on the types of nested distributions being considered, the parametric values of the parent distributions, the confidence levels used (if applicable), and the sample sizes. The parameter space investigated covers an extensive range of values which might arise in practice. For an illustrative example, the sensitivity of the results to the values of the location and

shape parameters is evaluated.

Selection of an appropriate criterion should depend upon the intended use of the model. The practical usefulness of the techniques is illustrated by observing the errors of the models in fitting the upper percentiles of the parent distribution. Two sets of air pollution data from an urban airshed are used to examine the similarities and differences in fitting two- and three-parameter distributions where there is a preference for the more parsimonious model.

The chapter also considers the relationship between the LR test and the two information criteria. The former is an hypothesis test which implicitly assumes that one of the distributions being tested is true, while the latter makes no such assumption and attempts to discriminate among alternatives in terms of the maximized log-likelihood value, with an allowance made for the number of parameters and observations used in estimation. Since the LR test performs quite well, it is useful to interpret the equivalence of the test and the information criteria at a given confidence level in terms of a generalised information criterion which relates directly to the critical region of the LR test.

The plan of the chapter is as follows. In Section 2 the distribution functions and log-likelihood equations are presented. The discrimination criteria and loss functions are given in Sections 3 and 4, respectively. Sections 5 and 6 contain discussions of the simulation procedure and Monte Carlo results, respectively. An empirical application on hourly pollutant observations of β-scattering and nitrogen dioxide is outlined in Section 7. Some concluding remarks are given in Section 8.

6.2 The Distributions

Standardized probability density functions for the three-parameter gamma, Weibull and lognormal distributions for a random sample are given by:

Gamma:

$$f(x) = \frac{1}{\beta\Gamma(\alpha)}(\frac{x-\gamma}{\beta})^{\alpha-1}exp[-(\frac{x-\gamma}{\beta})] \tag{6.1}$$

Weibull:

$$f(x) = \frac{\alpha}{\beta}(\frac{x-\gamma}{\beta})^{\alpha-1}exp[-(\frac{x-\gamma}{\beta})^{\alpha}] \tag{6.2}$$

Lognormal:

$$f(x) = \frac{1}{\alpha\sqrt{2\pi}}(x-\gamma)^{-1}exp\{-\frac{[log(x-\gamma)-\beta]^2}{2\alpha^2}\}. \tag{6.3}$$

In equations (1), (2) and (3), α represents the shape parameter, β the scale parameter, γ the location parameter, and Γ is the gamma function. The two-parameter versions of the density functions of the gamma, Weibull and lognormal distributions are the same as in (1), (2) and (3), with $\gamma = 0$ in each case. In the above equations, $\beta > 0$, $\alpha > 0$ and γ is less than the minimum observed sample value.

The properties of these three distributions and the asymptotic behaviour of estimators depend very heavily on the values of the parameters, particularly that of the shape parameter. Figures 6.1 and 6.2 show that the resulting density functions of the gamma and Weibull distributions are similar to the exponential distribution at $\alpha = 1$, reverse 'J' shaped for $\alpha < 1$, and 'bell' shaped for $\alpha > 1$. Figure 6.3 shows that the curves for the lognormal distribution change from nearly symmetric to heavily skewed as α is increased from 0.3 to 1.2. These values span a large range of shapes which arise in the analysis of real data, such as air pollutant concentrations. In order to assess the different criteria for discriminating among competing descriptions of the data, the shape parameter is examined over an extensive range of possible cases where the density functions vary from being skewed to symmetric.

The maximized value of the likelihood function is an essential statistic employed in many criteria used to discriminate among alternative models. For a sample $x_1, x_2, \ldots, x_n$ of n independently and identically distributed random observations, the log-likelihood functions for the three-parameter gamma, Weibull and lognormal distributions are given as follows:

Gamma:

$$logL = -n\alpha log\beta - nlog\Gamma(\alpha) + (\alpha-1)\sum_{i=1}^{n} log(x_i-\gamma) - \sum_{i=1}^{n}(\frac{x_i-\gamma}{\beta}) \tag{6.4}$$

Weibull:

$$logL = nlog\alpha - n\alpha log\beta + (\alpha - 1)\sum_{i=1}^{n} log(x_i - \gamma) - \sum_{i=1}^{n}(\frac{x_i - \gamma}{\beta})^{\alpha} \qquad (6.5)$$

Lognormal:

$$logL = -\frac{n}{2}log(2\pi\alpha^2) - \sum_{i=1}^{n} log(x_i - \gamma) - \frac{1}{2\alpha^2}\sum_{i=1}^{n}[log(x_i - \gamma) - \beta]^2. \qquad (6.6)$$

The parameters of the three log-likelihood functions are estimated by maximum likelihood methods. Since the general maximum likelihood procedure for the three-parameter gamma and Weibull distributions will frequently fail to converge when the (unknown) shape parameter is less than or equal to unity, a computationally efficient approach that circumvents this problem is used (for further details, see Bai et al. (1989)).

6.3 Discrimination Criteria

Let $x_1, x_2, \ldots, x_n$ represent a random sample of n observations. Interest here lies in discriminating among nested two- and three-parameter distributions in which the null hypothesis of interest is $H_0 : \gamma = 0$ against the alternative $H_1 : \gamma \neq 0$. The standard LR test can be employed for this problem. Denoting the maximized values of the two- and three-parameter variants of a particular log-likelihood function as $logL_0$ and $logL_1$, respectively, the LR test can be expressed as:

$$LR = -2(logL_0 - logL_1) \overset{a}{\sim} \chi^2(1) \qquad (6.7)$$

under the null hypothesis that the location parameter is zero. The AIC and SIC may be expressed, respectively, as:

Choose the $\left\{ \begin{array}{c} 2 \\ 3 \end{array} \right\}$ parameter distribution if

$$AIC : logL_0 - 2\left\{ \begin{array}{c} > \\ < \end{array} \right\} logL_1 - 3 \qquad (6.8)$$

$$SIC : logL_0 - logn \left\{ \begin{matrix} > \\ < \end{matrix} \right\} logL_1 - 3logn/2. \tag{6.9}$$

When H_0 holds for testing a two-parameter distribution against a three-parameter alternative, rearranging (8) and (9), and defining $\Delta L = logL_0 - logL_1$, corresponds to choosing the $\left\{ \begin{matrix} 2 \\ 3 \end{matrix} \right\}$ parameter distribution if

$$AIC : -\Delta L \left\{ \begin{matrix} < \\ > \end{matrix} \right\} 1 \tag{6.10}$$

$$SIC : -\Delta L \left\{ \begin{matrix} < \\ > \end{matrix} \right\} \frac{logn}{2}. \tag{6.11}$$

Since the information criteria and the LR test are based on the maximized value of the likelihood function, it is possible to compare the information criteria and hypothesis test in terms of the probability of accepting the underlying null distribution. By comparison with (10) and (11), the LR test will accept H_0 if

$$LR : -\Delta L < \frac{c}{2} \tag{6.12}$$

where c is the critical value of the $\chi^2(1)$ statistic. It is easy to see that an equivalence among the LR test, AIC and SIC can be found when the nested model is regarded as the true distribution. Use of the AIC criterion is equivalent to the LR statistic at the 84.2 per cent confidence level (i.e. when $c = 2$) and SIC, for a sample size of 365, is equivalent to the LR statistic at the 98.5 per cent confidence level (i.e. when $c = logn$). When the sample size is decreased to $n = 100$, SIC is equivalent to the LR statistic at the 96.81 per cent confidence level, but will be increased to the 99.14 per cent confidence level for $n = 1000$.

The equivalence demonstrated above could be used to construct generalised information criteria (GIC) which, when the nested distributions is true, is equivalent to the

LR test at different confidence levels. In this chapter, we use two such criteria, GIC1 and GIC2 which can be regarded as LR analogues at the 40 per cent and 99 per cent confidence levels, respectively, and indicate two extreme cases: the lowest and highest confidence levels that might reasonably be considered in applications. These two cases can also help to illustrate the tradeoff between the confidence level and power of a test. An appropriate confidence level for air pollutant concentrations will be recommended in a later section when examining real data.

The performance of two well-known procedures for testing goodness of fit are also considered, namely the chi-square (CHI) test and Kolmogorov-Smirnov (KS) test. Classifying the n observations into k categories, the chi-square statistic is of the form (see Pearson (1900)):

$$CHI = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} \tag{6.13}$$

which has an asymptotic χ^2 distribution with $(k - l - 1)$ degrees of freedom when H_0 holds. The p_i are hypothetical probabilities, the f_i are empirical frequencies and l is the number of parameters estimated for each distribution (for further details, see Kendall and Stuart (1979)). For the experiments conducted in Section 6 below, $k = 10$ and $l = 2$ or $l = 3$. The KS test, which is defined in terms of the maximum absolute difference between the sample distribution function $S_n(x)$ and the hypothetical distribution function $F_0(x)$ (see e.g. Bury (1975, p. 204)), is given by

$$D_n = \sup_x |S_n(x) - F_0(x)|. \tag{6.14}$$

Large observed values of the D_n statistic lead to rejection of the hypothesis $F_0(x)$.

6.4 Loss Functions

An assessment of the performance of different tests and criteria requires some form of loss function or performance criterion which should rely on the nature of the problem and the major purpose of the application. Standard performance criteria for assessing

nested hypothesis tests are size and power. In this Monte Carlo study, loss functions recommended for assessing air quality models have also been chosen (see Fox (1981)) to establish the effect of discrimination criteria on the intended use of the model. These functions are the relative bias (BIAS) and the relative root mean square error (RRMSE) which are evaluated at the upper percentiles of the distributions. For an estimate $\hat{q}_i$ of a quantity of interest q, these loss function are defined in terms of deviations from the true or parent value q in each replication of the Monte Carlo experiments. The definitions used for the loss functions are:

$$BIAS(q) = \frac{1}{N}\sum_{i=1}^{N}(\frac{\hat{q}_i - q}{q}) \tag{6.15}$$

$$RRMSE(q) = [\frac{1}{N}\sum_{i=1}^{N}(\frac{\hat{q}_i - q}{q})^2]^{0.5} \tag{6.16}$$

where N is the number of replications of the experiment. For present purposes, the quantity q denotes the upper percentiles of the underlying distributions.

6.5 Simulation Procedure

In order to assess the various criteria for discriminating between models over different independently and identically distributed random samples, simulation over an extensive range of possible cases is considered. For all parameter sets in the tables and figures reported here, one thousand simulation experiments are processed. The main sample size used is $n = 365$, since it represents a common case: a full year of 24-hour average observations. For two extreme cases associated with possible applications, $n = 100$ and $n = 1000$ are considered here as illustrative examples. The shape parameters take the values 0.5, 1, 2, 4, 6 for the gamma distribution; 0.5, 1, 2, 3, 4 for the Weibull distribution; and 0.3, 0.5, 0.7, 0.9, 1.2 for the lognormal distribution. It should be noted that the lognormal distribution has the opposite behaviour to the other two as the shape parameter is increased. In all of the cases investigated in this chapter, the

arbitrary scale parameter is set at unity. In most cases the location parameter is also set to unity, but the sensitivity of our results to other values is also examined.

The random sample generators used for the Monte Carlo experiments are DRNGAM for the gamma, DRNLNL for the lognormal and DRNWIB for the Weibull distribution. These are available as subroutines in the International Mathematical and Statistical Library (IMSL) in version 1.0 of April 1987. The same seed number (1234) is used to obtain the first random sample of the first of the 1000 simulations. Varying the initial seed produces similar results to those given in the chapter. For maximum likelihood estimation, a golden section search algorithm is used with final estimates being accepted when the relative error between two successive approximations is less than 10^{-6}. Two subroutines, namely DCHIGF and DKSONE, are chosen from IMSL to perform the CHI and KS tests. All results are obtained on a VAX8700 mainframe computer at ANU.

6.6 Monte Carlo Results

Consider initially an investigation of the performance of the discrimination criteria for random samples of size $n = 365$ from the gamma distribution. In this chapter, the scale parameter is always set at $\beta = 1$, and the values of the location parameter are $\gamma = 0$ or $\gamma = 1$. It should be noted that, for a fixed value of the location parameter, it becomes increasingly difficult to reject the false null hypothesis that $\gamma = 0$ as the value of the shape parameter increases (i.e. power decreases). Table 6.1 shows the results in two situations: first, the null hypothesis H_0 is true, so that the samples for each Monte Carlo experiment are taken from a two- parameter distribution; second, the alternative hypothesis H_1 is true so that the samples are taken from a three-parameter distribution.

When $\alpha \geq 2$, $\gamma = 0$ and $n = 365$, the empirical performance of the LR test is not significantly different from the nominal level of 0.05 given by asymptotic theory. The empirical probabilities vary only slightly with the shape parameter and with the initial

seed used for the random number generator. Acceptance rates for AIC, SIC, GIC1 and GIC2 are also similar to those expected from the derived equivalent (LR) confidence intervals reported in Section 3, namely the 84.2, 98.5, 40.0 and 99.0 per cent levels, respectively. The CHI test has rejection frequencies which are very similar to those predicted by theory, while the KS test rarely rejects the true null hypothesis.

The power of any of the first five tests is inversely related, in general, to the acceptance rate when $\alpha \geq 2$. The lower the confidence level imposed for acceptance of the null hypothesis, the higher is the power. Quantifying this inverse relationship for different parameter values is a major concern in terms of how often we can expect underfitting of two-parameter distributions to occur in samples taken from three-parameter parent distributions. For a fixed value of the location parameter, power decreases as the shape parameter is increased.

Notice that, for the gamma distribution when $\alpha = 0.5$ and $\alpha = 1.0$, the LR test and the four discrimination criteria tend to overfit, a three-parameter distribution being generally preferred when H_0 is true (i.e. $\gamma = 0$), especially for $\alpha = 0.5$, and always preferred when H_0 is false (i.e. $\gamma = 1$). This behaviour is due to the fact that the distributions approach the exponential when $\alpha \leq 1$, and likelihood values increase when the location parameter is set near the first order statistic. However, the CHI test has empirical sizes that are unaffected by whether the value of the shape parameter is less than or greater than unity, and the KS test still rarely rejects a true null hypothesis.

Figure 6.4 portrays the cumulative frequency over 1000 experiments of the differences between the maximized log-likelihood values of the two- and three-parameter gamma distributions when the samples are taken from a three-parameter parent distribution. The figure demonstrates why power decreases as the value of the shape parameter increases. Figure 6.5 shows the differences when the samples are taken from a two-parameter parent distribution. Whatever the value of the shape parameter, the differences are large in only a small proportion of the 1000 cases.

Consider the power of the LR test at the 95 per cent confidence level for different

values of the location parameter. Figure 6.6 shows the results for $\alpha = 2, 4$ and 6. For $\alpha = 2$, the power of the LR test is high for quite low location values; for example, power is 0.98 when $\gamma = 0.34$. Power is also 0.98 for the combinations $(\alpha = 4, \gamma = 2.4)$ and $(\alpha = 6, \gamma = 6.5)$.

Table 6.1 also provides the results for rejection probabilities of the null hypothesis and powers of the tests and discrimination criteria over a range of parameter values for the Weibull and lognormal distributions. The conformity with theory of the LR test and AIC, SIC, GIC1 and GIC2 is good for the Weibull distribution when $\alpha \geq 2$ and for the lognormal distribution for all values of α when the sample size is $n = 365$. The empirical sizes of the CHI and KS tests of the Weibull and lognormal distributions are very similar to those of the gamma distribution for all values of the shape parameter. Sizes for the CHI test are close to the nominal size of 0.05, while the sizes for the KS test are almost zero in all cases. Not surprisingly, the powers of the CHI and KS tests are much lower than for the LR test.

The power of the LR test, when applied at the 95 per cent confidence level for the Weibull and lognormal distributions, is shown in Figures 6.7 and 6.8, respectively, as a function of the shape and location parameters. Compared with the case of the gamma distribution in Figure 6.6, a similar pattern of power as a function of the shape and location parameters is observed for the Weibull distribution. High power will be obtained when the shape parameter is 2 for quite low values of the location parameter, as shown in Figure 6.7. For example, power is 0.98 when $\alpha = 2$ and $\gamma = 0.17$. When the value of the shape parameter is increased, large values of the location parameter will be required to maintain high power. For instance, power is also 0.98 for the combinations $(\alpha = 3, \gamma = 0.49)$ and $(\alpha = 4, \gamma = 1.48)$. Figure 6.8 provides similar results for the lognormal distribution, except that the lognormal has the opposite behavour to the other two as the shape parameter is increased. For power to be 0.98, the combinations of shape and location parameters are $(\alpha = 0.9, \gamma = 0.41)$, $(\alpha = 0.7, \gamma = 1.25)$ and $(\alpha = 0.5, \gamma = 3.91)$.

We now turn to an evaluation of the performance of the discrimination criteria for other sample sizes using the gamma distribution as a guide. Table 6.2 provides the analogous results to those in Table 6.1 where the sample sizes are $n = 100$ and $n = 1000$. As expected, at $n = 1000$ the criteria perform according to asymptotic theory in terms of correctly accepting two-parameter models, since a similar result is achieved at the lower sample size of $n = 365$. The empirical sizes for the KS test are still very low, and the powers of the CHI and KS tests are considerably lower than for the LR test. Again, power declines with the shape parameter for a fixed value of the location parameter, but at the larger sample size the power is much higher for any specific shape parameter and criterion. At $n = 100$, the acceptance rates of two-parameter distributions for the LR and CHI tests and the four criteria are lower than those predicted by theory, while power is consistently lower than at $n = 365$ for any specific shape parameter and criterion. The acceptance rates for the KS test vary with the value of the shape parameter, being too high when $\alpha = 2$ and too low when $\alpha = 4$ and $\alpha = 6$. The powers of the CHI and KS tests are considerably lower than those of the LR test for all values of the shape parameter.

6.7 Application to Models of Air Pollution

Hourly pollutant observations of β-scattering and nitrogen dioxide for Melbourne, Australia, are available at state site numbers 11 (Museum), 27 (Alphington), 34 (Dandenong) and 81 (Camberwell) for the years indicated in Tables 6.3 and 6.4. These data are converted into samples of 24-hour averages for those years and sites where the number of the resultant daily samples available is greater than 100. These data sets are used to illustrate an application of the discrimination criteria for the situation where the intended use of the model is predicting extreme concentrations and historical practice suggests there is a preference for two-parameter models.

Table 6.3 gives results for β-scattering when the two- and three-parameter lognormal distribution is estimated. β-scattering here refers to light scattering by suspended

aerosol as measured by an intergrating nephelometer (see Finlayson-Pitts and Pitts (1986)). Notice that, in general, the probabilities of rejecting the two-parameter lognormal model are very high and the maximized log-likelihood values are much lower for the two-parameter lognormal distribution than for its three-parameter counterpart. Indeed, the lognormal distribution yields much larger maximized log-likelihood values for the three-parameter models than the gamma and Weibull distributions in 18 of the 20 cases considered. Omitting the single case in 1977 for site 11 where the two- and three-parameter log-likelihoods are equal, the minimum value of the acceptance threshold for the null hypothesis is 0.9860 for site 11 in 1976. If the parent distribution is the three-parameter lognormal, then fitting the two-parameter lognormal to samples from this parent yields substantial errors which can be quantified by simulation. For example, for the three-parameter lognormal distribution, the RRMSE obtained by simulation over 1000 experiments is 0.092 for the maximum percentile (MAX1), 0.078 for the second-highest percentile (MAX2), and 0.058 for the 98'th percentile (98%), while for the the two-parameter lognormal these values are 0.136, 0.112 and 0.076, respectively. Admittedly, if we do not wish to risk obtaining errors of these magnitudes in such air quality applications, we should set our acceptance threshold for the null hypothesis below 98.6 per cent.

In order to fine-tune the estimate of where this threshold should be, given a preference for two-parameter models, consider the results for daily nitrogen dioxide samples in Table 6.4. The three-parameter gamma and lognormal distributions have the highest maximized log-likelihood values. However, the three-parameter lognormal distribution generally has a negative location parameter, which is regarded as physically unrealistic. Let us, therefore, assume that the gamma distribution is appropriate. The simulation results reported in Table 6.1 indicate that, even when the parent distribution is a two-parameter model, the probability of rejecting the two-parameter gamma distribution with the LR test is unity when the shape parameter is less than unity. In such cases, the true underlying model may not be determined even when the estimate of the location parameter for the three-parameter gamma distribution is very small.

Let us now re-examine the data sets and evaluate the errors in percentiles when we obtain an acceptance threshold below the value of .9860 found to be too high in the β-scattering case. The 1978 data set at site 11 yields an acceptance threshold of 0.9782. Again we can calculate the errors assuming that the three-parameter gamma is the underlying parent distribution. The RRMSE values for MAX1, MAX2 and 98% for fitting the three-parameter gamma distribution are 0.063, 0.060 and 0.055, respectively, compared with 0.081, 0.075 and 0.065, respectively, when fitting the two-parameter gamma distribution. These results provide further information as to where to set the confidence levels, given the errors that can be tolerated. If there is seen to be a strong need to use a two-parameter model, such as might be set by historical precedent, then from the results presented here, it can be observed how often and by how much the use of such a model is likely to exceed tolerable errors.

Any criterion used to discriminate between nested models will involve a trade-off between acceptance of the true null hypothesis and rejection of the false null hypothesis. What level of BIAS should be chosen for overfitting? This should depend on answers to the following two basic questions: (i) Under what conditions would it be inaccurate to assume that the true model is the two-parameter null? (ii) When would it be inaccurate to assume that the true model is the three-parameter alternative? Of course, the precise answers depend on the acceptable levels of inaccuracy. Basically, for the first question, inaccuracy is greatest for those parameter sets where the powers of the discrimination criteria are around unity. The answer to the second question is when the information content of the sample is too low to give reasonably efficient estimates of the three parameters.

The answers given above can be refined in specific cases. Consider, for example, predicting the upper percentiles of the underlying parent distribution. This is a motivation in analysing data sets for environmental quality. Environmental guidelines for air and water pollutants can be written in terms of allowable excesses of some extreme concentration. In this chapter we confine attention to the annual maximum concentration MAX1, the second highest value MAX2, and 98% values. More detailed

results for the comparative errors in fitting two- and three-parameter alternatives to the distributions with parameters within the range of Table 6.1 are given in Bai et al. (1990). However, some indication of the errors is warranted here. We use the gamma distribution as a guide and begin with the situation where the underlying parent is a three-parameter distribution. For $\alpha = 2$, the comparative errors begin to diverge for $\gamma > \frac{1}{2}$. For instance, when $\alpha = 2$, $\gamma = 1$ and $n = 365$, the RRMSE of MAX1, MAX2 and 98% is more than double that of the three-parameter estimates when the location parameter is not estimated but is set to zero. When the parent distribution is two-parameter and the sample size is 365, there is little additional error in fitting a three-parameter over a two-parameter model.

6.8 Concluding Remarks

The purpose of this chapter has been to discriminate between two- and three-parameter nested alternatives for the gamma, Weibull and lognormal distributions. Monte Carlo experiments were conducted to evaluate the likelihood ratio test, Akaike's information criterion, Schwarz's information criterion, the Chi-square test and the Kolmogorov-Smirnov test. The performance of the tests and criteria was shown to depend on the types of nested distributions under consideration, the parametric values of the parent distributions, the confidence levels used (if applicable), and the sample sizes. The practical usefulness of the techniques was illustrated by observing the errors of the models in fitting the upper percentiles of the parent distribution. Two sets of air pollution data, namely hourly pollutant observations of β-scattering and nitrogen dioxide, from an urban airshed were used to examine the similarities and differences in fitting two- and three-parameter distributions where historical practice suggests there is a preference for the more parsimonious model.

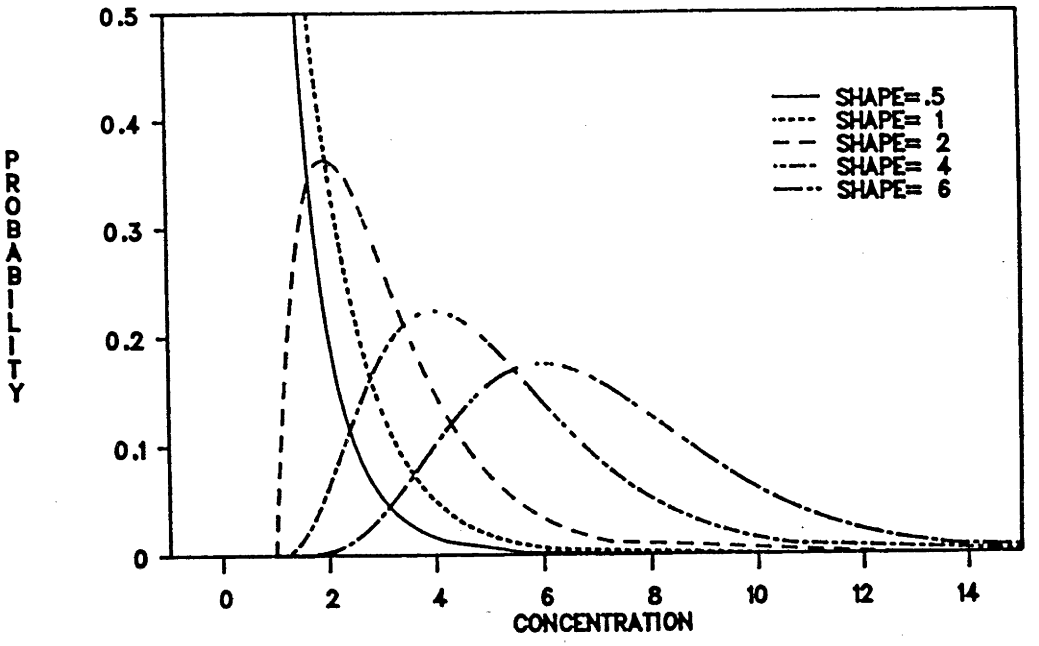

Figure 6.1: Profile of the gamma distribution for a range of shape parameters and unit scale and location parameters

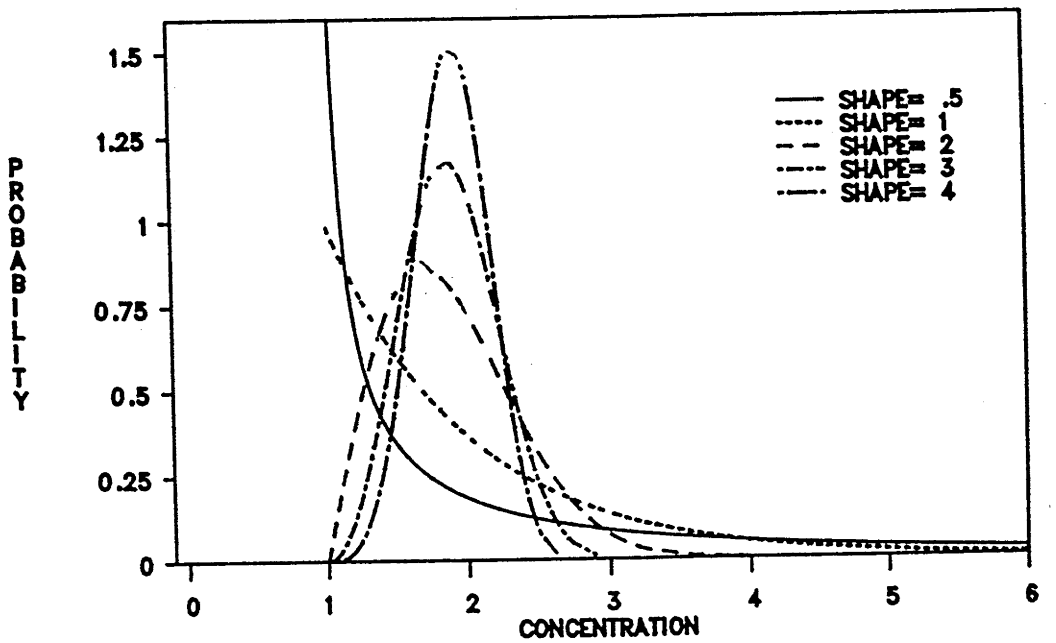

Figure 6.2: Profile of the Weibull distribution for a range of shape parameters and unit scale and location parameters

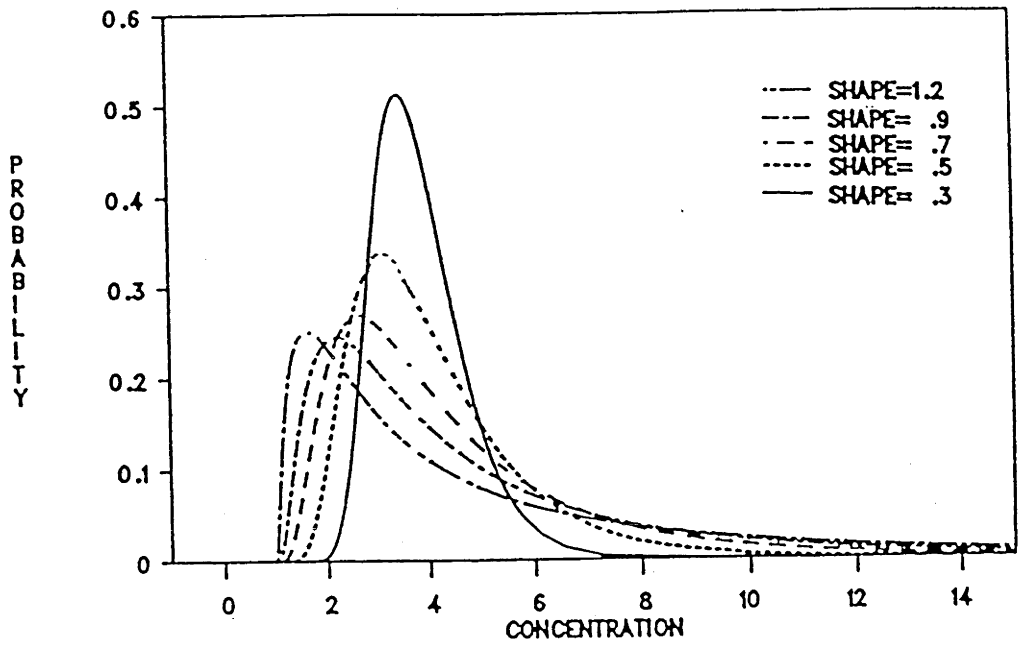

Figure 6.3: Profile of the lognormal distribution for a range of shape parameters and unit scale and location parameters

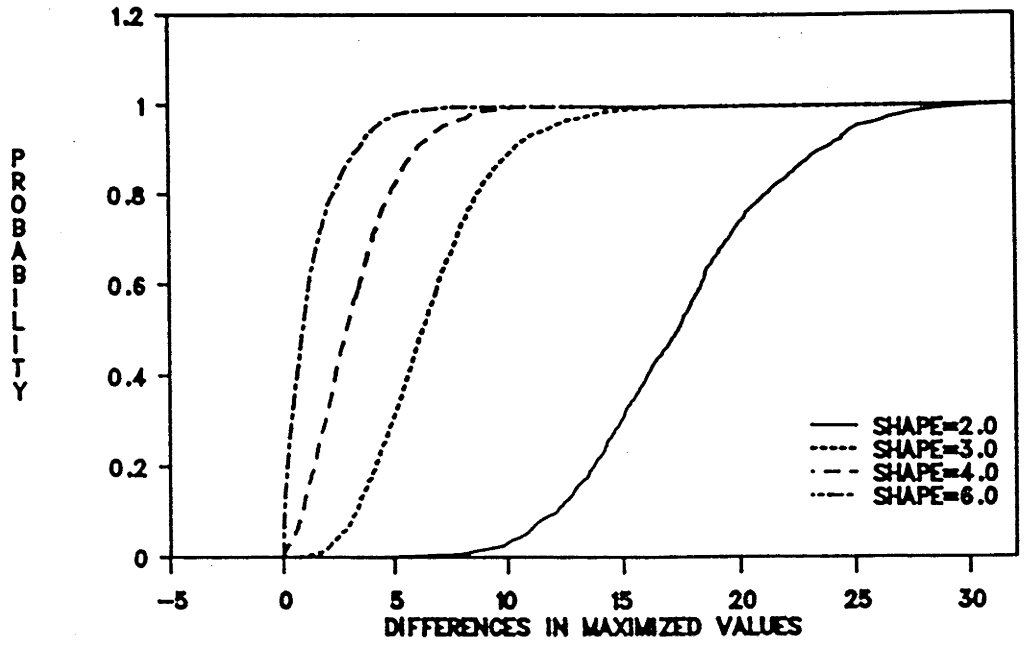

Figure 6.4: Cumulative frequency over 1000 experiments of the differences between the maximized log-likelihood values of the 2- and 3-parameter gamma distributions when the samples are taken from a 3-parameter parent distribution

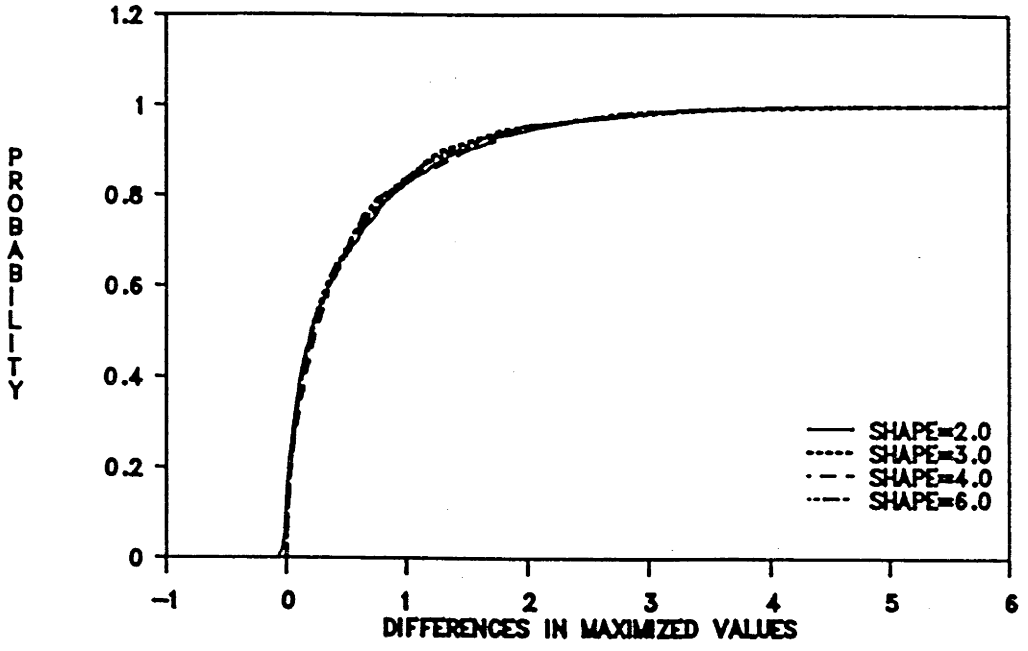

Figure 6.5: Cumulative frequency over 1000 experiments of the differences between the maximized log-likelihood values of the 2- and 3-parameter gamma distributions when the samples are taken from a 2-parameter parent distribution

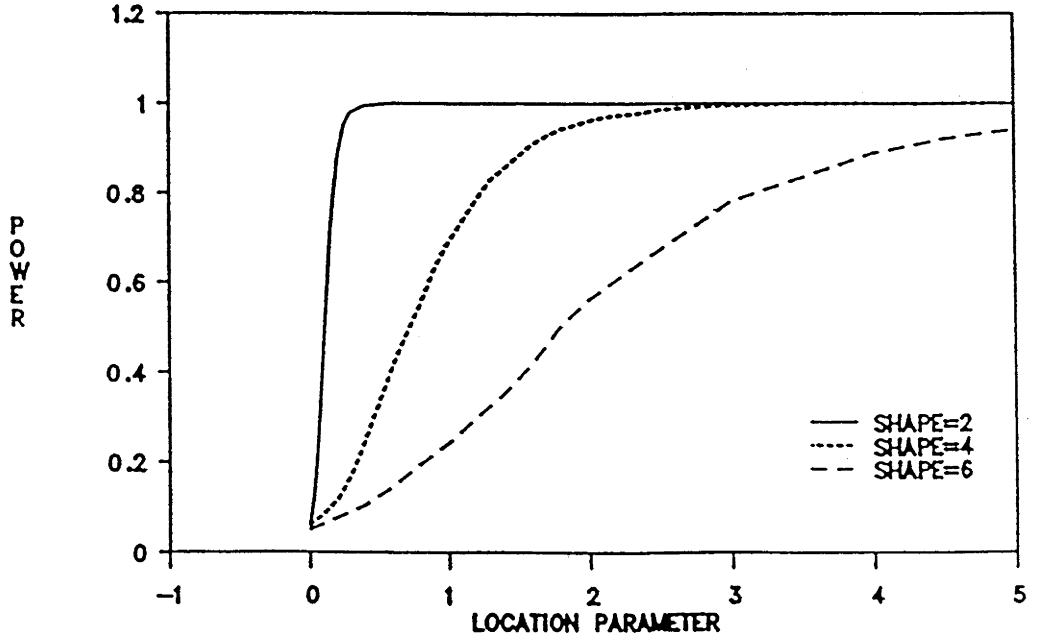

Figure 6.6: The power of the LR test at the 95 per cent confidence level for different values of the location parameter γ for testing between 2- and 3-parameter gamma distributions

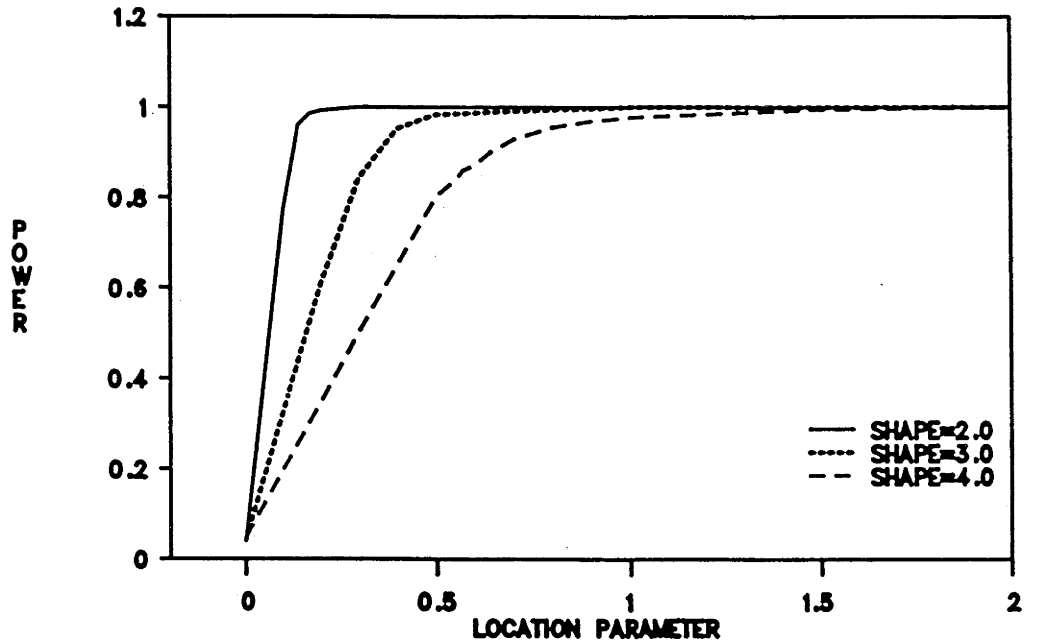

Figure 6.7: The power of the LR test at the 95 per cent confidence level for different values of the location parameter γ for testing between 2- and 3-parameter Weibull distributions

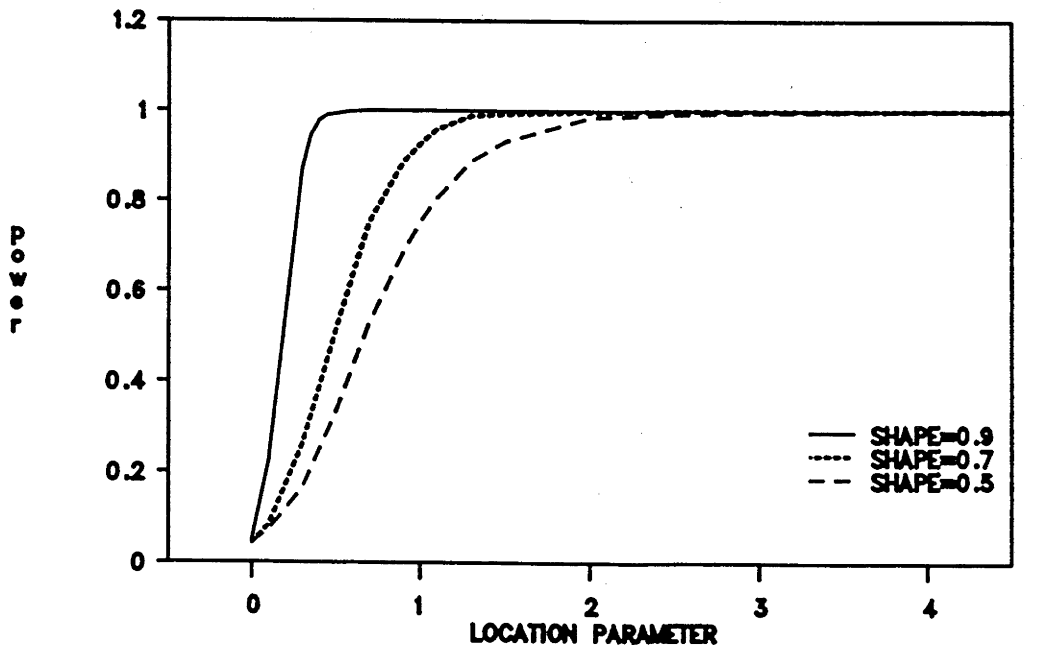

Figure 6.8: The power of the LR test at the 95 per cent confidence level for different values of the location parameter γ for testing between 2- and 3-parameter lognormal distributions

TABLE 6.1

Probabilities of rejecting the null hypothesis that $\gamma = 0$ using seven tests and discrimination criteria over 1000 replications of random samples of size $n = 365$ ($\beta = 1$)

		Gamma					Weibull					Lognormal				
		Values of Shape Parameter α					Values of Shape Parameter α					Values of Shape Parameter α				
True Distribution	Criteria	0.5	1.0	2.0	4.0	6.0	0.5	1.0	2.0	3.0	4.0	1.2	0.9	0.7	0.5	0.4
Two Parameter ($\gamma = 0$)	LR	1.000	0.332	0.061	0.061	0.050	1.000	0.217	0.041	0.045	0.049	0.053	0.049	0.048	0.043	0.043
	AIC	0.883	0.627	0.175	0.164	0.167	1.000	0.506	0.129	0.136	0.192	0.163	0.153	0.146	0.145	0.140
	SIC	0.904	0.148	0.017	0.020	0.015	1.000	0.087	0.010	0.010	0.043	0.017	0.012	0.016	0.015	0.016
	GIC1	1.000	0.884	0.573	0.632	0.589	1.000	0.843	0.574	0.581	0.601	0.588	0.598	0.596	0.597	0.597
	GIC2	1.000	0.107	0.010	0.013	0.012	0.999	0.059	0.007	0.003	0.008	0.011	0.009	0.009	0.010	0.010
	CHI	0.055	0.058	0.056	0.056	0.053	0.062	0.062	0.062	0.062	0.062	0.062	0.062	0.062	0.062	0.062
	KS	0.001	0.001	0.000	0.000	0.000	0.001	0.001	0.001	0.001	0.001	0.000	0.000	0.000	0.000	0.000
Three Parameter ($\gamma = 1$)	LR	1.000	1.000	1.000	0.690	0.243	1.000	1.000	1.000	1.000	0.977	1.000	1.000	0.997	0.925	0.752
	AIC	1.000	1.000	1.000	0.865	0.462	1.000	1.000	1.000	1.000	0.975	1.000	1.000	1.000	0.986	0.896
	SIC	1.000	1.000	1.000	0.479	0.128	1.000	1.000	1.000	0.999	0.925	1.000	1.000	0.993	0.815	0.575
	GIC1	1.000	1.000	1.000	0.987	0.840	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.996	0.992
	GIC2	1.000	1.000	1.000	0.420	0.103	1.000	1.000	1.000	0.999	0.912	1.000	1.000	0.992	0.781	0.412
	CHI	1.000	1.000	0.586	0.105	0.071	1.000	1.000	0.986	0.602	0.312	0.963	0.743	0.447	0.222	0.134
	KS	1.000	1.000	0.211	0.003	0.001	1.000	1.000	0.510	0.071	0.027	0.705	0.281	0.105	0.025	0.006

Note: The LR, CHI and KS tests have a nominal level of significance of 0.05

TABLE 6.2

Probabilities of rejecting the null hypothesis that $\gamma = 0$ for the gamma distribution using seven tests and discrimination criteria over 1000 replications of random samples of sizes $n = 100$ and $n = 1000$ ($\beta = 1$)

		$n = 100$			$n = 1000$		
		Shape Parameter α			Shape Parameter α		
True Distribution	Criteria	2.0	4.0	6.0	2.0	4.0	6.0
Two Parameter ($\gamma = 0$)	LR	0.088	0.066	0.051	0.042	0.043	0.048
	AIC	0.199	0.178	0.168	0.143	0.152	0.143
	SIC	0.056	0.039	0.041	0.005	0.013	0.014
	GIC1	0.648	0.628	0.609	0.546	0.576	0.575
	GIC2	0.026	0.018	0.014	0.005	0.016	0.014
	CHI	0.086	0.083	0.091	0.055	0.049	0.065
	KS	0.002	0.112	0.087	0.019	0.000	0.000
Three Parameter ($\gamma = 1$)	LR	0.930	0.256	0.114	1.000	0.987	0.569
	AIC	0.979	0.473	0.276	1.000	0.996	0.789
	SIC	0.898	0.189	0.079	1.000	0.931	0.279
	GIC1	0.999	0.844	0.695	1.000	1.000	0.963
	GIC2	0.767	0.090	0.040	1.000	0.942	0.299
	CHI	0.286	0.112	0.087	0.965	0.227	0.193
	KS	0.019	0.000	0.000	0.896	0.040	0.103

Note: The LR, CHI and KS tests have a nominal level of significance of 0.05

TABLE 6.3

Iaxiimized log-likelihood values, estimated parameter values and probabilities of rejecting the ull hypothesis that $\gamma = 0$ for the 3- and 2-parameter lognormal distributions fitted to n daily β-scattering samples over different years and sample sizes ($\beta = 1$)

Site	Year	n	Probabilities of rejecting $\gamma = 0$	Max(log L) Lognormal (3) Lognormal (2)	Shape α	Scale β	Location γ
11	1975	156	0.9993	-458.89	0.73	1.84	1.84
				-464.68	0.56	2.14	0.00
	1976	311	0.9860	-855.78	0.57	1.89	1.25
				-858.80	0.47	2.09	0.00
	1977	251	0.0000	-644.62	0.43	1.99	0.01
				-644.62	0.43	2.00	0.00
	1978	215	1.0000	-644.42	0.87	1.71	2.10
				-659.12	0.63	2.11	0.00
	1979	257	0.9999	-655.32	0.57	1.70	1.99
				-663.19	0.42	2.04	0.00
	1980	199	1.0000	-478.81	0.68	1.38	1.79
				-487.36	0.46	1.80	0.00
	1981	277	1.0000	-587.44	0.81	0.92	1.57
				-602.63	0.49	1.48	0.00
	1982	272	1.0000	-563.11	0.91	0.75	1.77
				-586.93	0.49	1.45	0.00
	1983	324	0.9999	-735.19	0.61	1.35	1.25
				-742.98	0.45	1.66	0.00
27	1979	291	1.0000	-812.74	0.61	1.86	1.69
				-821.60	0.49	2.13	0.00
	1980	304	1.0000	-686.81	0.69	1.21	3.30
				-722.44	0.37	1.95	0.00
	1981	302	1.0000	-755.21	0.72	1.41	2.74
				-790.78	0.46	1.98	0.00
	1982	279	1.0000	-686.98	0.72	1.37	1.94
				-708.51	0.49	1.82	0.00
	1983	326	0.9948	-824.74	0.55	1.70	1.18
				-828.64	0.45	1.92	0.00
34	1981	272	1.0000	-589.83	0.68	1.14	1.44
				-600.24	0.46	1.56	0.00
	1982	298	1.0000	-694.29	0.68	1.29	1.43
				-703.41	0.48	1.67	0.00
	1983	280	0.9991	-628.90	0.59	1.35	1.23
				-634.44	0.44	1.66	0.00
81	1981	160	1.0000	-319.99	0.76	0.85	2.89
				-343.30	0.37	1.72	0.00
	1982	312	1.0000	-985.45	0.78	1.99	3.05
				-1002.13	0.54	2.40	0.00
	1983	301	1.0000	-827.81	0.73	1.64	2.54
				-842.02	0.49	2.10	0.00

Note: Sites 11, 27, 34 and 81 are Museum, Alphington, Dandenong and Camberwell, respectively.

Table 6.4

Maxiimized log-likelihood values, estimated parameter values and probabilities of rejecting the nulll hypothesis that $\gamma = 0$ for the 3- and 2-parameter gamma distributions fitted to n daily nitrogen dioxide samples over different years and sample sizes ($\beta = 1$)

Sitee	Year	n	Probabilities of rejecting $\gamma = 0$	Max(log L) Lognormal (3) Lognormal (2)	Shape α	Scale β	Location γ
11	1975	116	0.5457	-253.42	2.68	1.50	0.25
				-253.70	3.14	1.37	0.00
	1976	280	0.1936	-687.36	2.06	2.36	0.03
				-687.39	2.10	2.33	0.00
	1977	297	0.5835	-678.94	3.44	1.43	0.13
				-679.27	3.20	1.49	0.00
	1978	227	0.9782	-613.74	1.32	4.26	0.25
				-616.37	1.57	3.74	0.00
	1979	271	0.9996	-586.23	1.76	1.98	0.53
				-592.42	2.60	1.54	0.00
	1980	175	0.7666	-365.56	2.30	1.52	0.18
				-366.27	2.65	1.39	0.00
	1981	276	0.9600	-518.99	2.05	1.33	0.14
				-521.10	2.35	1.22	0.00
	1982	292	0.9698	-773.78	1.59	3.47	0.22
				-776.13	1.80	3.18	0.00
	1983	313	0.5024	-741.46	2.67	1.81	0.12
				-741.69	2.85	1.74	0.00
27	1979	317	1.0000	-525.33	0.64	3.29	0.04
				-556.82	0.95	2.24	0.00
	1980	302	0.9787	-477.91	1.19	1.52	0.03
				-480.56	1.29	1.43	0.00
	1981	245	0.9632	-355.64	1.22	1.31	0.03
				-357.82	1.34	1.22	0.00
	1982	188	1.0000	-380.20	0.61	5.04	0.17
				-408.71	0.92	3.53	0.00
	1983	241	1.0000	-470.45	0.84	3.13	0.09
				-480.57	1.10	2.46	0.00
34	1981	193	1.0000	-244.10	0.67	2.07	0.04
				-260.46	1.15	1.24	0.00
	1982	256	0.9636	-426.33	0.96	2.04	0.04
				-428.52	1.23	1.62	0.00
	1981	139	0.1585	-219.45	1.52	1.23	0.01
				-219.47	1.55	1.21	0.00
	1982	251	0.9995	-536.92	1.35	2.37	0.25
				-543.01	1.74	1.99	0.00
	1983	230	0.8945	-477.96	1.17	2.53	0.03
				-479.27	1.24	2.42	0.00

Note: Sites 11, 27 and 34 are Museum, Alphington and Dandenong, respectively.

Chapter 7

The Effects of Misspecification in Estimating the Percentiles of Some Two- and Three-Parameter Distributions

7.1 Introduction

The gamma, Weibull and lognormal distributions have been used successfully in testing and modelling natural phenomena such as reliability and life testing (Mann et al. (1974) and Bain (1978)), hydrology (Stedinger (1980)), and air quality management (Jakeman and Taylor (1989) and Jakeman et al. (1986)). Two- and three-parameter versions of these distributions have been used because they are parsimonious in considering the shape, scale and location of the distribution, but still sufficiently flexible in fitting real data. Occasionally, there may be some prior information regarding the location of the distribution and parsimony considerations might yield a preference for the two-parameter variant. However, an estimated two-parameter distribution might be inadequate if the location parameter is not sufficiently close to zero. In general, it is not known which of the two- or three-parameter distributions is appropriate, and conventional wisdom regarding underfitting or overfitting may not be a good guide to selecting one of these distributions.

In considering whether the two- or three-parameter variant of a distribution should

be used to estimate the percentiles of the distribution, it is important to take account of the sorts of errors that might be made in fitting the distribution. Specifically, the consequences of misspecifying the distribution should be evaluated. Such misspecifications arise when a two- (three-) parameter distribution is estimated when the three- (two-) parameter version is correct. The purpose of this chapter is to assess the consequences of such misspecification in estimating the upper percentiles of the two- and three-parameter gamma, Weibull and lognormal distributions.

For a given distribution, the statistical decision might be to simply estimate the two- or three-parameter variant of the distribution, or to use discrimination and/or testing criteria to choose one of the two distributions. The three discrimination methods considered in this chapter are the likelihood ratio (LR) test, Akaike's information criterion (AIC), and Schwarz's information criterion (SIC) based on Bayesian methods. The primary aim of the experiments is to observe the magnitudes of the errors obtained by fitting the incorrect distribution (by overfitting or underfitting), by fitting the correct distribution, and by fitting a distribution that is selected by the LR method, AIC or SIC. The experiments are conducted for three different distributions and different parameter sets, especially for different values of the shape parameter.

7.2 Distribution Functions and Statistical Criteria

For a sample $x_1, x_2, \ldots, x_n$ of n independently and identically distributed random observations, the log- likelihood functions for the three-parameter gamma, Weibull and lognormal distributions are given as follows:

Gamma:

$$\log L = -n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log(x_i - \gamma) - \sum_{i=1}^{n} (\frac{x_i - \gamma}{\beta}) \qquad (7.1)$$

Weibull:

$$\log L = n \log \alpha - n\alpha \log \beta + (\alpha - 1) \sum_{i=1}^{n} \log(x_i - \gamma) - \sum_{i=1}^{n} (\frac{x_i - \gamma}{\beta})^{\alpha} \qquad (7.2)$$

Lognormal:

$$\log L = -\frac{n}{2}\log(2\pi\alpha^2) + \sum_{i=1}^{n}\log(x_i - \gamma) - \frac{1}{2\alpha^2}\sum_{i=1}^{n}[\log(x_i - \gamma) - \beta]^2 \qquad (7.3)$$

in which β represents the scale parameter, α the shape parameter, γ the location parameter, and Γ the gamma function. The two-parameter versions of the above functions are obtained by setting the location parameter γ to zero in each case. In the above equations, $\beta > 0$, $\alpha > 0$ and $\gamma < x_i < \infty$ for $i = 1, 2, \ldots, n$. The density functions of the gamma and Weibull functions approach the exponential at $\alpha = 1$, are "J" shaped for $\alpha < 1$ and "bell" shaped for $\alpha > 1$, whereas the density for the lognormal function changes from being nearly symmetric to being heavily skewed as α is increased from 0.4 to 0.9 to 1.2. These values accommodate a variety of shapes which arise in practice in analysing real data.

The parameters of the three log-likelihood functions are estimated by maximum likelihood methods. Since the general maximum likelihood procedure will frequently fail to converge when the shape parameter is less than or equal to unity, an approach that circumvents this problem is used (for further details, see Bai et al. (1989)). Denoting the maximized values of the two- and three-parameter variants of a particular log-likelihood as $logL_0$ and $logL_1$, respectively, the LR test can be expressed as:

$$LR = -2(logL_0 - logL_1) \overset{a}{\sim} \chi^2(1) \qquad (7.4)$$

when the null hypothesis that the location parameter is zero is true. The AIC and SIC may be expressed, respectively, as:

Choose the $\left\{ \begin{array}{c} 2 \\ 3 \end{array} \right\}$ parameter distribution if

$$\text{AIC: } logL_0 - 2\left\{ \begin{array}{c} > \\ < \end{array} \right\} logL_1 - 3$$

$$\text{SIC: } logL_0 - logn\left\{ \begin{array}{c} > \\ < \end{array} \right\} logL_1 - 3logn/2.$$

For purposes of assessing the performance of the LR, AIC and SIC, loss functions recommended for assessing air quality models are used (see Fox (1981)). These functions are the relative bias (BIAS) and the relative root mean square error (RRMSE) which are evaluated at the upper percentiles of the distribution. For an estimate $\hat{q}_i$ of a quantity of interest q, the performance criteria are defined in terms of deviations from q in each replication of the simulation experiments. The definitions used are as follows:

$$BIAS(q) = \frac{1}{N}\sum_{i=1}^{N}(\frac{\hat{q}_i - q}{q}) \tag{7.5}$$

$$RRMSE(q) = [\frac{1}{N}\sum_{i=1}^{N}(\frac{\hat{q}_i - q}{q})^2]^{0.5} \tag{7.6}$$

where N is the number of replications of the experiment. For present purposes, q denotes the upper percentile quantities of the underlying distributions.

7.3 Monte Carlo Experiments

In order to assess the effects of misspecification in estimating the percentiles of the three distributions, an extensive range of possible cases is considered. The shape parameter is examined over a wide range of possible values where the density functions are positively skewed: the shape parameter takes the values 0.5, 1, 2 and 6 for the gamma distribution; 0.5, 1, 2 and 3 for the Weibull distribution; and 0.4, 0.5, 0.9 and 1.2 for the lognormal distribution. In all cases considered in this chapter, the arbitrary scale parameter is set to unity, and the location parameter takes on the values 0 and 1 for the two- and three- parameter distributions, respectively. The lognormal distribution has the opposite behaviour to the gamma and Weibull distributions as the shape parameter is increased. For each entry in the tables, N=1000 replications of the experiments are processed. The sample size used is n=365, since it represents a common case, namely a full year of 24-hourly average observations. The level of significance used for the LR test is five percent. For each set of parameter values, BIAS and RRMSE are evaluated

for estimates of the exact 98'th percentile and the highest value which, for n=365, is equivalent to using the 99.9'th percentile.

The random sample generators used for the Monte Carlo experiments are DRNGAM, DRNWIB and DRNLNL for the gamma, Weibull and lognormal distributions, respectively. These are available as subroutines in the International Mathematical and Statistical Library (IMSL) in version 1.0 (April 1987). The same seed number (1234) is used to obtain the first random sample of the first of the 1000 replications. Varying the initial seed produces similar results to those reported in the chapter.

7.4 Monte Carlo Results

Results of the experiments for the two- and three-parameter gamma, Weibull and log-normal distributions are given in Tables 7.1-7.3, respectively. In all cases, the true model is either a two- or three-parameter distribution, the estimated quantities are the maximum value or the 98% value, and the performance criteria are BIAS and RRMSE. The LR test and the two discrimination criteria, AIC and SIC, are used. Both the three- and two-parameter distributions are estimated to examine the consequences of misspecifying the distribution, namely estimating the two- (three-) parameter distribution when the three- (two-) parameter variant is correct.

The following points should be noted from the experiments reported in Table 7.1.

(i) When the shape parameter is 0.5 and the three-parameter gamma distribution is correct ($\gamma = 1$), the BIAS and RRMSE are identical for the three discrimination criteria and the estimated three-parameter gamma distribution, for both the maximum and 98% quantities. Underfitting the correct distribution with a two-parameter version yields substantially larger values for BIAS and RRMSE. Identical qualitative results are obtained as the shape parameter is increased from 0.5 to 1, and then to 2. When the shape parameter is increased to 6 with $\gamma = 1$, the rankings in terms of RRMSE for both the maximum and 98% values are very similar for the three discrimination criteria and the two estimated distributions, with a slight preference for the

correctly estimated three-parameter gamma distribution. In terms of BIAS, however, the estimated three-parameter distribution is superior to the rest, followed by AIC which is known to favour the model with the larger number of parameters (namely, the three-parameter model). Interestingly, the BIAS in estimating the underfitted two-parameter gamma distribution is not substantially higher than for SIC, which favours the more parsimonious two-parameter model. On the basis of these results, it is clear that incorrectly underfitting a distribution can lead to substantially higher BIAS and RRMSE values, in general, as compared with fitting the correct distribution, and also relative to using discrimination criteria to select the appropriate distribution.

(ii) On the other hand, when the two-parameter gamma distribution is correct ($\gamma = 0$) and the shape parameter is 0.5, the estimated two-parameter distribution is understandably superior in terms of BIAS for both the maximum and 98% values, with the three discrimination criteria and the estimated three-parameter gamma distribution being very similar. In terms of RRMSE, however, the three discrimination criteria and the two estimated distributions are very similar, with only a very slight preference for the correctly estimated two-parameter gamma distribution, followed by SIC which favours parsimony. Similar qualitative results hold for both BIAS and RRMSE as the shape parameter is increased from 0.5 to 1. However, when the shape parameter is increased from 1 to 2, and from 2 to 6, the SIC and LR criteria have smaller BIAS values than the estimated two-parameter gamma distribution for both the maximum and 98% quantities. Using RRMSE, the three discrimination criteria and the two estimated distributions are quite similar. On the basis of these results, it is clear that BIAS values will not generally be increased substantially by incorrectly overfitting a three-parameter gamma distribution when the two-parameter version is correct, and that at least one of the discrimination criteria can be superior to the correctly estimated two-parameter distribution in terms of BIAS and RRMSE.

The main points from the Weibull distribution experiments given in Table 7.2 are as follows:

(i) When the three-parameter Weibull distribution is correct, the three discrimination criteria and the estimated three-parameter Weibull distribution are identical in terms of BIAS and RRMSE for both the maximum and 98% quantities for all values of the shape parameter (namely, 0.5, 1, 2 and 3). Underfitting the correct Weibull distribution with its two-parameter variant yields considerably larger BIAS and RRMSE for all values of the shape parameter, although the degree of error is reduced as the value of the shape parameter is increased. Thus, estimating the correct three-parameter distribution is not preferable in terms of BIAS and RRMSE as compared with using the three discrimination criteria, although underfitting the three-parameter Weibull distribution by setting the location parameter γ to zero yields much larger BIAS and RRMSE values in all cases considered. This result is broadly similar in qualitative terms to those obtained for the gamma distribution although, in the latter case, estimating the correct distribution is preferred in terms of BIAS for a large value of the shape parameter.

(ii) As compared with the case of the gamma distribution, overfitting the correct two-parameter Weibull distribution yields some surprising results in terms of BIAS, especially for larger values of the shape parameter. When the shape parameter is 0.5, the estimated two-parameter Weibull distribution has much lower BIAS values than those obtained using the discrimination criteria and the estimated three-parameter Weibull distribution for both the maximum and 98% quantities. As the shape parameter is increased from 0.5 to 1, the correctly estimated two-parameter distribution is still preferred for the maximum quantity but is inferior to SIC, which favours the lower dimensioned model, for the 98% quantity. When the shape parameter is increased to 2 or 3, all three discrimination criteria are preferred to the estimated two-parameter Weibull distribution in terms of BIAS. Indeed, when the shape parameter is set to 3, even the overfitted three-parameter distribution has lower BIAS than the correctly fitted two-parameter variant for both the maximum and 98% quantities. In terms of RRMSE, the correctly estimated two-parameter distribution is preferred in all cases, although its superiority is diminished as the shape parameter is increased. While there

is not a noticeable difference between the overfitted three-parameter distribution and the three discrimination criteria, SIC is always second best to the estimated two- parameter distribution, and the estimated three-parameter distribution is generally the worst. Qualitatively, the results are reasonably similar to those for the gamma distribution in that overfitting does not generally increase the BIAS and RRMSE values for the maximum and 98% quantities for the Weibull distribution. However, it is interesting to note that estimating the correct two-parameter Weibull distribution does not always yield the smallest BIAS value relative to the three discrimination criteria, or even to the overfitted three-parameter distribution.

Finally, Table 7.3 contains the results from experiments for the lognormal distribution. Since the lognormal distribution has the opposite behaviour to the gamma and Weibull distributions as the shape parameter is increased, it is useful for comparative purposes to examine the results as the shape parameter is decreased rather than increased. The principal points to note from the table are as follows:

(i) When the three-parameter lognormal distribution is correct ($\gamma = 1$) and the shape parameter is 1.2 or 0.9, the correctly fitted three-parameter distribution and the three discrimination criteria have identical BIAS and RRMSE values for both the maximum and 98% quantities, whereas the underfitted two-parameter lognormal distribution has substantially higher values for BIAS and RRMSE. Although the underfitted two-parameter distribution still has the largest BIAS and RRMSE values when the shape parameter is reduced to 0.5 or 0.4, the other four methods do not remain identical. When the shape parameter is 0.5, the LR method has by far the smallest BIAS for the maximum quantity while SIC has the largest; for the 98% quantity, AIC and the correctly fitted three-parameter distribution have the smallest BIAS, and SIC again has the largest. These rankings are not maintained when the shape parameter is reduced to 0.4. The smallest BIAS values for the maximum and 98% quantities are AIC and the estimated three-parameter lognormal distribution, respectively, with SIC the worst of the four methods in each case. On the basis of RRMSE, however, the three-parameter lognormal distribution has the smallest value, followed closely by

AIC, with SIC the worst of the four methods, for both the maximum and 98% quantities when the shape parameter is 0.5 or 0.4. The poor performance of SIC, which favours the more parsimonious two-parameter lognormal variant of the correct model, is consistent with the findings for the gamma and Weibull distributions, as is the result that underfitting the model will generally lead to much larger values of BIAS and RRMSE for both the maximum and 98% quantities. Moreover, estimating the correct distribution is the preferred strategy, at least in terms of RRMSE, and sometimes also for BIAS, relative to using the three discrimination criteria to determine which of the three- and two- parameter distributions should be used.

(ii) Similar observed patterns to the above do not hold when the correct model is the two-parameter lognormal distribution ($\gamma = 0$). For example, when the shape parameter is 1.2, the correctly fitted two-parameter lognormal distribution has the smallest BIAS, followed by the parsimony- inclined SIC, and lastly by the overfitted three-parameter lognormal distribution, for the maximum quantity; for the 98% quantity, however, SIC has BIAS equal to that of the estimated two-parameter distribution, followed by the LR method and lastly by the overfitted three-parameter lognormal distribution. Even when the shape parameter is reduced to 0.9, these rankings are not sustained. While the estimated three-parameter lognormal distribution has the largest BIAS for both the maximum and 98% quantities, in the former case SIC is best, followed by the estimated two-parameter lognormal distribution, and in the latter case the LR method is best, followed by SIC. Interesting results arise when the shape parameter is reduced to 0.5 or 0.4. In the former case, the LR method has lowest BIAS, followed by SIC and lastly by the overfitted three-parameter distribution for the maximum quantity; for the 98% quantity, AIC is best, followed by the overfitted three-parameter distribution and lastly by the correctly fitted two-parameter distribution. The results for BIAS are especially striking when the shape parameter is reduced to 0.4. The LR method has lowest BIAS, followed by SIC, for the maximum quantity, while AIC is best, followed by the LR method, for the 98% quantity. However, the correctly estimated two-parameter distribution has the largest BIAS in each case, even larger than those

of the overfitted three-parameter variant. Thus, overfitting the lognormal distribution for low values of the shape parameter can yield lower BIAS values than estimating the correct distribution. The results for RRMSE are not too dissimilar from the qualitative results for the gamma and Weibull distributions. For shape parameter values of 1.2, 0.9 or 0.5, and for both the maximum and 98% quantities, the correctly estimated two-parameter lognormal distribution always has the smallest RRMSE, followed in each case by SIC, which favours parsimony, and lastly by the overfitted three-parameter lognormal distribution. When the shape parameter is reduced to 0.4, however, the rankings for RRMSE for both the maximum and 98% quantities are SIC, LR and AIC, followed distantly by the two estimated distributions. Thus, in terms of RRMSE, the three discrimination criteria are preferred to simple estimation of the distribution, whether it be the correct distribution or an overfitted variant.

7.5 Concluding Remarks

In this chapter we have assessed the effects of missspecification in estimating the percentiles of the two- and three-parameter gamma, Weibull and lognormal distributions. In the experiments, the true model is either a two- or three-parameter distribution, the estimated quantities are the maximum observed value or the ninety-eighth percentile value, and the performance criteria are the BIAS and RRMSE associated with the estimated quantities. Three discrimination criteria are used, together with estimation of both the three- and two-parameter distributions, to evaluate the consequences of misspecifying the distribution. The shape parameter is examined over a wide range of possible values. The results for the gamma and Weibull distributions are generally similar in qualitative terms. In particular, incorrectly underfitting a distribution can lead to substantially higher BIAS and RRMSE values as compared with fitting the correct distribution, and also relative to using discrimination criteria to select the appropriate distribution. Incorrectly overfitting a distribution does not generally increase BIAS substantially, if at all, and the discrimination criteria can be superior to the cor-

rectly estimated two-parameter distribution. Results for the lognormal distribution are not entirely similar to the above, depending crucially on the value of the shape parameter. Incorrectly underfitting the three-parameter lognormal distribution yields significantly larger BIAS and RRMSE values, and correctly fitting the distribution seems to be optimal. However, for small values of the shape parameter, the correctly estimated two-parameter distribution can have larger BIAS values than for the incorrectly overfitted three-parameter lognormal distribution, and much larger values than those obtained using discrimination criteria to select the appropriate distribution.

TABLE 7.1

Estimates of BIAS and RRMSE at maximum and 98 % quantities from 1000 Monte Carlo simulations with sample size n = 365 for the gamma distribution ($\beta = 1$)

GAMMA DISTRIBUTION							
True Values	Percentile	Performance Criteria	Discrimination Criteria LR	AIC	SIC	Estimated Models 3-P	2-P
Gamma3	MAX	BIAS	0.0013	0.0013	0.0013	0.0013	-.3803
α=0.5		RRMSE	0.0673	0.0673	0.0673	0.0673	0.3820
γ=1.0	98	BIAS	0.0007	0.0007	0.0007	0.0007	-.2198
		RRMSE	0.0550	0.0550	0.0550	0.0550	0.2231
Gamma2	MAX	BIAS	0.0183	0.0164	0.0168	0.0183	-.0003
α=0.5		RRMSE	0.0836	0.0846	0.0844	0.0836	0.0808
γ=0.0	98	BIAS	0.0132	0.0118	0.0120	0.0132	-.0003
		RRMSE	0.0772	0.0782	0.0780	0.0772	0.0755
Gamma3	MAX	BIAS	0.0234	0.0234	0.0234	0.0234	-.2638
α=1.0		RRMSE	0.0600	0.0600	0.0600	0.0600	0.2665
γ=1.0	98	BIAS	0.0156	0.0156	0.0156	0.0156	-.1559
		RRMSE	0.0492	0.0492	0.0492	0.0492	0.1603
Gamma2	MAX	BIAS	0.0157	0.0282	0.0070	0.0429	-.0016
α=1.0		RRMSE	0.0729	0.0775	0.0687	0.0777	0.0620
γ=0.0	98	BIAS	0.0113	0.0205	0.0050	0.0314	-.0012
		RRMSE	0.0638	0.0667	0.0614	0.0671	0.0575
Gamma3	MAX	BIAS	0.0073	0.0073	0.0073	0.0073	-.1450
α=2.0		RRMSE	0.0527	0.0527	0.0527	0.0527	0.1499
γ=1.0	98	BIAS	0.0040	0.0040	0.0040	0.0040	-.0850
		RRMSE	0.0422	0.0422	0.0422	0.0422	0.0919
Gamma2	MAX	BIAS	0.0014	0.0050	-.0008	0.0106	-.0018
α=2.0		RRMSE	0.0552	0.0572	0.0546	0.0597	0.0532
γ=0.0	98	BIAS	0.0002	0.0026	-.0012	0.0062	-.0018
		RRMSE	0.0478	0.0487	0.0476	0.0500	0.0468
Gamma3	MAX	BIAS	-.0166	-.0062	-.0235	0.0042	-.0339
α=6.0		RRMSE	0.0492	0.0484	0.0488	0.0440	0.0461
γ=1.0	98	BIAS	-.0094	-.0038	-.0131	0.0019	-.0186
		RRMSE	0.0336	0.0335	0.0333	0.0315	0.0321
Gamma2	MAX	BIAS	0.0005	0.0027	-.0006	0.0045	-.0014
α=6.0		RRMSE	0.0383	0.0423	0.0367	0.0467	0.0358
γ=0.0	98	BIAS	0.0000	0.0011	-.0006	0.0021	-.0011
		RRMSE	0.0303	0.0321	0.0298	0.0341	0.0296

Note : α is the shape parameter, β the scale parameter and γ the location parameter. Gamma3 and Gamma2 denote the 3- and 2-parameter gamma distributions.

TABLE 7.2

Estimates of BIAS and RRMSE at maximum and 98 % quantities from 1000 Monte Carlo simulations with sample size n = 365 for the Weibull distribution ($\beta = 1$)

WEIBULL DISTRIBUTION							
True Values	Percentile	Performance Criteria	Discrimination Criteria LR	AIC	SIC	Estimated Models 3-P	2-P
Weibull3	MAX	BIAS	0.0041	0.0041	0.0041	0.0041	-.5739
α=0.5		RRMSE	0.1466	0.1466	0.1466	0.1466	0.5773
γ=1.0	98	BIAS	0.0021	0.0021	0.0021	0.0021	-.3002
		RRMSE	0.1165	0.1165	0.1165	0.1165	0.3116
Weibull2	MAX	BIAS	0.0382	0.0382	0.0382	0.0382	0.0000
α=0.5		RRMSE	0.1599	0.1599	0.1599	0.1599	0.1495
γ=0.0	98	BIAS	0.0239	0.0239	0.0239	0.0239	-.0005
		RRMSE	0.1290	0.1290	0.1290	0.1290	0.1238
Weibull3	MAX	BIAS	0.0241	0.0241	0.0241	0.0241	-.2768
α=1.0		RRMSE	0.0714	0.0714	0.0714	0.0714	0.2803
γ=1.0	98	BIAS	0.0135	0.0135	0.0135	0.0135	-.1250
		RRMSE	0.0521	0.0521	0.0521	0.0521	0.1323
Weibull2	MAX	BIAS	0.0111	0.0249	0.0035	0.0446	-.0028
α=1.0		RRMSE	0.0850	0.0912	0.0806	0.0904	0.0743
γ=0.0	98	BIAS	0.0063	0.0150	0.0017	0.0276	-.0021
		RRMSE	0.0667	0.0699	0.0647	0.0696	0.0617
Weibull3	MAX	BIAS	0.0017	0.0017	0.0017	0.0017	-.0955
α=2.0		RRMSE	0.0302	0.0302	0.0302	0.0302	0.0980
γ=1.0	98	BIAS	0.0005	0.0005	0.0005	0.0005	-.0419
		RRMSE	0.0217	0.0217	0.0217	0.0217	0.0461
Weibull2	MAX	BIAS	-.0008	0.0005	-.0016	0.0055	-.0021
α=2.0		RRMSE	0.0388	0.0399	0.0377	0.0426	0.0372
γ=0.0	98	BIAS	-.0009	-.0002	-.0013	0.0024	-.0015
		RRMSE	0.0316	0.0319	0.0311	0.0329	0.0309
Weibull3	MAX	BIAS	0.0008	0.0008	0.0008	0.0008	-.0466
α=3.0		RRMSE	0.0195	0.0195	0.0195	0.0195	0.0486
γ=1.0	98	BIAS	0.0001	0.0001	0.0001	0.0001	-.0206
		RRMSE	0.0137	0.0137	0.0137	0.0137	0.0238
Weibull2	MAX	BIAS	-.0011	-.0007	-.0014	0.0013	-.0015
α=3.0		RRMSE	0.0262	0.0276	0.0251	0.0298	0.0248
γ=0.0	98	BIAS	-.0009	-.0008	-.0011	0.0001	-.0011
		RRMSE	0.0211	0.0216	0.0207	0.0223	0.0206

Note : α is the shape parameter, β the scale parameter and γ the location parameter. Weibull3 and Weibull2 denote the 3- and 2-parameter Weibull distributions.

TABLE 7.3

Estimates of BIAS and RRMSE at maximum and 98 % quantities from 1000 Monte Carlo simulations with sample size n = 365 for the lognormal distribution ($\beta = 1$)

LOGNORMAL DISTRIBUTION							
True Values	Percentile	Performance Criteria	Discrimination Criteria LR	AIC	SIC	Estimated Models 3-P	2-P
LogN3	MAX	BIAS	0.0283	0.0283	0.0283	0.0283	-.4706
α=1.2		RRMSE	0.1742	0.1742	0.1742	0.1742	0.4758
γ=1.0	98	BIAS	0.0130	0.0130	0.0130	0.0130	-.2748
		RRMSE	0.1195	0.1195	0.1195	0.1195	0.2844
LogN2	MAX	BIAS	0.0110	0.0181	0.0050	0.0286	0.0016
α=1.2		RRMSE	0.1582	0.1645	0.1527	0.1760	0.1485
γ=0.0	98	BIAS	0.0041	0.0078	0.0009	0.0134	-.0009
		RRMSE	0.1154	0.1179	0.1133	0.1232	0.1116
LogN3	MAX	BIAS	0.0169	0.0169	0.0169	0.0169	-.3387
α=0.9		RRMSE	0.1344	0.1344	0.1344	0.1344	0.3444
γ=1.0	98	BIAS	0.0066	0.0066	0.0066	0.0066	-.1823
		RRMSE	0.0891	0.0891	0.0891	0.0891	0.1918
LogN2	MAX	BIAS	0.0042	0.0091	0.0005	0.0173	-.0009
α=0.9		RRMSE	0.1196	0.1262	0.1125	0.1377	0.1108
γ=0.0	98	BIAS	0.0006	0.0030	-.0012	0.0070	-.0018
		RRMSE	0.0869	0.0892	0.0840	0.0942	0.0835
LogN3	MAX	BIAS	0.0001	0.0047	-.0107	0.0052	-.1453
α=0.5		RRMSE	0.0865	0.0794	0.0988	0.0784	0.1514
γ=1.0	98	BIAS	-.0015	0.0008	-.0067	0.0010	-.0689
		RRMSE	0.0515	0.0485	0.0565	0.0481	0.0775
LogN2	MAX	BIAS	-.0004	0.0036	-.0014	0.0056	-.0020
α=0.5		RRMSE	0.0668	0.0752	0.0641	0.0849	0.0613
γ=0.0	98	BIAS	-.0013	0.0005	-.0017	0.0012	-.0019
		RRMSE	0.0479	0.0508	0.0472	0.0545	0.0464
LogN3	MAX	BIAS	-.0102	-.0007	-.0255	0.0032	-.1006
α=0.4		RRMSE	0.0782	0.0685	0.0885	0.0633	0.1066
γ=1.0	98	BIAS	-.0060	-.0016	-.0130	0.0003	-.0463
		RRMSE	0.0438	0.0400	0.0478	0.0380	0.0543
LogN2	MAX	BIAS	-.0006	0.0024	-.0015	-.0967	-.1016
α=0.4		RRMSE	0.0546	0.0617	0.0520	0.1156	0.1108
γ=0.0	98	BIAS	-.0012	0.0000	-.0016	-.1390	-.1407
		RRMSE	0.0386	0.0410	0.0379	0.1441	0.1443

Note : α is the shape parameter, β the scale parameter and γ the location parameter. LogN3 and LogN2 denote the 3- and 2-parameter lognormal distributions.

Chapter 8

Discrimination Procedures for Fitting Nested and Non-Nested Distributions

8.1 Introduction

In recent years there has been an increased demand for statistical techniques to draw inferences regarding the probability distribution of an unknown parent population based on sample information. Applications occur in many areas, such as reliability and life-testing (Mann et al. (1974) and Bain (1978)), hydrology (Stedinger (1980)), and air quality management (Jakeman and Taylor (1989)). Discrimination techniques are used when the null model has both a more general as well as non-nested alternatives.

For a particular sample of observations, it is generally preferable to have a greater choice among alternatives rather than less. It is found in this chapter that some members of non-nested families can be good approximations to others. This occurs especially among three-parameter probability distributions. When distributions are very similar, it can be difficult to discriminate among alternatives. It would be useful, therefore, to consider selection criteria that are powerful in discriminating among very similar distributions.

In this chapter, discrimination among nested and non-nested distributions is considered. Traditional approaches are re-examined. These include the standard likelihood

ratio procedure which tests the null hypothesis against a more general alternative, and some well-known discrimination criteria. The particular problem considered is to discriminate among a set of probability distributions in three situations. The first two cases are concerned only with non-nested distributions, but both nested and non-nested distributions are examined in the third case. For illustrative purposes, the two- and three-parameter gamma, Weibull and lognormal distributions are used to demonstrate a new discrimination procedure and compare its performance with existing criteria. These three distributions are well known and have been used successfully for various problems. Monte Carlo experiments are employed to examine their performance over a range of parameter values found to be relevant in assessing environmental quality data.

Extensive investigation of a number of popular criteria indicates that there is no ideal existing method of discrimination which is satisfactory for all cases of interest. The results show that the performance of each criterion depends strongly on the relevant probability distribution, the range of parameters and the confidence level (if applicable). Of course, performances are much improved when the sample size is increased, but sufficiently large sample sizes are not always possible in real applications.

To complement the inadequacies of existing criteria, a generalized information criterion (GIC) is proposed. A desirable feature of GIC is to determine whether or not a distribution is significantly superior to others under consideration. There is no restriction on the number of candidates in the discrimination procedure, as long as the true distribution is among the alternatives considered. The GIC procedure is particularly useful when the alternative distributions considered are very similar.

The plan of the chapter is as follows. In Section 2 several existing procedures for testing and discriminating among nested and non-nested distributions are discussed, as are several goodness-of-fit tests. Some practical problems in discrimination and testing are outlined in Section 3. Discrimination criteria and asymptotic tests are presented in Section 4, and a generalised information criterion is developed in Section

5. Three distribution functions and several statistical criteria are given in Section 6. The Monte Carlo procedure is outlined in Section 7, and the results of the experiments for discriminating between two, three and five non-nested distributions are discussed in Sections 8, 9 and 10, respectively. An extension of the Monte Carlo experiments to the case of one nested and five non-nested distributions is presented and discussed in Section 11. Some concluding remarks are given in Section 12.

8.2 Available Procedures

Let $x_1, x_2, \ldots, x_n$ represent n observations of a random variable with density function $f(x)$. If the null hypothesis specifies $H_0 : f(x) = f_0(x;\theta)$ against the alternative $H_1 : f(x) = f_1(x;\phi)$, where θ and ϕ are $p \times 1$ and $q \times 1$ parameter vectors ranging over suitable domains, the nesting requirement is that $f_0(x;\theta)$ and $f_1(x;\phi)$ are members of the same family of distributions. However, if an arbitrary member of one family cannot be obtained as a limit member of the other, $f_0(x;\theta)$ and $f_1(x;\phi)$ are separate (or non-nested) distributions.

Goodness-of-fit criteria used for assessing the appropriateness of a given distribution can be classified into three categories. Standard tests involve inference about either the unknown population distribution or parameters based on the available sample information. Taking into account the variability associated with samples from a particular distribution, such hypothesis tests allow confidence levels to be established. In the nested case, the typical procedure used is the likelihood ratio (LR) test with an asymptotic χ^2 distribution. In the non-nested case, a well-known test is the Cox (1961,1962) test, which is based upon the centred likelihood ratio statistic (see McAleer and Pesaran (1986)). Other tests include those of Atkinson (1970), Epps et al. (1982), and Horowitz and McAleer (1988).

In recent years, many information criteria have appeared in the literature to complement the hypothesis testing approach. The two most frequently used information criteria are those of Akaike (1974) and Schwarz (1978). The primary advantage of

information criteria is that they are very flexible. They can be used in both the nested and non-nested cases, and the true model does not have to be a candidate among the tested models. Information criteria can also be used to discriminate among a set of more than two candidates, unlike the hypothesis tests which are generally limited to two distributions. Moreover, the criteria are usually simple to use. The disadvantage of information criteria is that the results they yield generally do not convey any probabilistic statements, and it is difficult to analyse quantitatively their power and robustness, as can be done for hypothesis testing.

There are other familiar tests, different from the tests mentioned above, which emphasize testing the differences in fit between the proposed distribution and the sampling distribution. Such goodness-of-fit tests can be assessed by fitting either the cumulative distribution function or the probability density function, and the model will be tested for significant deviations from the sampling distribution. In general, there is no restriction to nested or non-nested cases for such tests and they allow consideration of a set of alternatives. Although such goodness-of-fit tests are also based on asymptotic theory, in finite samples their results might differ from the hypothesis tests discussed earlier. Two well-known goodness-of-fit tests considered in this chapter are the chi-square (Pearson (1900)) and the Kolmogorov-Smirnov (see Bury (1975)) tests.

8.3 Practical Problems

The purpose of standard approaches for testing and discrimination is to determine either the 'true' or 'best' distribution based upon a random sample of observations. In the first category, the Cox (1961, 1962) procedure is used to test the null hypothesis against a non-nested alternative. One of the two distributions is presumed to be true. A representative example in the second category is Akaike's information criterion (AIC), which chooses from among a set of models the distribution with the highest value of the log-likelihood function subject to a penalty for the number of parameters estimated.

If the null hypothesis is true, the sampling distribution will approach the population

distribution as the sample size tends to infinity, but false null hypotheses will be rejected based on the evidence of large departures between the hypothesized and sampling distributions. As the sample size increases without limit, the false distribution will be rejected with probability approaching one.

Information criteria are constructed in a different way. For example, Schwarz's information criterion (SIC) is based on Bayesian theory, while Akaike's information criterion can be considered from the theory of cross-validation (Stoica et al. (1986)). The aim of existing information criteria is to choose the distribution which 'best' fits the data, after imposing a penalty for the number of parameters and/or observations used in estimation. In contrast to hypothesis tests, the true distribution need not be included among the models considered. However, when the true distribution is under consideration, the 'best' distribution chosen should be the true one as the sample size tends to infinity. Under this assumption, the sampling errors tends to zero, and the underlying and sampling distribution become identical. In this sense, the performances of information criteria and hypothesis tests should be comparable. These discrimination and testing criteria can be used successfully over a wide range of situations. Simulation experiments have shown that good empirical sizes and powers can be obtained for many distributions, including many one- and two-parameter distributions.

There is no single dominant criterion, and some of these deficiencies will be discussed in the following sections. Attention is paid mainly to general problems and emphasis is placed on discrimination among parametric probability distributions.

Example 1: Consider real data from air pollutant measurements of 24-hour averages of nitrogen dioxide concentrations sampled at the city monitoring station in Canberra, Australia. The number of available daily observations for the 1979 full year is 271. To test the data, the three-parameter gamma distribution is chosen as the null hypothesis against the non-nested alternative three-parameter Weibull distribution. Fitting three-parameter gamma and Weibull distributions to the data, the maximized values of the log-likelihood functions are -375.99 and -375.33, respectively. Applying the bounded-

size likelihood ratio (BLR) test of Horowitz and McAleer (1988) at the 95 per cent confidence level, with a critical value of 1.35, the Weibull alternative does not fit significantly better than the gamma null. Thus, the gamma distribution cannot be rejected against the Weibull alternative.

Example 2: Suppose a random sample is generated from an underlying two- or three-parameter gamma distribution with sample size $n = 365$ (which represents one year of 24-hour observations), and this is repeated in 1000 Monte Carlo simulation experiments. The three-parameter gamma, Weibull and lognormal distributions are considered as candidates for fitting the data. For each experiment, the method of maximum likelihood is used to estimate the three parameters for each of these three distributions. The simplest discrimination procedure is to choose the distribution with the highest value of the log-likelihood function. This is equivalent to using AIC and SIC for selecting the distributions when the number of parameters of the alternative distributions is the same. The results of the simulations are as follows:

(1) When the parameters of the underlying gamma distribution are given as shape = 2, scale = 1 and location = 1 in all 1000 simulation experiments, the three-parameter gamma distribution is selected 596 times, the three-parameter Weibull distribution 355, and the three-parameter lognormal distribution 49. That is, only 59.6 per cent of the selections are correct and 40.4 per cent are incorrect.

(2) When the parameters of the underlying gamma distribution are given as shape = 6, scale = 1 and location = 1 in 1000 simulation experiments, the three-parameter gamma distribution is selected 526 times, the three-parameter Weibull distribution 234, and the three-parameter lognormal distribution 240. That is, only 52.6 per cent of selections are correct and 47.4 per cent are incorrect.

This example simply shows that, in finite samples, the incorrect distribution may frequently have a higher maximized value of the log-likelihood function than the true one. Using information criteria such as AIC may lead to a large number of incorrect selections. Indeed, many other existing selection criteria have similar problems. More

details will be given in the following sections.

For a particular sample of observations, the population from which the data are drawn is unknown. If the sample is finite, sampling error exists. When there is strong statistical evidence in favour of an hypothesized distribution, it might only be that the specific sampling distribution is close to the population. However, there may be alternative distributions which could provide better representations of the data.

One of the problems of standard hypothesis tests arises from the requirement that one of the models considered must be the true distribution, while the other must be false. As discussed already, a difference of performance between two distributions will be used to infer that one is true and the other false. The initial null hypothesis and the alternative used are chosen arbitrarily. There is no strong reason why the sample must be drawn from one of the tested distributions. If both of the distributions initially hypothesized are wrong, the test will not necessarily indicate such a result. Moreover, from simulation results, the true distribution does not always perform the best, even in large but finite samples.

Another problem of standard hypothesis tests is that they are generally restricted to two distributions. Initial consideration of the correct distribution could be over a wide range of possible cases, such as those which have already been used in practice. This is a practical rather than a statistical problem. Attempting to overcome the problem is a major concern of this chapter. Most information criteria such as AIC and SIC can be used without the limitation of the second problem. There is no requirement that the number of distributions considered be limited. The task of such criteria is to choose the 'best' fitting distribution among all candidates, with some penalty for parsimony and the number of observations used. For a particular sample of observations, the 'best' fit to the distribution can be determined. Generally, the true model is the best fitting distribution if the sample size is very large.

8.4 Discrimination Criteria and Asymptotic Tests

Traditional hypothesis testing involves testing the 'true' distribution. However, emphasis can also be placed on determining false distributions. Assuming that the true distribution exists among a set of distributions, strong evidence can be sought to indicate that any distribution is false. In the ideal case, all false distributions will be rejected and the true distribution accepted. In the worst case, with little significant differences between the distributions, no superior distribution will emerge. Performances are similar and false distributions cannot be recognized. Between these extreme cases, some false distributions are determined and some of them are not, so that empirical distributions can be allocated to two categories: the superior and the badly fitting categories. The true distribution should frequently appear in the superior distribution category and have a high probability of appearing as the superior distribution in each sample case. Such a procedure is used to dismiss the worst cases, and retain the superior distributions. Consequently, this procedure can avoid some problems encountered in the use of existing criteria.

The rationale used above leads to the construction of a new discrimination criterion. This criterion would perform the filtering function to dismiss the badly fitting distributions. It should also retain distributions with similar performances, which are superior to those rejected. The criterion is designed to test a number of distributions simultaneously and to apply in general situations. The key issue is how to distinguish between the superior and badly fitting distributions. To determine a critical value, the equivalence between some well-known information criteria and hypothesis tests can be used. Since most information criteria and hypothesis tests are based on the maximized value of the likelihood function, it is possible to compare information criteria and hypothesis tests in terms of the probability of accepting the underlying null distribution.

Consider comparisons between the well-known AIC, SIC, the LR test, and the BLR test of Horowitz and McAleer (1988). In discriminating between a 2-parameter distribution and a 3-parameter alternative for a random sample of size n, and defining

$logL_0$ and $logL_1$ as the maximized values of the log-likelihood functions of the 2- and 3-parameter distributions, respectively, the AIC and SIC may be expressed as:

Choose the 2-parameter distribution if

$$AIC : -logL_0 + 2 < -logL_1 + 3 \tag{8.1}$$

$$SIC : -logL_0 + logn < -logL_1 + 3logn/2. \tag{8.2}$$

Rearranging (1) and (2), and defining $\Delta L = logL_0 - logL_1$, the respective criteria will select the 2-parameter null distribution as follows :

$$AIC : -\Delta L < 1 \tag{8.3}$$

$$SIC : -\Delta L < \frac{logn}{2}. \tag{8.4}$$

Suppose it is desired to test the nested null hypothesis H_0: 2-parameter distribution, against the more general alternative hypothesis H_1: 3-parameter distribution. The LR test defines the rejection region of the null hypothesis as:

$$LR : -\Delta L > \frac{c}{2} \tag{8.5}$$

where c is the critical value of the χ^2 distribution with one degree of freedom.

In the non-nested case with a similar hypothesis H_0: 2-parameter distribution, against the non-nested alternative H_a: 3-parameter distribution, the BLR test defines the rejection region as :

$$BLR : -\Delta L > z^* \tag{8.6}$$

where $z^* > 0$ and the asymptotic upper bound on the significance level is given by the cumulative standard normal distribution function evaluated at $\Phi[-\sqrt{2z^*}]$.

By comparison with (3) and (4), the LR and BLR test statistics will accept H_0 if

$$LR : -\Delta L < \frac{c}{2} \tag{8.7}$$

$$BLR : -\Delta L < z^*. \tag{8.8}$$

The probability of accepting the true model could be obtained by using the critical values of the χ^2 and the standard normal distributions, respectively. An equivalence among AIC, SIC and the LR and BLR tests can easily be established. For any sample size, use of AIC is equivalent to use of the LR statistic at the 84.2 per cent confidence level for nested distributions, and to use of the BLR statistic at the 92.1 per cent confidence level for non-nested distributions. Similarly, for a sample size of 365, SIC is equivalent to the LR test at the 98.5 per cent confidence level in the nested case, and is equivalent to the BLR test at the 99.2 per cent confidence level in the non-nested case. Such equivalences can be used to construct GIC which, when the null hypothesis holds, is equivalent to use of the LR or BLR tests at different confidence levels for nested and non-nested distributions, respectively. Essentially, GIC can be based on any asymptotic test which uses the maximized values of the log-likelihood functions of the appropriate distributions.

8.5 A Generalized Information Criterion (GIC)

A new GIC procedure is proposed for the discrimination of distributional structures among a set of alternatives. The GIC is based on the equivalence between some well-known information criteria and hypothesis tests, and attempts to determine the false distributions based on sample information. Large differences between the maximized values of log-likelihood functions will lead to rejection of the distribution with the lower value. Discriminated distributions are separated into two categories, the superior and badly fitting categories. The distributions in the superior category perform within an acceptable tolerance level and there are no significant differences among their

performances. The GIC procedure may provide several alternatives rather than one particular distribution for a particular set of data. In the event of there being several sets of data, the distribution with the highest probability of acceptance in the superior category will be chosen.

Let $x_1, x_2, \ldots, x_n$ be n independently and identically distributed random observations. Denote $logL_j$ as the maximized log-likelihood value of distribution j ($j = 1, 2, \ldots, m$), with the ordering given as

$$logL_1 > logL_2 > \cdots > logL_m. \tag{8.9}$$

Then distribution j will be rejected in favour of distribution 1 if

$$GIC : -2logL_1 + T_0 < -2logL_j \tag{8.10}$$

where $T_0 > 0$. The value T_0 is the tolerance level required in order to reject distributions as being significantly different from each other, and is equivalent to the rejection region discussed in the previous section. For the nested case, T_0 can be expressed as

$$T_0 = c \tag{8.11}$$

whereas in the non-nested case, T_0 is given by

$$T_0 = 2z^*. \tag{8.12}$$

The motivation behind the GIC procedure is straightforward. First, for a given sample, select the distribution with the highest maximized log-likelihood value among $logL_j$ ($j = 1, 2, \ldots, m$). This distribution then belongs to the category of superior distributions and is also used as a standard for further inference. Second, reject the false alternatives among the remaining distributions in terms of the given tolerance. The distributions which perform within an acceptable tolerance level of the best fitting distribution are retained in the superior category, and the distribution with the highest

probability of acceptance over different sets of data will be chosen from the superior category.

8.6 Distribution Functions and Statistical Criteria

For a sample $x_1, x_2, \ldots, x_n$ of n independently and identically distributed random observations, the log-likelihood functions for the 3-parameter gamma, Weibull and lognormal distributions are given as follows:

Gamma:

$$logL = -n\alpha log\beta - nlog\Gamma(\alpha) + (\alpha - 1)\sum_{i=1}^{n} log(x_i - \gamma) - \sum_{i=1}^{n}(\frac{x_i - \gamma}{\beta}) \tag{8.13}$$

Weibull:

$$logL = nlog\alpha - n\alpha log\beta + (\alpha - 1)\sum_{i=1}^{n} log(x_i - \gamma) - \sum_{i=1}^{n}(\frac{x_i - \gamma}{\beta})^{\alpha} \tag{8.14}$$

Lognormal:

$$logL = -\frac{n}{2}log(2\pi\alpha^2) - \sum_{i=1}^{n} log(x_i - \gamma) - \frac{1}{2\alpha^2}\sum_{i=1}^{n}[log(x_i - \gamma) - \beta]^2 \tag{8.15}$$

in which β represents the scale parameter, α the shape parameter, γ the location parameter, and Γ is the gamma function. The 2-parameter versions of the above functions are obtained by setting the location parameter γ to zero in each case. In the above equations, $\beta > 0$, $\alpha > 0$ and $\gamma < x_i < \infty$ for $i = 1, 2, \ldots, n$. The density functions of the gamma and Weibull functions approach the exponential at $\alpha = 1$, are "J" shaped for $\alpha < 1$ and "bell" shaped for $\alpha > 1$, whereas the density for the lognormal function changes from being nearly symmetric to being heavily skewed as α is increased from 0.4 to 1.2. These values accommodate a variety of shapes which arise in practice in analysing real data.

The parameters of the three log-likelihood functions are estimated by maximum likelihood methods. Since the general maximum likelihood procedure will frequently fail to converge when the shape parameter is less than or equal to unity, an approach that circumvents this problem is used (for further details, see Bai et al. (1989)).

In the application to air quality assessment and control, the simplest non-nested discrimination problem is to test an hypothesis that the pollutant population has a specified distribution against the alternative that the distribution belongs to a non-nested alternative. Standard non-nested procedures can then be applied.

The choice of an appropriate test for a specific practical problem depends upon the properties of the test. The Cox test is generally complicated to use, involving the calculation of the asymptotic variances of the differences of the log-likelihoods between two distributions. Kent (1986) points out the problem of possible degeneracy of the Cox test. To introduce a consistent test, Epps et al. (1982) propose a method based on the comparison of the theoretical and empirical moment generating functions. However, since the empirical moment generating function will depart largely from the theoretical moment generating function, particularly for positively skewed distributions, the test will be inefficient. In order to avoid the disadvantages of such tests, Horowitz and McAleer (1988) developed the BLR test which is tractable, easy to use and applies for general parametric forms of distributions. Denoting the maximized values of the 2- and 3-parameter variants of a particular log-likelihood as $logL_0$ and $logL_1$, respectively, the BLR test is given in equation (8), when the null hypothesis is true that the location parameter is zero. The AIC and SIC are given in equations (3) and (4), respectively, with the 3-parameter distribution being selected when the inequalities are reversed.

In this chapter, two well-known criteria for testing goodness of fit are also considered. These are the chi-square (CHI) test and the Kolmogorov-Smirnov (KS) test. Classifying the n observations into k categories, the chi-square statistic is of the form (see Pearson (1900)):

$$CHI = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} \tag{8.16}$$

which has an asymptotic χ^2 distribution with $(k - l - 1)$ degrees of freedom when H_0 holds. The p_i are hypothetical probabilities, the f_i are empirical frequencies and l is the number of parameters estimated for each distribution (for further details, see

Kendall and Stuart (1979)). For the experiments conducted in this chapter, $k = 10$ and $l = 2$ or $l = 3$. The KS test, which is defined in terms of the maximum absolute difference between the sample distribution function $S_n(x)$ and the hypothetical distribution function $F_0(x)$ (see e.g. Bury (1975, p. 204)), is given by

$$D_n = \sup_x |S_n(x) - F_0(x)|. \tag{8.17}$$

Large observed values of the D_n statistic lead to rejection of the null hypothesis $F_0(x)$.

Based on the loss function recommended for assessing air quality models (see Fox (1981)), two performance criteria defined in terms of the relative root mean square error (RRMSE) are also used. The first criterion is based on the mean of RRMSE in the upper (U) percentiles, and the second is based on the mean of RRMSE in the entire or full (F) percentiles of the distribution. For an estimate $\hat{q}_{ij}$ of a quantity of interest q, these performance criteria are defined in terms of deviations from q in the percentiles of interest, where i denotes the specific percentiles estimated and j corresponds to the replication of experiments. The definitions of the upper percentile error and full percentile error are as follows:

$$UPE = \frac{1}{N}\sum_{j=1}^{N}[\frac{1}{1+(1-p)n}\sum_{i=pn}^{n}(\frac{\hat{q}_{ij}-q}{q})^2]^{0.5} \tag{8.18}$$

$$FPE = \frac{1}{N}\sum_{j=1}^{N}[\frac{1}{n}\sum_{i=1}^{n}(\frac{\hat{q}_{ij}-q}{q})^2]^{0.5} \tag{8.19}$$

where p is the location of the p-quantile which is chosen at the 98 per cent level for UPE and 100 per cent for FPE, N is the number of replications of the experiment, and n is the sample size. For present purposes, q denotes the percentile quantities related to the upper and full percentile errors.

8.7 Monte Carlo Experiments

In order to assess the performances of discrimination criteria, an extensive range of possible cases is considered. The shape parameter is examined over a wide range of possible values where the density functions are positively skewed: the shape parameter takes the values 2 and 6 for the gamma distribution; 2 and 4 for the Weibull distribution; and 0.5 and 0.9 for the lognormal distribution. In all cases considered in this chapter, the arbitrary scale parameter is set to unity, and the location parameter takes on the values 0 and 1 for the 2- and 3-parameter distributions, respectively. The lognormal distribution has the opposite behaviour to the gamma and Weibull distributions as the shape parameter is increased. For each entry in the tables, $N = 1000$ replications of the experiments are processed. The sample size used is $n = 365$, since it represents a common case, namely a full year of 24-hourly average observations.

The random sample generators used for the Monte Carlo experiments are DRNGAM, DRNWIB and DRNLNL for the gamma, Weibull and lognormal distributions, respectively. These are available as subroutines in the International Mathematical and Statistical Library (IMSL) in version 1.0 (April 1987). The same seed number (1234) is used to obtain the first random sample of the first of the 1000 replications. Varying the initial seed produces similar results to those reported in the chapter. Two subroutines, namely DCHIGF and DKSONE, are chosen from IMSL to perform the CHI and KS tests. All results are obtained on a VAX8700 mainframe computer at ANU.

8.8 Discrimination Between Two Non-Nested Distributions

The most common use of hypothesis tests and discrimination criteria arises in the case of having two candidate models. To compare the performance of different tests and criteria, each pair of possible combinations of the two-and three-parameter gamma, Weibull and lognormal non-nested distributions will be considered. The major pur-

pose of the Monte Carlo simulation experiments presented here is to re-examine the performance of different discrimination criteria, and to draw some useful conclusions as a guide for practice. It must be pointed out that some distributions are very similar in fitting the data, so that some of the discrimination criteria will not perform well. Theoretically, such simulation experiments could be used to expose the difficulties in discrimination of distributions that are very similar and give a necessary warning that care should be exercised. For further discrimination among a group of distributions, the results obtained in this section may be used to assist in the analysis of complicated problems, such as selection among six distributions, which will be discussed in Section 11.

The results of the Monte Carlo simulation experiments are shown in Tables 1 to 3. The main points are given as follows.

(i) When the three-parameter gamma is the true distribution, the BLR test performs very well in accepting the true null hypothesis, except when the three-parameter Weibull distribution is the alternative, in which case the probability of accepting the true null is below the nominal level of significance. On the other hand, in terms of rejecting the false null hypothesis, the BLR test provides low probabilities in cases where the false null is either the three-parameter Weibull or lognormal distribution. For the Weibull distribution, power is increased when the shape parameter is increased to 6, but power is decreased in the lognormal case. Generally, the BLR test has high power in rejecting the false two-parameter null model, except in the single case when the two-parameter lognormal distribution is false and the value of the shape parameter of the alternative three-parameter gamma distribution is 6.

In terms of the probability of accepting the correct distribution, the remaining criteria perform differently when the value of the shape parameter of the true distribution is 2. The KS statistic accepts the true null with high probability, but CHI has slightly lower probabilities of accepting the true null, when compared with the confidence level. Note that the confidence level of the KS and CHI tests used here is 98 per cent, which is

a value chosen from empirical air pollution applications. The same confidence level for the KS and CHI tests will be also used throughout this chapter. The ranking of these criteria are AIC and SIC (which are identical because they have the same number of parameters), and then FPE. The worst is UPE, whose probability is below 50 per cent. When the shape parameter is increased to 6, all criteria except UPE have significantly improved performances, even though they are still not satisfactory.

(ii) The results of discrimination between two non-nested models are quite similar to the above when the true distribution is the two-parameter gamma. The BLR test provides high probabilities of accepting the true null hypothesis, in general. However, when testing the two-parameter gamma against the three-parameter Weibull, or testing the two-parameter gamma against the three-parameter lognormal distribution, for values of the shape parameter being 2 and 6, the BLR test performs poorly. If the alternative is the two-parameter Weibull distribution, the power in rejecting the false null is quite low when the shape parameter is 2, but improves appreciably when the shape parameter is increased to 6. Similar results can be obtained when the two-parameter lognormal is the alternative; that is, the BLR test performs very well when the value of the shape parameter is 2, but becomes worse when the value of the shape parameter is increased to 6.

Of the remaining criteria, SIC works very well in most cases, with high probabilities of accepting the true null hypothesis. AIC is generally good but sometimes has low probabilities of accepting the true null, such as when the alternative is the three-parameter gamma distribution. Other methods are unstable. For example, FPE performs very well if the alternative is the two-parameter lognormal distribution, and performs well when the alternative is the two-parameter Weibull with shape parameter taking the value 6. However, SIC is worst when the alternative is either the three-parameter Weibull or lognormal distribution, with the shape parameter taking the values 2 and 6. Based on these results, SIC is the preferred criterion, in general.

(iii) When the three-parameter Weibull distribution is correct, the probabilities

of accepting the true null hypothesis for the BLR test all exceed the nominal level of significance. However, in terms of rejecting the false null hypothesis, the BLR test performs differently according to various pairs of tested distributions, and the values of the shape parameters of the true distribution. Testing between the three-parameter Weibull and gamma distributions when the shape parameter is 2, the BLR test has low power in rejecting the false gamma null hypothesis, especially at the 99 per cent confidence level. An improvement in the performance of the BLR test can be seen when the shape parameter is increased to 4, but the power in rejecting the false model is still relatively low at the 95 and 99 per cent confidence levels when compared with some other pairs of non-nested models. Similar results are also obtained when testing between the three-parameter Weibull and lognormal distributions, in which the power of the BLR test is increased when the value of the shape parameter of the Weibull distribution is increased. In testing the three-parameter Weibull versus the two-parameter gamma distribution when the shape-parameter is 2, the BLR test has very good power in rejecting the false model, but power decreases significantly when the shape parameter is increased to 4. This is also the case for testing the three-parameter Weibull distribution against the two-parameter lognormal distribution, in which the power of the BLR test changes when the value of the shape parameter is increased.

Of the others, AIC performs consistently well in most cases, except for the single case where the probability of accepting the true null hypothesis decreases to 0.89 for discriminating between the three-parameter Weibull and lognormal distributions when the value of the shape parameter is 4. SIC produces the same results as AIC when discriminating among three-parameter distributions, and SIC performs worse when discriminating between the three-parameter Weibull and either the two-parameter gamma or lognormal alternatives. The lowest probability of accepting the true null is observed when the shape parameter equals 4 for the case of the three-parameter Weibull against the two-parameter gamma distribution. Compared with AIC, FPE also performs consistently well in most cases but also decreases the probability of accepting the true null hypothesis when the shape parameter is 4 in the case of the three-parameter Weibull

against the two-parameter lognormal distribution. Surprisingly, UPE is the best in cases where the others are worsening, but has the lowest probability of accepting the true null hypothesis when the shape parameter is 2 for the three-parameter Weibull versus the two-parameter gamma distribution. The performance of KS is good in terms of accepting the true null, but CHI has a probability of acceptance of the true null slightly below the confidence level.

(iv) For the same discrimination procedures but with the sample coming from the two-parameter Weibull distribution, the results obtained in the simulation experiments are similar to the above. In terms of the probability of accepting the true null hypothesis, the BLR test provides very good performance in all cases. However, when testing the two-parameter Weibull distribution against either the three-parameter gamma or lognormal distributions, the power in rejecting the false null hypothesis is very poor. When the shape parameter is increased from 2 to 4, the performance of the BLR test improves when testing against the three-parameter gamma, and deteriorates against the three-parameter lognormal distribution. The probability of rejecting the false null hypothesis is low at the 99 per cent confidence level when the shape parameter is 2, but improves considerably when the shape parameter is increased to 4. For the last pair of the two-parameter Weibull versus the two-parameter lognormal distribution, the probability of rejecting the false null hypothesis for the BLR test is 100 per cent.

On the other hand, AIC and SIC provide high probabilities of accepting the true null hypothesis in all cases. Similar results will be obtained when testing against the two-parameter distributions and, as expected, AIC will have a lower probability of accepting the true distribution than SIC if the alternative is the three-parameter distribution. UPE is comparable to AIC and SIC in all cases, and is sometimes the best. FPE has a lower probability of accepting the two-parameter Weibull distribution against both the three-parameter gamma and lognormal distributions when the shape parameter is 2, and has a high probability for the former when the shape parameter is increased to 4, but becomes worse for the latter. KS and CHI have high probabilities of accepting the true null hypothesis in testing against the two-parameter lognormal

distribution when the shape parameter equals 2 or 4. Similar results will be obtained against the two- and three-parameter gamma distributions if the value of the shape parameter is 4.

On the basis of these results, great care should be exercised in testing the two-parameter Weibull distribution against both the three-parameter gamma and lognormal distributions using different criteria. The user should also be careful in the case of the two-parameter Weibull versus the two-parameter gamma distribution. AIC, SIC and UPE are generally reliable. FPE will perform very well for the two-parameter Weibull versus either the two-parameter gamma or lognormal distribution. Care should be exercised in using the KS and CHI tests because they often have low powers for rejecting the false null, although their probabilities of accepting the true null are often very high.

Finally, the simulation experiments are conducted to discriminate among the two- and three-parameter gamma, Weibull and lognormal distributions, based on samples taken from the lognormal distribution. The results are included in Table 3, and the main points should be noted as follows.

(v) When the three-parameter lognormal distribution is correct, the BLR test provides accurate acceptances of the true null hypothesis in most cases. The exception is the three-parameter lognormal versus the three-parameter gamma distribution when the shape parameter is 2, with the probabilities being below the given confidence level. In terms of the power of rejecting the false null hypothesis, the BLR test has low probabilities of accepting the true null hypothesis in the case of the three-parameter lognormal null against the three-parameter gamma alternative when the value of the shape parameter is 0.5. However, the probabilities are increased significantly when the shape parameter of the three-parameter lognormal distribution is set at 0.9. AIC, SIC, and FPE perform very well in most cases, but there are exceptions when the three-parameter lognormal null is tested against the three-parameter gamma alternative. CHI and KS consistently provide high probabilities of accepting the true null.

Unfortunately, UPE works poorly in all pairs of tested distributions. The results show that care should be exercised in applying the discrimination procedures to the three-parameter lognormal null versus the gamma alternative distribution when the value of the shape parameter is 0.5. UPE is not reliable here, reflecting the fact that the estimation of the upper percentiles is unstable.

(vi) Similar observed patterns are obtained when the true distribution is the two-parameter lognormal distribution. The results for accepting the true null and rejecting the false null are similar to the case when the true distribution is the three parameter version, with the probabilities of accepting the true null generally being lower than those given in point (v) above. The BLR test still has problems in rejecting a false null hypothesis when the shape parameter is 0.5, but improves considerably when the shape parameter is increased to 0.9, in testing the two-parameter Weibull against both the three-parameter gamma and lognormal distributions. This behaviour is also found in testing against the two-parameter gamma distribution. Similarly, AIC, SIC and FPE work well in most cases, except for the two-parameter lognormal against the three-parameter gamma distribution. Once again, the performance of KS is good in accepting the true null, but CHI has slightly lower probabilities of accepting the true null. For SIC, its probability also decreases substantially against the two-parameter gamma distribution. Once again, UPE performs poorly in all cases.

In summary, different discrimination criteria for selecting between two distributions are re-examined. The experimental results show that there is no criterion which consistently performs well in all cases. Great difficulties arise whenever three-parameter distributions are involved because the extra (location) parameter significantly improves empirical fitting of the data. It is argued here that efforts made to develop a new criterion cannot help resolve this problem since differences between the true and false distributions are too small.

The large sample sizes used in the experiments are commonly found in practice, so the problems encountered here have general meaning for application of such statistical

techniques to real situations. Based on the results of these simulation experiments, the user should exercise care in employing these criteria to discriminate between two three-parameter distributions, especially when the gamma distribution is involved. Extensive care should be exercised in situations when three-parameter distributions are involved because false three-parameter non-nested distributions can frequently perform better than true two-parameter distributions. In general, AIC is good in most cases but might perform worse than SIC when the true distribution has only two parameters and better than SIC when the true distribution has three parameters. FPE has very good performances and is generally quite similar to AIC and SIC. UPE is not generally reliable, but is superior in some cases where the others are performing badly. However, UPE is particularly important for air quality management because this criterion indicates errors at the upper percentiles. Unfortunately, KS and CHI generally have low powers for rejecting the false null, so that care should be exercised in applications.

8.9 Discrimination Among Three Non-Nested Distributions

Discrimination among three non-nested distributions is more difficult than between two, as was reported in the previous section. Standard hypothesis tests are generally not valid if the number of tested distributions exceeds two. On the other hand, an additional candidate will affect other criteria used in selecting the true distribution. The emphasis here is placed on discrimination among the two- or three-parameter gamma, Weibull and lognormal distributions. Monte Carlo experiments are used to examine the performance of the discrimination criteria as each distribution is taken to be true. The discrimination procedures might be expected to expose certain distributions as being good approximations to others over different parameter ranges, even when the distributions come from different parametric families.

One of the purposes of the experiments is to examine the outcomes arising from the situation where the data are generated from, say, the three-parameter gamma distribu-

tion, and discrimination is made between it and its two-parameter counterpart. Such considerations often arise in practice, and has especial importance for fitting models to air quality data, as will be discussed later. For each experiment, selection will proceed among the two- and three-parameter gamma, Weibull and lognormal distributions. Instead of using the BLR test, the Kullback-Leibler (KL) information criterion based on the maximized log-likelihood values will be used. In this case, AIC and SIC will be equivalent to KL since the penalties for AIC and SIC are identical. The sample size is 365, which is considerably large for practical purposes. As mentioned previously, GIC is primarily designed to discriminate among three or more distributions simultaneously, in which case the LR test is not applicable. Thus, GIC is expected to complement the existing discrimination criteria.

The results of experiments using different criteria are shown in Table 4. The main points from the table are as follows.

(i) When the three-parameter gamma distribution is correct, all of the criteria in Table 4 perform poorly. The best is KL (or equivalently, AIC and SIC), although it is only correct 59.6 per cent of the time when the shape parameter is 2, and 52.6 per cent of the time when the shape parameter is increased to 6. FPE is reasonably close to KL. KS and CHI perform well in accepting the true null but they also accept the false three-parameter models frequently. UPE is worst and selects the false three-parameter Weibull distribution 576 times in 1000 experiments. When the value of the shape parameter is 2, the three-parameter Weibull distribution has a rather high probability of 35.5 per cent of being the best fitting distribution, but when the shape parameter is increased to 6, the three-parameter lognormal distribution becomes a good approximation to the true distribution. Somewhat surprisingly, when the value of the shape parameter is 2, the two-parameter lognormal is the best fitting of the two-parameter distributions. Although the two-parameter gamma becomes the best when the value of the shape parameter is increased to 6, the two-parameter lognormal distribution is still a good approximation to the true three-parameter gamma distribution.

(ii) The performances of the discrimination criteria are improved substantially when the true distribution is the two-parameter gamma. KL, AIC and SIC are the best when the value of the shape parameter is 2, but FPE is best when the shape parameter is 6. KS and CHI have high probabilities of accepting the two- and three-parameter gamma and Weibull distributions when the values of the shape parameter are 2 and 6. UPE remains the worst. For the other three-parameter distributions, the ranks are similar to the case where the true distribution is the three-parameter gamma, which suggests that changing the value of the location parameter from 1 to 0 does not effect the results qualitatively.

(iii) Most of the criteria perform very well when the three-parameter Weibull distribution is true. In this case, UPE becomes the best when the value of the shape parameter equals 2 or 4. The high probabilities of accepting the true distribution indicate that both three-parameter gamma and lognormal distributions are not good approximations to the three-parameter Weibull distribution. Once again, the two-parameter lognormal distribution is the best when the shape parameter is 2, but it is replaced by the two-parameter gamma distribution when the shape parameter is 4. These results suggest that the member of the same family of distributions with two parameters is not necessarily as good as a member from a non-nested counterpart. The two-parameter lognormal distribution is the best approximation to the three-parameter Weibull distribution if the data are heavily skewed, as is often encountered in air pollution applications.

(iv) Similar observed patterns are obtained when the true distribution is the two-parameter Weibull. Most of the criteria perform well in accepting the true distribution. Only KS and CHI have low powers for rejecting the false two- and three-parameter gamma, and the three-parameter Weibull and lognormal distributions. UPE performs the best of the criteria.

(v) When the value of the shape parameter is 0.9 for the underlying three-parameter lognormal distribution, most of the criteria perform very well, with the exception of

UPE. However, when the shape parameter is decreased to 0.5, power is reduced substantially. KS and CHI have significantly higher probabilities for rejecting the false distributions when the value of the shape parameter is 0.9, as compared with the cases when the true distributions are the two- or three-parameter gamma and Weibull distributions. However, the false three-parameter gamma distribution will be frequently accepted by KS and CHI tests if the shape parameter is decreased to 0.5. UPE has the worst. It is interesting to note that, of the two-parameter distributions, the two-parameter lognormal distribution is the best approximation to the three-parameter lognormal. Based on these results, it is clear that, no matter which three-parameter distribution is true, the two-parameter lognormal will always be a good representation of the data if selection is restricted to two-parameter distributions and the sample data are quite skewed. These simulation results are consistent with the recommendations of many air pollution specialists (for example, Larsen (1971,1974) and Benarie (1980)) that the two-parameter lognormal distribution is the best for fitting urban air pollutant concentrations. However, care should still be exercised because the two-parameter version might not be the best if alternative three-parameter distributions are under serious consideration.

(vi) When the true distribution is the two-parameter lognormal, most criteria apart from UPE perform very well. The improvement in rejecting the false models using the KS and CHI tests can also been seen when the value of the shape parameter is 0.9, as compared with the cases where the true distributions are the two- or three-parameter gamma and Weibull distributions. Their performance in rejecting false models will deteriorate when the shape parameter is reduced to 0.5. Similarly, the result of selection among three-parameter distributions has no significant changes when the data are from the two-parameter lognormal distribution.

From the results of the experiments, the criteria used are not always consistent, especially for the underlying two-and three-parameter gamma distributions. The same problem is found when the three-parameter lognormal is the true distribution with the value of the shape parameter taken as 0.5. A major consequence of these findings is

that the criteria will reject the true distribution frequently because of a slightly lower value of the maximized log-likelihood function relative to that of the false distribution. To avoid this outcome, GIC is recommended. The principle of GIC is to reject the distribution if the evidence shows that it performs very poorly. With small differences in the performances of the distributions, GIC will not initially reject the distributions but simply determine if they are similar for the particular sample. The true distribution will be included in the selection and should have the highest probability of acceptance. The results given in Table 5 emphasize the following points.

(a) GIC works consistently very well in accepting the true distribution as compared with the remaining criteria. In all cases, the probability of acceptance is over 94 per cent and, in many cases, the true distribution is always accepted.

(b) If the distribution is false, GIC will reject it when the value of its log-likelihood function is lower than the tolerance level. In the event that a false distribution is accepted by GIC, it means that such a distribution fits the data as closely as the true one in terms of the given critical region. In all cases, the probability of accepting the false model is less than 90 per cent, which is much lower than the result obtained for the true model. Some three-parameter distributions, such as the Weibull and lognormal, have large numbers of acceptances when the true distribution is the three-parameter gamma. This only serves to indicate that they could be good approximations for the underlying three-parameter gamma distribution over certain ranges of the shape parameter.

(c) When selection is among three-parameter distributions but the parent distribution is a two-parameter version, its three-parameter counterpart is consistently selected. This is not always the case when selection is among two-parameter distributions and the underlying distribution is a three-parameter version. Since there is no true model among the distributions under consideration, GIC is not guaranteed to select the correct model.

In conclusion, experiments have been conducted to discriminate among three non-

nested distributions. Such discrimination has employed some well-known criteria and tests, such as KL, AIC, SIC, CHI and KS, as well as some criteria recommended for application to air quality data, such as UPE and FPE. The results of the experiments have shown that the criteria are not consistent in all cases where the underlying distributions and parameter values are altered, so that the true distribution can be rejected quite frequently. Therefore, GIC was applied in the same experiments, and was shown to perform very well for various underlying distributions and parameter values. GIC is useful in avoiding over-rejection of the true distribution, which is important in applications. Moreover, GIC indicates that some distributions may fit the data equally well. In addition, through Monte Carlo experiments, it was shown that the two-parameter lognormal is a good approximation to the three-parameter gamma, Weibull and lognormal distributions when selection is restricted to two-parameter distributions and the sample is heavily skewed. In terms of acceptance of the true distribution, it is necessary to consider three-parameter distributions against the two-parameter lognormal in such cases, because further discrimination might yield different results. Such a suggestion should be very useful when applying statistical techniques to air quality applications.

8.10 Discrimination Among Five Non-Nested Distributions

Similar statistical techniques used for discrimination among three non-nested distributions can easily be extended to discriminate among five non-nested distributions simultaneously. That is, for any given model, there will be a maximum of four non-nested alternative distributions. Additional distributions generally affect the discrimination criteria in terms of selecting the true distribution, but the nature of such distributions is also important, namely whether they are good approximations to each other. For example, from the results of the previous two sections, it is known that the three-parameter Weibull distribution is often a good approximation to the three-parameter gamma distribution for certain parameter values, but the two-parameter lognormal dis-

tribution does not fit either the two- or three-parameter gamma distribution especially well.

The advantage of including five distributions as candidates for a particular sample is that this could help to avoid the problem found in the previous section, namely where an incorrect selection occurs when the procedure is restricted to only two-parameter distributions, but a heavily skewed three-parameter distribution is true. In principle, criteria such as KL, AIC, SIC, CHI, KS, UPE and FPE have no restrictions on the number of distributions considered. GIC is also constructed especially for discrimination when the number of distributions exceeds two. The task here is to investigate whether they are capable of providing satisfactory results in applications.

At first, the Monte Carlo experiments are used to re-examine the different criteria except GIC, and the results are given in Table 6. The following points should be noted.

(i) Consider the case of the three-parameter gamma distribution being correct. The major concern here is with a pair of two-parameter distributions. When the value of the shape parameter equals 2, KL and AIC remain the same, and SIC and FPE are changed slightly. KS and CHI are good in terms of accepting the true distribution but false distributions, such as the three-parameter gamma and Weibull, are also frequently accepted by these two tests. However, when the shape parameter is increased to 6, all of the criteria have substantially decreased probabilities of accepting the true distribution. KL is the only criterion with a probability of acceptance in excess of 50 per cent. The worst is SIC, which favours the two-parameter lognormal distribution. UPE consistently shows that the three-parameter Weibull distribution is the best, which means that the three-parameter Weibull distribution provides lower errors at the upper percentiles. From the outcomes of these experiments, the qualitative findings are quite different for each discrimination criterion.

(ii) The results from the experiments will largely differ from the above when the true distribution is the two-parameter gamma. SIC changes from being the worst to the best, especially when the value of the shape parameter is increased from 2 to 6.

Unfortunately, AIC and the remaining criteria provide probabilities which are all under 50 per cent. When the value of the shape parameter is 6, KL and FPE are all in favour of the three-parameter lognormal, whereas UPE still selects the three-parameter Weibull distribution. Although SIC and AIC have relatively high probabilities of accepting the true distribution, the results show that the three-parameter distributions have much stronger influence than their two-parameter counterparts.

(iii) When the three-parameter Weibull distribution is correct with the shape parameter equal to 2, KL and FPE have the same probabilities as compared with point (iii) of Section 9 in considering a pair of two-parameter distributions. AIC has only a very small change in the probability of accepting the true distribution but the probability of SIC has decreased significantly. UPE accepts the two-parameter gamma distribution rather than the true distribution. However, when the value of the shape parameter is increased to 4, UPE becomes the best criterion, with a high probability of accepting the true distribution, while KL, AIC, SIC and FPE have worsened substantially. The KS and CHI tests have poor performances in rejecting the false distribution when the shape parameter is 2, but their powers in rejecting the false three-parameter gamma is significantly increased when the value of the shape parameter is increased to 6. This is not the case for the remaining false distributions.

(iv) Considering the two-parameter Weibull distribution as true, SIC performs well when the value of the shape parameter is 2 or 4. AIC and UPE also perform quite well, but KL provides a relatively low probability of accepting the true distribution. KS and FPE perform quite well in accepting the true distribution, but are less powerful in rejecting the false models. Based on these results, the selections are strongly influenced by the addition of two- or three-parameter distributions. All of the criteria perform poorly and become unreliable when the values of the parameters and the nested true distribution are varied.

(v) When the three-parameter lognormal distribution is correct, all of the criteria remain the same when the value of the shape parameter is equal to 0.9, which indicates

that the additional two-parameter distributions do not affect the outcome of the discrimination criteria. Such stability remains when the shape parameter is decreased to 0.5. This is the only situation in which the results of the experiments are not changed when the number of distributions considered is increased from three to five.

(vi) There are noticeable differences in the results when the true distribution is changed to the two-parameter lognormal. The probabilities of accepting the true distribution have decreased slightly for all the criteria when the value of the shape parameter is equal to 0.9, but have decreased substantially when the value is reduced to 0.5, apart from SIC. Such changes arise mainly from the three-parameter gamma distribution, which indicates that it is a good approximation to the two-parameter lognormal distribution. Compared with the case where the gamma and Weibull distributions are true, when the true model is the lognormal distribution, the results of discrimination among five distributions are not much altered from discrimination among three distributions.

As expected, discrimination among five distributions is more difficult than among three. Compared with the latter case, the criteria used here perform either poorly or unreliably, and the results generally become much worse after involving the additional distributions, especially in the case when the two- or three-parameter gamma distribution is true. The gamma distribution has a probability of acceptance below 50 per cent for most of the criteria when the value of the shape parameter equals 6. Such performances are not satisfactory and the use of these criteria is subsequently in doubt. Therefore, there is still a need to provide more reliable selection procedures than those used above.

The results from the experiments in using GIC are shown in Table 7. It is interesting to note that GIC performs very well, consistently providing the highest probability of accepting the true distribution in all cases. The main points should be noted as follows.

(a) When the three-parameter gamma is the true distribution, the three-parameter Weibull distribution is a good approximation when the values of the shape parameter

are 2 or 6. When the distribution becomes more symmetric at $\alpha = 6$, the two- and three-parameter lognormal distributions become close to the three-parameter gamma distribution. When the true distribution is the two-parameter gamma, the two- and three-parameter Weibull are quite good if the sample is skewed, but the three-parameter Weibull and lognormal are superior when the distribution is more symmetric.

(b) The additional two-parameter distributions do not change the discrimination results when the Weibull becomes the true distribution in both the two- and three-parameter cases. If the sample is heavily skewed when the value of the shape parameter is 2, the three-parameter gamma is close to both the two- and three-parameter Weibull distributions. When the value of the shape parameter is 4, the three-parameter lognormal distribution will approach the three-parameter gamma distribution.

(c) There is no difficulty in discriminating between the underlying lognormal distributions for both the two- and three-parameter cases when the value of the shape parameter is 0.9. In this respect, the lognormal distribution is different from the gamma and Weibull distributions. However, the result is changed when the value of the shape parameter is decreased to 0.5, and the three-parameter gamma distribution becomes a very good approximation to the lognormal distribution.

(d) The most important point to note here is that the three-parameter distributions will generally be affected only slightly by the presence of the two-parameter distributions, but two-parameter distributions will be more heavily influenced by the addition of the three-parameter distributions. Based on the Monte Carlo simulations, when the three-parameter distribution is true, the probability of accepting the true distribution for GIC will be greater than 90 per cent in large samples. On the other hand, if the true distribution has two parameters, the probability would be slightly in excess of 85 per cent in large samples, such as when the two-parameter gamma is the true distribution. Such results indicate that if the two-parameter distribution has the highest probability of acceptance among the remaining distributions, this two-parameter distribution is likely to be the true one, although the probability of accepting the three-parameter

distribution is close to that of its two-parameter counterpart. This result arises because three-parameter distributions generally fit the data better than two-parameter models.

In summary, different criteria have been re-examined through Monte Carlo experiments, and the weaknesses of the criteria used here have been highlighted. The proposed GIC has performed very well and consistently provides high probabilities of acceptance of the true distribution in all cases. It is, therefore, recommended that GIC should always be used in discriminating among five non-nested distributions.

8.11 Discrimination Among Six Nested and Non-Nested Distributions

It is a difficult task to discriminate among a number of distributions simultaneously. When these distributions contain both nested and non-nested relations, the problems become even more complicated. As discussed in Section 3, the gamma, Weibull and lognormal distributions are quite similar to each other for certain parameter values, so that each could be a candidate for a particular set of data. This is the usual case encountered in applications. For illustrative purposes, we introduced some simple procedures for discriminating among two- and three-parameter gamma, Weibull and lognormal distributions related to air quality applications.

Basically, the selection criteria used in the previous section can be applied in fairly general situations, where both nested and non-nested distributions can be considered. Discrimination procedures for six distributions are similar to those for three or five distributions. In this section, we consider discriminating between a null model and five other distributions, namely a nested alternative and four non-nested alternatives. However, the simulation results are quite different even though the sampling experiments are the same. The reason is simply that the greater is the number of alternative distributions considered, the greater are the chances of selecting the incorrect distribution in small samples. On the other hand, all criteria perform poorly when there

are strong similarities among some three-parameter distributions such as the three-parameter gamma and Weibull distributions. This problem arises generally for all existing selection criteria. The performances of each criterion are discussed below.

(i) Consider the situation when the true distribution is the 3-parameter gamma, and the two-parameter gamma distribution becomes one of the six candidates. Before the experiments are conducted, it might be predicted that the results should change only slightly since the two-parameter distribution might be expected to have little impact on the three-parameter distributions. In Table 8, when the value of the shape parameter is 2, all of the criteria remain largely unchanged, but most of them except KL have largely decreased probabilities of accepting the true distribution when the value of the shape parameter is increased to 6. According to the results of the nested discrimination procedure (see Bai et al. (1990)), the two-parameter gamma distribution becomes a very good approximation to the three-parameter version as the distribution tends to symmetry. KL is the only criterion with a probability exceeding 50 per cent, and the others perform very poorly. AIC and SIC favour the two-parameter gamma, but UPE favours the three-parameter Weibull distribution.

(ii) Similar patterns are obtained when the true distribution is the two-parameter gamma. While SIC is best, followed by AIC, they are steady for both values of the shape parameter. Other criteria have much lower probabilities, especially KL, which has decreased substantially. The results show that the three-parameter gamma distribution has a strong influence on the outcomes of the discrimination procedures for the underlying two-parameter gamma distribution.

(iii) When the three-parameter Weibull distribution is true with the shape parameter set at 2, most of the criteria are very stable and are little influenced by the addition of the two-parameter Weibull distribution. Similar results can also be seen when the shape parameter is increased to 4, except that UPE has substantially decreased probabilities of accepting the true distribution. Surprisingly, KS and CHI have high powers for rejecting the false two-parameter Weibull distribution when the shape parameter is

2, but power will be decreased when the value of the shape parameter is increased to 4.

(iv) Adding the three-parameter Weibull distribution has significantly altered the results of most criteria when the true distribution is the two-parameter Weibull distribution. KL now performs the worst and UPE has decreased by more than one-half. AIC has also reduced probabilities, and only SIC remains similar to the results obtained before the addition of the three-parameter Weibull distribution. It is interesting to note that the powers of KS and CHI are high for rejecting the false two-parameter lognormal distribution when the values of the shape parameter are 2 and 4. However, this is not the case for the remaining false distributions.

(v) When the three-parameter lognormal distribution is correct, the results are similar to those obtained before adding the nested two-parameter lognormal distribution. This simply indicates that all of the two-parameter distributions, including the nested members, are not close to the three-parameter lognormal distribution. Only the three-parameter gamma distribution is close to the three-parameter lognormal.

(vi) The results of discrimination among six distributions could be expected to defer to those obtained in point (vi) of the previous section when the true distribution is the two-parameter lognormal and the three-parameter lognormal distribution is involved. The influence from the three-parameter lognormal distribution can be seen from KL, which has a zero probability of accepting the true distribution but favours the nested three-parameter alternative with very high probability. This is because the three-parameter lognormal distribution will be chosen even if it has only a slightly higher maximized log-likelihood value than the two-parameter distribution. SIC and AIC have much improved results since they penalize the number of parameters and/or observations used, but they still have decreased probabilities in comparison with the results obtained previously. The other criteria have substantially decreased probabilities, providing very poor performances in selecting the true distribution.

Based on these results, the selection of the underlying two-parameter lognormal

distribution is also affected significantly by adding the nested three-parameter alternative. It is not surprising that all of the criteria perform very poorly because they did not work well even without involving the nested cases. Table 8 provides further evidence that these criteria are not capable of discriminating among distributions which are quite similar to each other.

Of course, discriminating among six distributions simultaneously is also a difficult task for GIC. The reason is simply that the three-parameter distribution will perform better than, or at least similarly to, the two-parameter distribution when the latter distribution is true. As mentioned previously, GIC selects the 3-parameter distribution if its probability of acceptance is over 90 per cent. There is no doubt that this will be satisfied by the three-parameter distribution even though the true distribution has only two parameters.

When using the LR test, if the value of the log-likelihood function for the three-parameter distribution is close to that of the two-parameter distribution, it is inferred that the sample comes from the two-parameter distribution. By using GIC, if the probability of acceptance of the two-parameter distribution is close to that of the three-parameter nested alternative, it means that the values of the log-likelihood functions for these two nested distributions are very close and the data are inferred as being generated from a two-parameter distribution. From the empirical results obtained in the previous section, when the two-parameter distribution has a probability of acceptance in excess of 85 per cent, the two-parameter distribution would be expected to be the true distribution, regardless of the performance of the three-parameter alternative. However, when the two-parameter distribution has a much lower probability of acceptance, it means that the values of the log-likelihood functions for the two- and three-parameter distributions have departed significantly from each other, so that the sample can be determined as being generated from a three-parameter distribution. Based on the results in Table 8, GIC will select the three-parameter distribution with the highest probability of acceptance, which should be over 90 per cent. This suggestion is based on the results from experiments for the discrimination procedures in both

the nested and non-nested cases.

Based on the analysis above, the matter of selecting the true distribution from among the six distributions is straightforward. The results are presented in Table 9, and the following points should be noted.

(a) When the three-parameter gamma distribution is true, with the value of the shape parameter set at 2, the result is unchanged from previous findings, with the two-parameter gamma distribution having no influence on the selection at all. When the shape parameter is increased to 6, the probability of accepting the two-parameter gamma distribution is still less than 85 per cent. Other distributions have low probabilities of being accepted and the three-parameter gamma distribution is selected. On the other hand, although involving the three-parameter gamma distribution, the probability of other distributions being selected is unchanged when the true distribution is the two-parameter gamma. The two-parameter gamma distribution has a high probability of acceptance which is comparable with the three-parameter version, namely over 85 per cent, and the result indicates that the values of the log-likelihood function are very similar. In such cases, the two-parameter gamma distribution will be selected and the three-parameter version will be rejected.

(b) Similar results will also be obtained when the Weibull distribution is correct. The results are the same as previously when the true distribution is the three-parameter Weibull for both values of the shape parameter. When the true distribution is the two-parameter Weibull, its acceptance is very similar to that of its three-parameter counterpart because these two distributions have similar log-likelihood values.

(c) When the lognormal distribution is correct, similar patterns will be obtained as above. If the true distribution is the three-parameter lognormal, the probability of accepting the true distribution remains unchanged for both values of the shape parameter. When the two-parameter lognormal distribution is correct, the results do not change significantly. Based on the similar probabilities of acceptance of the two- and three-parameter lognormal distributions, the log-likelihood values of these two

distributions are very similar, thereby providing evidence that the sample comes from the two-parameter distribution.

(d) The best feature of GIC is that it determines the probability of accepting each distribution consistently for particular sample sets, while additional non-nested distributions will not significantly affect the probabilities of existing candidates if the true distribution is already under consideration. Based on the Monte Carlo experiments, the discrimination criteria listed in Table 8 fail to accept the true distribution frequently. However, the results show that GIC is a useful tool in discriminating among six nested and non-nested distributions in that it has a high probability of accepting the true distribution.

8.12 Concluding Remarks

In this chapter, some important issues in discriminating among nested and non-nested distributions have been raised. Existing problems in the literature are reported and discrimination procedures for non-nested distributions are re-examined. Such procedures are generally used to discriminate between two distributions, such as in classical nested hypothesis testing. The emphasis here is placed on discrimination among a set of three, five or six distributions where standard hypothesis testing procedures are not valid. From the practical point of view it is argued that, because the true distribution is unknown, the more distributions that are considered, the greater will be the chances of selecting the most appropriate distribution. An example is given in Section 10 whereby the two-parameter lognormal distribution is selected when discrimination is restricted to three two-parameter distributions, but the true distribution has three parameters. However, it is shown in this chapter that, in using different criteria, the greater the number of distributions that are considered, the greater will be the difficulty in selecting the true distribution. How to balance usefulness against selection of the true model is a major concern of this chapter.

From the Monte Carlo simulations, some well-known discrimination criteria and

tests such as KL, AIC, SIC, CHI, and KS, as well as some new criteria such as UPE and FPE, are shown to be neither consistent nor reliable. Some of these criteria, such as KL, AIC and SIC perform poorly in terms of accepting the true distribution, and such probabilities are decreased substantially as the number of distributions is increased. The KS and CHI tests are often accurate in accepting the true distribution, but generally have low powers in rejecting false distributions. To solve this problem, GIC is proposed specifically for discriminating among a set of distributions, where the number of distributions exceeds two. It has been shown that GIC performs exceedingly well and provides high probabilities of accepting the true distribution. Compared with other criteria, GIC is consistent and reliable, and it is also simple to use. When the true distribution is included among the distributions, additional candidates will not significantly affect the probability of accepting the true distribution.

TABLE 8.1

Sizes and powers of the BLR test and probabilities of accepting the true model using two tests, two discrimination criteria and two performance criteria over 1000 replications of random samples of size $n = 365$ ($\beta = 1$)

True Model	False Model	Significance Level For BLR Test	BLR Test Size	BLR Test Power	Tests and Criteria AIC	SIC	CHI	KS	UPE	FPE
G3	W3	.10	0.195	0.447						
α=2.0		.05	0.111	0.348	0.645	0.645	0.960	1.000	0.424	0.635
γ=1.0		.01	0.015	0.123						
G3	LN3	.10	0.030	0.903						
α=2.0		.05	0.017	0.858	0.951	0.951	0.960	1.000	0.950	0.937
γ=1.0		.01	0.003	0.659						
G3	W2	.10	0.000	1.000						
α=2.0		.05	0.000	1.000	1.000	1.000	0.960	1.000	0.779	1.000
γ=1.0		.01	0.000	1.000						
G3	LN2	.10	0.000	1.000						
α=2.0		.05	0.000	0.998	0.999	0.976	0.960	1.000	0.551	0.993
γ=1.0		.01	0.000	0.982						
G2	W3	.10	0.301	0.351						
α=2.0		.05	0.209	0.272	0.738	0.960	0.964	0.990	0.442	0.491
γ=0.0		.01	0.057	0.096						
G2	LN3	.10	0.046	0.860						
α=2.0		.05	0.034	0.803	0.956	0.995	0.964	0.990	0.946	0.801
γ=0.0		.01	0.006	0.560						
G2	W2	.10	0.072	0.746						
α=2.0		.05	0.044	0.679	0.864	0.864	0.964	0.990	0.508	0.618
γ=0.0		.01	0.012	0.452						
G2	LN2	.10	0.000	0.998						
α=2.0		.05	0.000	0.995	1.000	1.000	0.964	0.990	1.000	1.000
γ=0.0		.01	0.000	0.983						

TABLE 8.1 continued

True Model	False Model	Significance Level For BLR Test	BLR Test Size	BLR Test Power	Tests and Criteria AIC	SIC	CHI	KS	UPE	FPE
G3	W3	.90	0.152	0.662						
α=6.0		.95	0.109	0.581	0.766	0.766	0.966	1.000	0.521	0.807
γ=1.0		.99	0.025	0.380						
G3	LN3	.90	0.038	0.270						
α=6.0		.95	0.008	0.105	0.760	0.760	0.966	1.000	0.822	0.657
γ=1.0		.99	0.000	0.001						
G3	W2	.90	0.000	1.000						
α=6.0		.95	0.000	1.000	1.000	1.000	0.966	1.000	0.757	1.000
γ=1.0		.99	0.000	1.000						
G3	LN2	.90	0.016	0.735						
α=6.0		.95	0.003	0.555	0.665	0.160	0.966	1.000	0.917	0.939
γ=1.0		.99	0.000	0.187						
G2	W3	.90	0.217	0.590						
α=6.0		.95	0.157	0.508	0.805	0.959	0.966	0.990	0.524	0.669
γ=0.0		.99	0.053	0.334						
G2	LN3	.90	0.180	0.139						
α=6.0		.95	0.083	0.034	0.860	0.989	0.966	0.990	0.667	0.442
γ=0.0		.99	0.016	0.000						
G2	W2	.90	0.004	0.988						
α=6.0		.95	0.003	0.982	0.991	0.991	0.966	0.990	0.704	0.973
γ=0.0		.99	0.000	0.957						
G2	LN2	.90	0.041	0.892						
α=6.0		.95	0.028	0.861	0.934	0.934	0.966	0.990	0.986	1.000
γ=0.0		.99	0.007	0.756						

Note: α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the three- and two-parameter gamma distributions, W3 and W2 the three- and two-parameter Weibull distributions, and LN3 and LN2 the three- and two-parameter lognormal distributions.

TABLE 8.2

Probabilities of accepting the true model and rejecting the false model using three tests, two discrimination criteria and two performance criteria over 1000 replications of random samples of size $n = 365$ ($\beta = 1$)

True Model	False Model	Confidence Level For BLR Test	BLR Test Size	BLR Test Power	Tests and Criteria AIC	SIC	CHI	KS	UPE	FPE
W3	G3	.90	0.031	0.856						
α=2.0		.95	0.017	0.805	0.924	0.924	0.959	1.000	0.936	0.918
γ=1.0		.99	0.005	0.529						
W3	LN3	.90	0.018	0.940						
α=2.0		.95	0.011	0.910	0.968	0.968	0.959	1.000	0.953	0.949
γ=1.0		.99	0.004	0.791						
W3	G2	.90	0.000	1.000						
α=2.0		.95	0.000	0.998	0.999	0.982	0.959	1.000	0.476	0.999
γ=1.0		.99	0.000	0.987						
W3	LN2	.90	0.006	0.969						
α=2.0		.95	0.002	0.961	0.968	0.845	0.959	1.000	0.879	0.953
γ=1.0		.99	0.000	0.879						
W2	G3	.90	0.066	0.790						
α=2.0		.95	0.044	0.716	0.941	0.991	0.973	0.999	0.932	0.722
γ=0.0		.99	0.011	0.412						
W2	LN3	.90	0.031	0.909						
α=2.0		.95	0.016	0.870	0.975	0.996	0.973	0.999	0.952	0.886
γ=0.0		.99	0.004	0.724						
W2	G2	.90	0.022	0.941						
α=2.0		.95	0.016	0.920	0.966	0.966	0.973	0.999	0.987	0.983
γ=0.0		.99	0.005	0.853						
W2	LN2	.90	0.000	1.000						
α=2.0		.95	0.000	1.000	1.000	1.000	0.973	0.999	1.000	1.000
γ=0.0		.99	0.000	1.000						

TABLE 8.2 continued

True Model	False Model	Confidence Level For BLR Test	BLR Test Size	BLR Test Power	Tests and Criteria AIC	SIC	CHI	KS	UPE	FPE
W3	G3	.90	0.020	0.919						
α=4.0		.95	0.012	0.895	0.957	0.957	0.967	1.000	0.980	0.957
γ=1.0		.99	0.002	0.861						
W3	LN3	.90	0.041	0.744						
α=4.0		.95	0.019	0.611	0.890	0.890	0.967	1.000	0.921	0.845
γ=1.0		.99	0.002	0.209						
W3	G2	.90	0.008	0.951						
α=4.0		.95	0.004	0.934	0.945	0.789	0.967	1.000	0.987	0.994
γ=1.0		.99	0.000	0.817						
W3	LN2	.90	0.001	0.995						
α=4.0		.95	0.000	0.989	0.993	0.948	0.967	1.000	0.999	1.000
γ=1.0		.99	0.000	0.959						
W2	G3	.90	0.055	0.870						
α=4.0		.95	0.041	0.860	0.948	0.993	0.973	0.999	0.974	0.925
γ=0.0		.99	0.008	0.853						
W2	LN3	.90	0.090	0.634						
α=4.0		.95	0.056	0.488	0.929	0.993	0.973	0.999	0.908	0.558
γ=0.0		.99	0.009	0.129						
W2	G2	.90	0.000	0.996						
α=4.0		.95	0.000	0.995	0.998	0.998	0.973	0.999	1.000	0.997
γ=0.0		.99	0.000	0.991						
W2	LN2	.90	0.000	1.000						
α=4.0		.95	0.000	1.000	1.000	1.000	0.973	0.999	1.000	1.000
γ=0.0		.99	0.000	1.000						

Note: α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the three- and two-parameter gamma distributions, W3 and W2 the three- and two-parameter Weibull distributions, and LN3 and LN2 the three- and two-parameter lognormal distributions.

TABLE 8.3

Probabilities of accepting the true model and rejecting the false model using three tests, two discrimination criteria and two performance criteria over 1000 replications of random samples of size $n = 365$ ($\beta = 1$)

True Model	False Model	Confidence Level For BLR Test	BLR Test Size	BLR Test Power	Tests and Criteria AIC	SIC	CHI	KS	UPE	FPE
LN3	G3	.90	0.022	0.951						
α=0.9		.95	0.016	0.946	0.966	0.966	0.954	1.000	0.684	0.930
γ=1.0		.99	0.006	0.919						
LN3	W3	.90	0.014	0.973						
α=0.9		.95	0.009	0.965	0.981	0.981	0.954	1.000	0.650	0.939
γ=1.0		.99	0.004	0.941						
LN3	G2	.90	0.000	1.000						
α=0.9		.95	0.000	1.000	1.000	1.000	0.954	1.000	0.830	1.000
γ=1.0		.99	0.000	1.000						
LN3	W2	.90	0.000	1.000						
α=0.9		.95	0.000	1.000	1.000	1.000	0.954	1.000	0.787	1.000
γ=1.0		.99	0.000	1.000						
LN2	G3	.90	0.028	0.944						
α=0.9		.95	0.022	0.935	0.975	0.991	0.964	0.990	0.692	0.942
γ=0.0		.99	0.010	0.909						
LN2	W3	.90	0.020	0.969						
α=0.9		.95	0.016	0.958	0.980	0.996	0.964	0.990	0.659	0.978
γ=0.0		.99	0.004	0.935						
LN2	G2	.90	0.003	0.996						
α=0.9		.95	0.003	0.996	0.997	0.997	0.964	0.990	0.747	0.991
γ=0.0		.99	0.000	0.995						
LN2	W2	.90	0.000	1.000						
α=0.9		.95	0.000	1.000	1.000	1.000	0.964	0.990	0.715	0.998
γ=0.0		.99	0.000	0.999						

TABLE 8.3 continued

True Model	False Model	Confidence Level For BLR Test	BLR Test Size	BLR Test Power	Tests and Criteria AIC	SIC	CHI	KS	UPE	FPE
LN3	G3	.90	0.174	0.598						
α=0.5		.95	0.102	0.515	0.718	0.718	0.954	1.000	0.515	0.678
γ=1.0		.99	0.019	0.316						
LN3	W3	.90	0.050	0.896						
α=0.5		.95	0.032	0.875	0.931	0.931	0.954	1.000	0.649	0.910
γ=1.0		.99	0.014	0.809						
LN3	G2	.90	0.000	1.000						
α=0.5		.95	0.000	1.000	1.000	0.998	0.954	1.000	0.769	1.000
γ=1.0		.99	0.000	0.998						
LN3	W2	.90	0.000	1.000						
α=0.5		.95	0.000	1.000	1.000	1.000	0.954	1.000	0845	1.000
γ=1.0		.99	0.000	1.000						
LN2	G3	.90	0.226	0.522						
α=0.5		.95	0.169	0.437	0.793	0.968	0.964	0.990	0.534	0.608
γ=0.0		.99	0.045	0.260						
LN2	W3	.90	0.067	0.876						
α=0.5		.95	0.048	0.844	0.937	0.980	0.964	0.990	0.646	0.913
γ=0.0		.99	0.023	0.783						
LN2	G2	.90	0.016	0.951						
α=0.5		.95	0.008	0.939	0.973	0.973	0.964	0.990	0.750	0.932
γ=0.0		.99	0.007	0.873						
LN2	W2	.90	0.000	1.000						
α=0.5		.95	0.000	1.000	1.000	1.000	0.964	0.990	0.756	1.000
γ=0.0		.99	0.000	0.999						

Note: α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the three- and two-parameter gamma distributions, W3 and W2 the three- and two-parameter Weibull distributions, and LN3 and LN2 the three- and two-parameter lognormal distributions.

TABLE 8.4

Probabilities of accepting the estimated model for two tests, three discrimination criteria and two performance criteria for three non-nested distributions over 1000 replications of random samples of size $n = 365$ ($\beta = 1$)

True Model	Estimated Model	Tests and Criteria KL	AIC	SIC	CHI	KS	UPE	FPE
G3	G3	0.596	0.596	0.596	0.960	1.000	0.374	0.572
α=2.0	W3	0.355	0.355	0.355	0.916	0.999	0.576	0.365
γ=1.0	LN3	0.049	0.049	0.049	0.934	0.999	0.050	0.063
G3	G2	0.001	0.001	0.001	0.545	0.781	0.317	0.001
α=2.0	W2	0.000	0.000	0.000	0.003	0.074	0.022	0.000
γ=1.0	LN2	0.999	0.999	0.999	0.923	0.985	0.661	0.999
G2	G3	0.597	0.597	0.597	0.958	1.000	0.368	0.344
α=2.0	W3	0.354	0.354	0.354	0.917	0.999	0.587	0.483
γ=0.0	LN3	0.049	0.049	0.049	0.934	0.999	0.045	0.173
G2	G2	0.864	0.864	0.864	0.964	0.990	0.508	0.618
α=2.0	W2	0.136	0.136	0.136	0.915	0.996	0.492	0.382
γ=0.0	LN2	0.000	0.000	0.000	0.396	0.654	0.000	0.000
G3	G3	0.526	0.526	0.526	0.966	1.000	0.348	0.464
α=6.0	W3	0.234	0.234	0.234	0.865	0.999	0.479	0.193
γ=1.0	LN3	0.240	0.240	0.240	0.969	1.000	0.173	0.343
G3	G2	0.567	0.567	0.567	0.962	0.989	0.696	0.757
α=6.0	W2	0.000	0.000	0.000	0.180	0.736	0.147	0.000
γ=1.0	LN2	0.433	0.433	0.433	0.948	0.989	0.157	0.243
G2	G3	0.526	0.526	0.526	0.966	1.000	0.341	0.323
α=6.0	W3	0.234	0.234	0.234	0.865	0.999	0.485	0.185
γ=0.0	LN3	0.240	0.240	0.240	0.969	1.000	0.174	0.492
G2	G2	0.925	0.925	0.925	0.966	0.990	0.683	0.949
α=6.0	W2	0.009	0.009	0.009	0.516	0.910	0.296	0.027
γ=0.0	LN2	0.066	0.066	0.066	0.834	0.967	0.021	0.024

TABLE 8.4 continued

True Model	Estimated Model	Tests and Criteria KL	AIC	SIC	CHI	KS	UPE	FPE
W3	G3	0.073	0.073	0.073	0.923	0.999	0.050	0.071
α=2.0	W3	0.924	0.924	0.924	0.959	1.000	0.936	0.918
γ=1.0	LN3	0.003	0.003	0.003	0.929	1.000	0.014	0.011
W3	G2	0.081	0.081	0.081	0.945	0.986	0.615	0.151
α=2.0	W2	0.000	0.000	0.000	0.038	0.490	0.194	0.000
γ=1.0	LN2	0.919	0.919	0.919	0.949	0.990	0.191	0.849
W2	G3	0.074	0.074	0.074	0.923	0.999	0.044	0.282
α=2.0	W3	0.923	0.923	0.923	0.958	1.000	0.944	0.660
γ=0.0	LN3	0.003	0.003	0.003	0.929	1.000	0.012	0.058
W2	G2	0.034	0.034	0.034	0.807	0.928	0.013	0.017
α=2.0	W2	0.966	0.966	0.966	0.973	0.999	0.987	0.983
γ=0.0	LN2	0.000	0.000	0.000	0.027	0.142	0.000	0.000
W3	G3	0.029	0.029	0.029	0.087	0.129	0.004	0.008
α=4.0	W3	0.890	0.890	0.890	0.967	1.000	0.921	0.845
γ=1.0	LN3	0.081	0.081	0.081	0.936	0.993	0.075	0.147
W3	G2	0.656	0.656	0.656	0.851	0.968	0.064	0.821
α=4.0	W2	0.344	0.344	0.344	0.813	0.973	0.936	0.179
γ=1.0	LN2	0.000	0.000	0.000	0.713	0.906	0.000	0.000
W2	G3	0.029	0.029	0.029	0.086	0.119	0.004	0.022
α=4.0	W3	0.890	0.890	0.890	0.967	1.000	0.924	0.720
γ=0.0	LN3	0.081	0.081	0.081	0.935	0.993	0.072	0.258
W2	G2	0.002	0.002	0.002	0.384	0.629	0.000	0.003
α=4.0	W2	0.998	0.998	0.998	0.973	0.999	1.000	0.997
γ=0.0	LN2	0.000	0.000	0.000	0.027	0.142	0.000	0.000

TABLE 8.4 continued

True Model	Estimated Model	Tests and Criteria KL	AIC	SIC	CHI	KS	UPE	FPE
LN3	G3	0.031	0.031	0.031	0.311	0.604	0.055	0.056
α=0.9	W3	0.003	0.003	0.003	0.133	0.705	0.296	0.017
γ=1.0	LN3	0.966	0.966	0.966	0.954	1.000	0.649	0.927
LN3	G2	0.000	0.000	0.000	0.000	0.001	0.000	0.000
α=0.9	W2	0.000	0.000	0.000	0.000	0.000	0.045	0.000
γ=1.0	LN2	1.000	1.000	1.000	0.391	0.713	0.955	1.000
LN2	G3	0.030	0.030	0.030	0.311	0.606	0.059	0.035
α=0.9	W3	0.003	0.003	0.003	0.124	0.696	0.294	0.000
γ=0.0	LN3	0.967	0.967	0.967	0.954	1.000	0.647	0.965
LN2	G2	0.003	0.003	0.003	0.177	0.256	0.038	0.009
α=0.9	W2	0.000	0.000	0.000	0.022	0.295	0.247	0.000
γ=0.0	LN2	0.997	0.997	0.997	0.964	0.990	0.715	0.991
LN3	G3	0.269	0.269	0.269	0.918	0.996	0.387	0.310
α=0.5	W3	0.013	0.013	0.013	0.471	0.924	0.098	0.012
γ=1.0	LN3	0.718	0.718	0.718	0.954	1.000	0.515	0.678
LN3	G2	0.000	0.000	0.000	0.347	0.606	0.038	0.000
α=0.5	W2	0.000	0.000	0.000	0.000	0.003	0.007	0.000
γ=1.0	LN2	1.000	1.000	1.000	0.874	0.965	0.955	1.000
LN2	G3	0.269	0.269	0.269	0.917	0.996	0.388	0.275
α=0.5	W3	0.013	0.013	0.013	0.476	0.924	0.100	0.007
γ=0.0	LN3	0.718	0.718	0.718	0.954	1.000	0.512	0.718
LN2	G2	0.027	0.027	0.027	0.790	0.915	0.323	0.095
α=0.5	W2	0.000	0.000	0.000	0.022	0.295	0.013	0.000
γ=0.0	LN2	0.973	0.973	0.973	0.964	0.990	0.664	0.905

Note: α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the three- and two-parameter gamma distributions, W3 and W2 the three- and two-parameter Weibull distributions, and LN3 and LN2 the three- and two-parameter lognormal distributions.

TABLE 8.5

Probabilities of accepting the estimated model from three non-nested distributions using GIC over 1000 replications of random samples of size $n = 365$ ($\beta = 1$)

True Model	Estimated Model G3	W3	LN3	Estimated Model G2	W2	LN2
G3 α=2.0 γ=1.0	0.943	0.795	0.240	0.003	0.000	1.000
G2 α=2.0 γ=0.0	0.943	0.794	0.242	0.977	0.450	0.011
G3 α=6.0 γ=1.0	0.947	0.527	0.860	0.846	0.000	0.774
G2 α=6.0 γ=0.0	0.947	0.527	0.860	0.986	0.030	0.197
W3 α=2.0 γ=1.0	0.342	0.992	0.136	0.538	0.000	0.998
W2 α=2.0 γ=0.0	0.338	0.992	0.137	0.115	0.991	0.000
W3 α=4.0 γ=1.0	0.128	0.989	0.644	0.781	0.490	0.018
W2 α=4.0 γ=0.0	0.128	0.989	0.644	0.007	1.000	0.000
LN3 α=0.9 γ=1.0	0.069	0.042	0.991	0.000	0.000	1.000
LN2 α=0.9 γ=0.0	0.069	0.042	0.991	0.005	0.000	1.000
LN3 α=0.5 γ=1.0	0.605	0.096	0.961	0.000	0.000	1.000
LN2 α=0.5 γ=0.0	0.605	0.096	0.961	0.091	0.000	0.992

Note: α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the three- and two-parameter gamma distributions, W3 and W2 the three- and two-parameter Weibull distributions, and LN3 and LN2 the three- and two-parameter lognormal distributions.

TABLE 8.6

Probabilities of accepting the estimated model for two tests, three discrimination criteria and two performance criteria for five non-nested distributions over 1000 replications of random samples of size $n = 365$ ($\beta = 1$)

True Model	Estimated Model	Tests and Criteria KL	AIC	SIC	CHI	KS	UPE	FPE
G3	G3	0.596	0.596	0.584	0.960	1.000	0.238	0.569
α=2.0	G2	-	-	-	-	-	-	-
γ=1.0	W3	0.355	0.355	0.355	0.916	0.999	0.467	0.365
	W2	0.000	0.000	0.000	0.003	0.074	0.061	0.000
	LN3	0.049	0.049	0.046	0.934	0.999	0.047	0.061
	LN2	0.000	0.000	0.015	0.923	0.985	0.187	0.005
G2	G3	-	-	-	-	-	-	-
α=2.0	G2	0.463	0.663	0.835	0.964	0.990	0.392	0.394
γ=0.0	W3	0.445	0.194	0.027	0.917	0.999	0.274	0.182
	W2	0.014	0.100	0.134	0.915	0.996	0.284	0.287
	LN3	0.078	0.043	0.004	0.934	0.999	0.050	0.137
	LN2	0.000	0.000	0.000	0.396	0.654	0.000	0.000
G3	G3	0.526	0.362	0.064	0.966	1.000	0.341	0.460
α=6.0	G2	-	-	-	-	-	-	-
γ=1.0	W3	0.234	0.234	0.170	0.865	0.999	0.424	0.193
	W2	0.000	0.000	0.000	0.180	0.736	0.055	0.000
	LN3	0.240	0.105	0.021	0.969	1.000	0.135	0.314
	LN2	0.000	0.299	0.745	0.948	0.989	0.045	0.033
G2	G3	-	-	-	-	-	-	-
α=6.0	G2	0.317	0.682	0.893	0.966	0.990	0.319	0.298
γ=0.0	W3	0.280	0.185	0.036	0.865	0.999	0.381	0.210
	W2	0.000	0.000	0.007	0.516	0.910	0.083	0.001
	LN3	0.403	0.091	0.003	0.969	1.000	0.211	0.490
	LN2	0.000	0.042	0.061	0.834	0.967	0.006	0.001

TABLE 8.6 continued

True Model	Estimated Model	Tests and Criteria KL	AIC	SIC	CHI	KS	UPE	FPE
W3	G3	0.073	0.064	0.025	0.923	0.999	0.028	0.064
α=2.0	G2	0.000	0.000	0.004	0.945	0.986	0.516	0.001
γ=1.0	W3	0.924	0.923	0.822	0.959	1.000	0.361	0.918
	W2	-	-	-	-	-	-	-
	LN3	0.003	0.003	0.000	0.929	1.000	0.010	0.009
	LN2	0.000	0.010	0.149	0.949	0.990	0.085	0.008
W2	G3	0.110	0.034	0.000	0.923	0.999	0.053	0.233
α=2.0	G2	0.002	0.027	0.034	0.807	0.928	0.001	0.003
γ=0.0	W3	-	-	-	-	-	-	-
	W2	0.885	0.936	0.964	0.973	0.999	0.929	0.707
	LN3	0.003	0.003	0.002	0.929	1.000	0.017	0.057
	LN2	0.000	0.000	0.000	0.027	0.142	0.000	0.000
W3	G3	0.029	0.014	0.000	0.087	0.129	0.004	0.007
α=4.0	G2	0.000	0.029	0.198	0.851	0.968	0.003	0.000
γ=1.0	W3	0.890	0.880	0.769	0.967	1.000	0.921	0.845
	W2	-	-	-	-	-	-	-
	LN3	0.081	0.077	0.033	0.936	0.993	0.072	0.148
	LN2	0.000	0.000	0.000	0.713	0.906	0.000	0.000
W2	G3	0.073	0.038	0.002	0.086	0.119	0.006	0.032
α=4.0	G2	0.000	0.000	0.001	0.384	0.629	0.000	0.000
γ=0.0	W3	-	-	-	-	-	-	-
	W2	0.810	0.927	0.992	0.973	0.999	0.908	0.557
	LN3	0.117	0.035	0.005	0.935	0.993	0.086	0.411
	LN2	0.000	0.000	0.000	0.027	0.142	0.000	0.000

TABLE 8.6 continued

True Model	Estimated Model	Tests and Criteria KL	AIC	SIC	CHI	KS	UPE	FPE
LN3	G3	0.031	0.031	0.031	0.311	0.604	0.055	0.056
α=0.9	G2	0.000	0.000	0.000	0.000	0.001	0.000	0.000
γ=1.0	W3	0.003	0.003	0.003	0.133	0.705	0.296	0.017
	W2	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	LN3	0.966	0.966	0.966	0.954	1.000	0.649	0.927
	LN2	-	-	-	-	-	-	-
LN2	G3	0.036	0.022	0.007	0.311	0.606	0.060	0.058
α=0.9	G2	0.000	0.000	0.001	0.177	0.256	0.001	0.000
γ=0.0	W3	0.005	0.004	0.001	0.124	0.696	0.279	0.000
	W2	0.000	0.000	0.000	0.022	0.295	0.001	0.000
	LN3	-	-	-	-	-	-	-
	LN2	0.959	0.974	0.991	0.964	0.990	0.659	0.942
LN3	G3	0.269	0.269	0.269	0.918	0.996	0.387	0.310
α=0.5	G2	0.000	0.000	0.000	0.347	0.606	0.015	0.000
γ=1.0	W3	0.013	0.013	0.013	0.471	0.924	0.083	0.012
	W2	0.000	0.000	0.000	0.000	0.003	0.000	0.000
	LN3	0.718	0.718	0.718	0.954	1.000	0.515	0.678
	LN2	-	-	-	-	-	-	-
LN2	G3	0.343	0.186	0.013	0.917	0.996	0.357	0.373
α=0.5	G2	0.000	0.008	0.026	0.790	0.915	0.034	0.013
γ=0.0	W3	0.013	0.013	0.010	0.476	0.924	0.068	0.006
	W2	0.000	0.000	0.000	0.022	0.295	0.007	0.000
	LN3	-	-	-	-	-	-	-
	LN2	0.644	0.793	0.951	0.964	0.990	0.534	0.608

Note: α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the three- and two-parameter gamma distributions, W3 and W2 the three- and two-parameter Weibull distributions, and LN3 and LN2 the three- and two-parameter lognormal distributions.

TABLE 8.7

Probabilities of accepting the estimated model from five non-nested distributions using GIC over 1000 replications of random samples of size $n = 365$ ($\beta = 1$)

True Model	Estimated Model G3	G2	W3	W2	LN3	LN2
G3 α=2.0 γ=1.0	0.943	-	0.795	0.000	0.240	0.004
G2 α=2.0 γ=0.0	-	0.871	0.831	0.411	0.294	0.002
G3 α=6.0 γ=1.0	0.947	-	0.527	0.000	0.860	0.639
G2 α=6.0 γ=0.0	-	0.881	0.554	0.008	0.862	0.170
W3 α=2.0 γ=1.0	0.342	0.004	0.992	-	0.136	0.060
W2 α=2.0 γ=0.0	0.440	0.111	-	0.980	0.186	0.000
W3 α=4.0 γ=1.0	0.128	0.097	0.989	-	0.644	0.017
W2 α=4.0 γ=0.0	0.146	0.001	-	0.970	0.735	0.000
LN3 α=0.9 γ=1.0	0.069	0.000	0.042	0.000	0.991	-
LN2 α=0.9 γ=0.0	0.080	0.000	0.048	0.000	-	0.982
LN3 α=0.5 γ=1.0	0.605	0.000	0.096	0.000	0.961	-
LN2 α=0.5 γ=0.0	0.669	0.065	0.096	0.000	-	0.913

Note: α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the three- and two-parameter gamma distributions, W3 and W2 the three- and two-parameter Weibull distributions, and LN3 and LN2 the three- and two-parameter lognormal distributions.

TABLE 8.8

Probabilities of accepting the estimated model for two tests, three discrimination criteria and two performance criteria for one nested and five non-nested distributions over 1000 replications of random samples of size $n = 365$ ($\beta = 1$)

True Model	Estimated Model	Tests and Criteria KL	AIC	SIC	CHI	KS	UPE	FPE
G3	G3	0.596	0.596	0.584	0.960	1.000	0.238	0.569
α=2.0	G2	0.000	0.000	0.000	0.545	0.781	0.130	0.000
γ=1.0	W3	0.355	0.355	0.355	0.916	0.999	0.376	0.365
	W2	0.000	0.000	0.000	0.003	0.074	0.022	0.000
	LN3	0.049	0.049	0.046	0.934	0.999	0.047	0.061
	LN2	0.000	0.000	0.015	0.923	0.985	0.187	0.005
G2	G3	0.520	0.092	0.010	0.958	1.000	0.243	0.089
α=2.0	G2	0.078	0.607	0.827	0.964	0.990	0.174	0.359
γ=0.0	W3	0.343	0.169	0.025	0.917	0.999	0.258	0.150
	W2	0.010	0.096	0.134	0.915	0.996	0.283	0.278
	LN3	0.049	0.036	0.004	0.934	0.999	0.042	0.124
	LN2	0.000	0.000	0.000	0.396	0.654	0.000	0.000
G3	G3	0.526	0.106	0.001	0.966	1.000	0.223	0.328
α=6.0	G2	0.000	0.383	0.528	0.962	0.989	0.250	0.216
γ=1.0	W3	0.234	0.192	0.059	0.865	0.999	0.300	0.191
	W2	0.000	0.000	0.000	0.180	0.736	0.055	0.000
	LN3	0.240	0.025	0.000	0.969	1.000	0.127	0.232
	LN2	0.000	0.294	0.412	0.948	0.989	0.045	0.033
G2	G3	0.521	0.075	0.001	0.966	1.000	0.227	0.172
α=6.0	G2	0.005	0.645	0.892	0.966	0.990	0.181	0.252
γ=0.0	W3	0.234	0.172	0.036	0.865	0.999	0.351	0.171
	W2	0.000	0.000	0.007	0.516	0.910	0.083	0.001
	LN3	0.240	0.078	0.003	0.969	1.000	0.152	0.404
	LN2	0.000	0.030	0.061	0.834	0.967	0.006	0.000

TABLE 8.8 continued

True Model	Estimated Model	Tests and Criteria KL	AIC	SIC	CHI	KS	UPE	FPE
W3	G3	0.073	0.064	0.025	0.923	0.999	0.028	0.064
α=2.0	G2	0.000	0.000	0.004	0.945	0.986	0.396	0.001
γ=1.0	W3	0.924	0.923	0.822	0.959	1.000	0.299	0.918
	W2	0.000	0.000	0.000	0.038	0.490	0.182	0.000
	LN3	0.003	0.003	0.000	0.929	1.000	0.010	0.009
	LN2	0.000	0.010	0.149	0.949	0.990	0.085	0.008
W2	G3	0.071	0.030	0.000	0.923	0.999	0.041	0.175
α=2.0	G2	0.001	0.017	0.034	0.807	0.928	0.001	0.001
γ=0.0	W3	0.788	0.117	0.004	0.958	1.000	0.436	0.280
	W2	0.137	0.833	0.960	0.973	0.999	0.510	0.487
	LN3	0.003	0.003	0.002	0.929	1.000	0.012	0.057
	LN2	0.000	0.000	0.000	0.027	0.142	0.000	0.000
W3	G3	0.029	0.014	0.000	0.087	0.129	0.004	0.008
α=4.0	G2	0.000	0.029	0.198	0.851	0.968	0.003	0.000
γ=1.0	W3	0.890	0.874	0.707	0.967	1.000	0.539	0.845
	W2	0.000	0.006	0.062	0.813	0.973	0.382	0.000
	LN3	0.081	0.077	0.033	0.936	0.993	0.072	0.147
	LN2	0.000	0.000	0.000	0.713	0.906	0.000	0.000
W2	G3	0.029	0.022	0.001	0.086	0.119	0.004	0.018
α=4.0	G2	0.000	0.000	0.000	0.384	0.629	0.000	0.000
γ=0.0	W3	0.829	0.129	0.017	0.967	1.000	0.499	0.473
	W2	0.061	0.818	0.977	0.973	0.999	0.427	0.315
	LN3	0.081	0.031	0.005	0.935	0.993	0.070	0.194
	LN2	0.000	0.000	0.000	0.027	0.142	0.000	0.000

TABLE 8.8 continued

True Model	Estimated Model	Tests and Criteria KL	AIC	SIC	CHI	KS	UPE	FPE
LN3	G3	0.031	0.031	0.031	0.311	0.604	0.013	0.056
α=0.9	G2	0.000	0.000	0.000	0.000	0.001	0.000	0.000
γ=1.0	W3	0.003	0.003	0.003	0.133	0.705	0.150	0.017
	W2	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	LN3	0.966	0.966	0.966	0.954	1.000	0.627	0.927
	LN2	0.000	0.000	0.000	0.391	0.713	0.210	0.000
LN2	G3	0.030	0.020	0.007	0.311	0.606	0.056	0.029
α=0.9	G2	0.000	0.000	0.001	0.177	0.256	0.001	0.000
γ=0.0	W3	0.003	0.003	0.001	0.124	0.696	0.249	0.000
	W2	0.000	0.000	0.000	0.022	0.295	0.001	0.000
	LN3	0.967	0.147	0.011	0.954	1.000	0.348	0.514
	LN2	0.000	0.830	0.980	0.964	0.990	0.345	0.457
LN3	G3	0.269	0.268	0.235	0.918	0.996	0.304	0.310
α=0.5	G2	0.000	0.000	0.000	0.347	0.606	0.014	0.000
γ=1.0	W3	0.013	0.013	0.011	0.471	0.924	0.068	0.012
	W2	0.000	0.000	0.000	0.000	0.003	0.000	0.000
	LN3	0.718	0.716	0.625	0.954	1.000	0.513	0.677
	LN2	0.000	0.003	0.129	0.874	0.965	0.101	0.001
LN2	G3	0.269	0.173	0.013	0.917	0.996	0.322	0.253
α=0.5	G2	0.000	0.006	0.026	0.790	0.915	0.034	0.005
γ=0.0	W3	0.013	0.013	0.010	0.476	0.924	0.068	0.006
	W2	0.000	0.000	0.000	0.022	0.295	0.007	0.000
	LN3	0.718	0.107	0.010	0.954	1.000	0.297	0.430
	LN2	0.000	0.701	0.941	0.964	0.990	0.272	0.306

Note: α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the three- and two-parameter gamma distributions, W3 and W2 the three- and two-parameter Weibull distributions, and LN3 and LN2 the three- and two-parameter lognormal distributions.

TABLE 8.9

Probabilities of accepting the estimated model from one nested and five non-nested distributions using GIC over 1000 replications of random samples of size $n = 365$ ($\beta = 1$)

True Model	Estimated Model G3	G2	W3	W2	LN3	LN2
G3 α=2.0 γ=1.0	0.943	0.000	0.795	0.000	0.240	0.004
G2 α=2.0 γ=0.0	0.943	0.867	0.794	0.390	0.241	0.001
G3 α=6.0 γ=1.0	0.947	0.721	0.527	0.000	0.860	0.639
G2 α=6.0 γ=0.0	0.947	0.882	0.527	0.006	0.860	0.116
W3 α=2.0 γ=1.0	0.342	0.004	0.992	0.000	0.136	0.060
W2 α=2.0 γ=0.0	0.338	0.062	0.992	0.970	0.137	0.000
W3 α=4.0 γ=1.0	0.128	0.097	0.989	0.048	0.644	0.017
W2 α=4.0 γ=0.0	0.128	0.000	0.989	0.956	0.644	0.000
LN3 α=0.9 γ=1.0	0.069	0.000	0.042	0.000	0.991	0.000
LN2 α=0.9 γ=0.0	0.069	0.000	0.042	0.000	0.991	0.965
LN3 α=0.5 γ=1.0	0.605	0.000	0.096	0.000	0.961	0.093
LN2 α=0.5 γ=0.0	0.605	0.040	0.096	0.000	0.961	0.904

Note: α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the three- and two-parameter gamma distributions, W3 and W2 the three- and two-parameter Weibull distributions, and LN3 and LN2 the three- and two-parameter lognormal distributions.

Chapter 9

Estimating the Percentiles of Some Misspecified Non-nested Distributions

9.1 Introduction

Three distributions which have been used extensively in testing and modelling natural phenomena in areas such as reliability and life testing (Mann et al. (1974) and Bain (1978)), hydrology (Stedinger (1980)), and air quality management (Jakeman and Taylor (1989) and Jakeman et al. (1986)) are the gamma, Weibull and lognormal distributions. Two- and three-parameter versions of these distributions have been used because they are parsimonious in considering the shape, scale and location of the distribution, but still sufficiently flexible in fitting real data. Occasionally, there may be some prior information regarding the form of the distribution and parsimony considerations might yield a preference for the two-parameter variant. However, an estimated two- or three-parameter distribution might be inadequate empirically if the distribution has been specified incorrectly. In practice, it is not known which of the three distributions is appropriate and the conventional practice of selecting a particular distribution for convenience or on the basis of previous studies pays little or no attention to the consequences of misspecifying the distribution. While an examination of the consequences of underfitting or overfitting a particular distribution is useful within the context of having selected the correct distribution, it has no bearing on the pos-

sible effects of misspecifying the distribution altogether in the presence of non-nested alternative distributions.

In considering whether a particular specification should be used to estimate the percentiles of a distribution, it is important to take account of the sorts of errors that might be made in fitting the distribution. Specifically, the consequences of misspecifying the distribution should be evaluated. Such misspecifications arise when a particular distribution is estimated but a non-nested alternative distribution is correct. The purpose of this chapter is to assess the consequences of such misspecification in estimating the upper percentiles of the two- and three-parameter non-nested gamma, Weibull and lognormal distributions.

For a given problem, a typical statistical decision is to simply estimate the two- or three-parameter variant of a particular distribution without examining the possibility that the estimated distribution is misspecified. The primary aim of the experiments in this chapter is to observe the magnitudes of the errors obtained by fitting an incorrect non-nested distribution relative to fitting the correct distribution. The experiments are conducted for three different distributions and different parameter sets, especially for different values of the shape parameter.

The plan of the chapter is as follows. In Section 2 the distribution functions and performance criteria are presented. The Monte Carlo experiments are outlined in Section 3 and the results are discussed in Section 4. Some concluding remarks are given in Section 5.

9.2 Distribution Functions and Performance Criteria

For a sample $x_1, x_2, \ldots, x_n$ of n independently and identically distributed random observations, the log-likelihood functions for the three-parameter gamma, Weibull and lognormal distributions are given in equations (7.1)-(7.3) and are reproduced here for convenience:

Gamma:

$$\log L = -n\alpha \log \beta - n \log \Gamma(\alpha) + (\alpha - 1) \sum_{i=1}^{n} \log(x_i - \gamma) - \sum_{i=1}^{n} (\frac{x_i - \gamma}{\beta}) \qquad (9.1)$$

Weibull:

$$\log L = n \log \alpha - n\alpha \log \beta + (\alpha - 1) \sum_{i=1}^{n} \log(x_i - \gamma) - \sum_{i=1}^{n} (\frac{x_i - \gamma}{\beta})^{\alpha} \qquad (9.2)$$

Lognormal:

$$\log L = -\frac{n}{2} \log(2\pi\alpha^2) + \sum_{i=1}^{n} \log(x_i - \gamma) - \frac{1}{2\alpha^2} \sum_{i=1}^{n} [\log(x_i - \gamma) - \beta]^2 \qquad (9.3)$$

in which β represents the scale parameter, α the shape parameter, γ the location parameter, and Γ the gamma function. The two- parameter versions of the above functions are obtained by setting the location parameter γ to zero in each case. In the above equations, $\beta > 0$, $\alpha > 0$ and $\gamma < x_i < \infty$ for $i = 1, 2, \ldots, n$. The density functions of the gamma and Weibull functions approach the exponential at $\alpha = 1$, are "J" shaped for $\alpha < 1$ and "bell" shaped for $\alpha > 1$, whereas the density for the lognormal function changes from being nearly symmetric to being heavily skewed as α is increased from 0.4 to 0.9 to 1.2. These values accommodate a variety of shapes which arise in practice in analysing real data.

The parameters of the three log-likelihood functions are estimated by maximum likelihood methods. Since the general maximum likelihood procedure will frequently fail to converge when the (unknown) shape parameter is less than or equal to unity, a computationally efficient approach that circumvents this problem is used (for further details, see Bai et al. (1989)). For purposes of evaluating the performance of the misspecified two- and three-parameter versions of the estimated distributions, loss functions recommended for assessing air quality models are used (see Fox (1981)). These functions are the relative bias (BIAS) and the relative root mean square error (RRMSE) which are evaluated at the upper percentiles of the distribution. For an estimate $\hat{q}_i$ of a quantity of interest q, the performance criteria are defined in terms of deviations from q in each replication of the simulation experiments. The definitions are given in equations (3.18) and (3.19), namely

$$BIAS(q) = \frac{1}{N}\sum_{i=1}^{N}(\frac{\hat{q}_i - q}{q}) \tag{9.4}$$

$$RRMSE(q) = [\frac{1}{N}\sum_{i=1}^{N}(\frac{\hat{q}_i - q}{q})^2]^{0.5} \tag{9.5}$$

where N is the number of replications of the experiment. For present purposes, q denotes the upper percentile quantities of the underlying distributions.

9.3 Monte Carlo Experiments

In order to assess the effects of misspecification in estimating the percentiles of the three distributions, an extensive range of possible cases is considered. The shape parameter is examined over a range of possible values where the density functions are positively skewed: the shape parameter takes the values 2 and 6 for the gamma distribution, 2 and 4 for the Weibull distribution, and 0.5 and 0.9 for the lognormal distribution. In all cases considered in this chapter, the arbitrary scale parameter is set to unity, and the location parameter takes on the values and 1 for the two- and three-parameter distributions, respectively. The lognormal distribution has the opposite behaviour to the gamma and Weibull distributions as the shape parameter is increased. For each entry in the tables, N = 1000 replications of the experiments are processed. The sample size used is n = 365, since it represents a common case, namely a full year of 24- hourly average observations. For each set of parameter values, BIAS and RRMSE are evaluated for estimates of the exact ninety-eighth percentile and the highest value which, for n = 365, is equivalent to using the ninety-ninth percentile.

The random sample generators used for the Monte Carlo experiments are DRNGAM, DRNWIB and DRNLNL for the gamma, Weibull and lognormal distributions, respectively. These are available as subroutines in the International Mathematical and Statistical Library (IMSL) in version 1.0 (April 1987). The same seed number (1234) is used to obtain the first random sample of the first of the 1000 replications. Varying the initial seed produces similar results to those reported in the chapter.

9.4 Monte Carlo Results

The results of the experiments in which the three- and two- parameter gamma distributions (G3 and G2, respectively), the three- and two-parameter Weibull distributions (W3 and W2, respectively), and the three- and two-parameter lognormal distributions (LN3 and LN2, respectively) are true are given in Tables 9.1 - 9.3, respectively. In all cases, the true model is either a three- or two-parameter distribution, the estimated quantities are the maximum value (MAX) or the ninety-eighth percentile value (98), and the performance criteria are BIAS and RRMSE. Both the three- and two-parameter versions of non-nested alternative distributions are estimated to examine the consequences of misspecifying the true distribution.

9.4.1 Gamma Distribution is True

The following points should be noted from the experiments reported in Table 9.1, in which the true model is either G3 or G2 and the misspecified models that are estimated are W3, W2, LN3 and LN2.

(i) When the shape parameter is 2 ($\alpha = 2$) and G3 is correct ($\gamma = 1$), LN2 has the lowest BIAS and RRMSE for both the maximum and 98 per cent quantities, followed by W3. Although the BIAS values for the correctly specified G3 are much closer to zero than are the BIAS values for LN2, the RRMSE values for

LN2 are very close to those of G3. The largest BIAS and RRMSE values for the maximum and 98 per cent quantities are LN3 and W2, respectively. When the shape parameter is increased to 6 ($\gamma = 6$) with $\gamma = 1$, the lowest BIAS and RRMSE values for both the maximum and 98 per cent quantities are obtained for LN3, followed by W3. In the case of $\alpha = 6$, the BIAS and RRMSE values for LN2 are the worst rather than the best, as was the case when $\alpha = 2$. The BIAS and RRMSE values for the correctly specified G3 are the lowest for both the maximum and 98 per cent quantities. Overall, the BIAS and RRMSE values of the incorrectly specified LN2 are highly sensitive to the value of the shape parameter whereas those for W3 are not. On the basis of these

results, it is clear that, in the absence of knowledge that G3 is correct, W3 is to be preferred of the four misspecified distributions.

(ii) However, when G2 is correct ($\gamma = 0$) and the shape parameter is 2 ($\alpha = 2$), the distribution with lowest BIAS and RRMSE is W3, followed by W2, LN3 and LN2. The lowest values of BIAS and RRMSE are, not surprisingly, for the correctly specified G2. When the shape parameter is increased to 6 ($\alpha = 6$), the lowest BIAS and RRMSE values hold for the correct G2. Of the misspecified distributions, the lowest are for LN3, followed by W3, W2 and LN2 On the basis of these results, it is clear that LN2 is the worst when G2 is correct, with W2 and LN3 being highly sensitive to the value of the shape parameter. As in the case of G3 being correct, W3 is again the most reliable for G2 in terms of BIAS and RRMSE for both the maximum and 98 per cent quantities.

9.4.2 Weibull Distribution is True

In Table 9.2 the true model is either W3 or W2 and the misspecified models that are estimated are G3, G2, LN3 and LN2. The main points are as follows

(i) When the three-parameter Weibull distribution is correct and $\alpha = 2$, the lowest BIAS and RRMSE values for the maximum quantity are obtained for G2, followed by LN2, G3 and LN3. It is worth observing that the only case in the three tables where the BIAS and RRMSE values for the correctly specified model exceed that of any of the misspecified models occurs in Table 9.2, where W3 is true and $\alpha = 2$. In this case, the RRMSE for G2 is slightly less than that for W3. For the 98 per cent quantity, however, the smallest BIAS and RRMSE values hold for LN2, followed by G2, LN3 and G3. As the shape parameter is increased to 4 ($\alpha = 4$) with W3 as the true distribution, the lowest BIAS and RRMSE values are again obtained for W3. The two lowest values of BIAS and RRMSE for the four misspecified distributions are G2 and LN3, followed by LN2 and G3. Overall, when W3 is true, LN3 and LN2 are sensitive to the value of the shape parameter, G3 has largest BIAS and RRMSE values, and G2 is the most

reliable of the four misspecified distributions.

(ii) In the case where W2 is the true distribution ($\alpha = 0$) and the shape parameter is 2 ($\alpha = 2$), the lowest BIAS and RRMSE values are for the true W2 distribution. For the maximum quantity, the low BIAS and RRMSE values hold for G3, followed by LN3, G2 and LN2, for the four misspecified distributions. For the 98 per cent quantity, the lowest BIAS and RRMSE values are obtained for G3 an LN3, ahead of G2 and LN2. When the shape parameter is increased to 4 ($\alpha = 4$), several of the rankings for $\alpha = 2$ are altered. The misspecified G3 is found to be highly sensitive to the value of the shape parameter, going from the lowest to the highest BIAS an RRMSE values for both the maximum and 98 per cent quantities. The true W2 always has the lowest BIAS and RRMSE values, with G2 being the best of the four misspecified distributions. followed by LN3 and LN2. While G2 is also slightly sensitive to the value of the shape parameter as it is increased, G2 still appears to be the most reliable of the misspecified distributions when W2 is true, just as in the case where W3 is true.

9.4.3 Lognormal Distribution is True

Finally, Table 9.3 contains the results from experiments when either LN3 or LN2 is true and the misspecified models that are estimated are G3, G2, W3 and W2. Since the lognormal distribution has the opposite behaviour to the gamma and Weibull distributions as the shape parameter is increased, for purposes of comparison it is useful to examine the results as the shape parameter is decreased rather than increased. The correctly specified distribution always has the smallest BIAS and RRMSE values for both the maximum and 98 per cent quantities. The principal points to note from the table are as follows.

(i) When LN3 is correct ($\alpha = 1$) and the shape parameter is 0.9 ($\alpha = 0.9$), the lowest BIAS and RRMSE values for both the maximum and 98 per cent quantities are found for W3, followed by G3, W2 and G2. The rankings are altered as the shape parameter is reduced to $\alpha = 0.5$, with the best performance being given by G3, followed

by W3. For the maximum quantity, W2 has the highest BIAS and RRMSE values, whereas G2 has the highest values for the 98 per cent quantity. IN terms of reliability, both G3 and W3 perform the best of the four misspecified distributions when LN3 is the true distribution.

(ii) The rankings for the case where LN2 is the true distribution ($\alpha = 0$) is very similar to that given above for LN3 being true. When the shape parameter is 0.9 ($\alpha = 0.9$), W3 heads the rankings in terms of lowest BIAS and RRMSE values for both the maximum and 98 per cent quantities, followed by G3, W2 and G2. Just as in the case where LN3 is true, when the shape parameter is reduced to 0. ($\alpha = 0.5$), the rankings are now given as G3, W3, G2 and W2 where LN2 is the true distribution. The least sensitive misspecified distributions to changes in the shape parameter are G3 and W3, while G2 and W2 typically have higher BIAS and RRMSE values for both the maximum and 98 per cent quantities.

9.5 Concluding Remarks

In this chapter we have assessed the effects of misspecification in estimating the percentiles of the two- and three-parameter gamma, Weibull and lognormal distributions. In the experiments, the true model is either a two- or three-parameter distribution, the estimated quantities are the maximum observed value or the ninety-eighth percentile value, the performance criteria are the BIAS and RRMSE associated with the estimated quantities, and four misspecified non-nested alternative distributions are estimated. The two- and three- parameter versions of two non-nested distributions are estimated to examine the consequences of misspecifying the true distribution, namely estimating the two- or three-parameter distribution when a two- or three-parameter version of a non-nested distribution is correct. The shape parameter is examined over a range of possible values where the density functions are positively skewed. When the two- or three-parameter gamma distribution is true, the three-parameter Weibull distribution is found to be the most reliable misspecified distribution in terms of lowest

BIAS and RRMSE values and lack of sensitivity to the value of the shape parameter. For the case where the two- or three-parameter Weibull distribution is true, the most reliable misspecified distribution is the two-parameter gamma distribution, which is somewhat unusual in that it is clearly preferred to its three-parameter counterpart. Finally, the three-parameter gamma and Weibull distributions are the most reliable misspecified distributions when the two- or three-parameter lognormal distribution is true.

TABLE 9.1
Estimates of BIAS and RRMSE at maximum and 98 % quantities from 1000 Monte Carlo simulations with sample size n = 365 for the gamma distribution ($\beta = 1$)

GAMMA DISTRIBUTION								
			Five Estimated Models					
True Values	Percentile	Performance Criteria	True Model		Four Misspecified Models			
			G3	G2	W3	W2	LN3	LN2
G3	MAX	BIAS	0.007	-	-0.074	-0.206	0.226	0.023
α=2.0		RRMSE	0.053	-	0.091	0.210	0.250	0.059
γ=1.0	98	BIAS	0.004	-	-0.027	-0.089	0.065	-0.022
		RRMSE	0.042	-	0.050	0.097	0.086	0.046
G2	MAX	BIAS	-	-0.002	-0.083	-0.110	0.252	0.904
α=2.0		RRMSE	-	0.053	0.101	0.123	0.278	0.927
γ=0.0	98	BIAS	-	-0.002	-0.031	-0.047	0.076	0.359
		RRMSE	-	0.047	0.058	0.067	0.100	0.372
G3	MAX	BIAS	0.004	-	-0.077	-0.138	0.046	0.110
α=6.0		RRMSE	0.044	-	0.085	0.141	0.074	0.119
γ=1.0	98	BIAS	0.002	-	-0.025	-0.053	0.011	0.041
		RRMSE	0.031	-	0.039	0.060	0.036	0.052
G2	MAX	BIAS	-	-0.001	-0.082	-0.128	0.048	0.229
α=6.0		RRMSE	-	0.036	0.091	0.132	0.079	0.237
γ=0.0	98	BIAS	-	-0.001	-0.027	-0.050	0.012	0.096
		RRMSE	-	0.030	0.042	0.058	0.039	0.104

Note : α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the 3- and 2-parameter gamma distributions, W3 and W2 the 3- and 2-parameter Weibull distributions, and LN3 and LN2 the 3- and 2-parameter lognormal distributions.

TABLE 9.2
Estimates of BIAS and RRMSE at maximum and 98 % quantities from 1000 Monte Carlo simulations with sample size n = 365 for the Weibull distribution ($\beta = 1$)

WEIBULL DISTRIBUTION								
			Five Estimated Models					
True Values	Percentile	Performance Criteria	True Model		Four Misspecified Models			
			W3	W2	G3	G2	LN3	LN2
W3	MAX	BIAS	0.002	-	0.079	-0.007	0.097	0.060
α=2.0		RRMSE	0.030	-	0.088	0.024	0.109	0.066
γ=1.0	98	BIAS	0.000	-	0.027	-0.015	0.026	0.011
		RRMSE	0.022	-	0.037	0.024	0.037	0.023
W2	MAX	BIAS	-	-0.002	0.110	0.221	0.135	0.995
α=2.0		RRMSE	-	0.037	0.122	0.226	0.151	1.011
γ=0.0	98	BIAS	-	-0.002	0.041	0.104	0.040	0.414
		RRMSE	-	0.031	0.055	0.110	0.056	0.422
W3	MAX	BIAS	0.000	-	0.610	0.065	0.036	0.094
α=4.0		RRMSE	0.014	-	1.881	0.066	0.285	0.095
γ=1.0	98	BIAS	0.000	-	0.541	0.028	0.017	0.040
		RRMSE	0.010	-	1.998	0.029	0.210	0.041
W2	MAX	BIAS	-	-0.001	0.998	0.213	0.059	0.411
α=4.0		RRMSE	-	0.019	3.056	0.215	0.462	0.416
γ=0.0	98	BIAS	-	-0.001	0.930	0.102	0.029	0.189
		RRMSE	-	0.015	3.418	0.104	0.360	0.192

Note : α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the 3- and 2-parameter gamma distributions, W3 and W2 the 3- and 2-parameter Weibull distributions, and LN3 and LN2 the 3- and 2-parameter lognormal distributions.

TABLE 9.3
Estimates of BIAS and RRMSE at maximum and 98 % quantities from 1000 Monte Carlo simulations with sample size n = 365 for the lognormal distribution ($\beta = 1$)

LOGNORMAL DISTRIBUTION								
			Five Estimated Models					
True Values	Percentile	Performance Criteria	True Model LN3	True Model LN2	Four Misspecified Models G3	G2	W3	W2
LN3 α=0.9 γ=1.0	MAX	BIAS	0.017	-	-0.403	-0.492	-0.374	-0.466
		RRMSE	0.134	-	0.407	0.494	0.381	0.470
	98	BIAS	0.007	-	-0.156	-0.232	-0.125	-0.174
		RRMSE	0.089	-	0.170	0.240	0.146	0.190
LN2 α=0.9 γ=0.0	MAX	BIAS	-	-0.001	-0.413	-0.449	-0.383	-0.427
		RRMSE	-	0.111	0.417	0.452	0.390	0.432
	98	BIAS	-	-0.002	-0.165	-0.200	-0.131	-0.163
		RRMSE	-	0.084	0.180	0.210	0.155	0.180
LN3 α=0.5 γ=1.0	MAX	BIAS	0.005	-	-0.137	-0.240	-0.194	-0.271
		RRMSE	0.078	-	0.145	0.242	0.200	0.275
	98	BIAS	0.001	-	-0.044	-0.105	-0.059	-0.090
		RRMSE	0.048	-	0.059	0.110	0.072	0.100
LN2 α=0.5 γ=0.0	MAX	BIAS	-	-0.002	-0.149	-0.213	-0.210	-0.267
		RRMSE	-	0.061	0.157	0.217	0.217	0.271
	98	BIAS	-	-0.002	-0.050	-0.092	-0.067	-0.095
		RRMSE	-	0.046	0.066	0.100	0.082	0.106

Note : α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the 3- and 2-parameter gamma distributions, W3 and W2 the 3- and 2-parameter Weibull distributions, and LN3 and LN2 the 3- and 2-parameter lognormal distributions.

Part IV

APPLICATIONS TO AIR QUALITY MANAGEMENT

Chapter 10

Estimation and Discrimination of Alternative Air Pollution Models

10.1 Introduction

The use of probability distributions as descriptions of pollutant concentrations offers many advantages, as discussed in Chapter 2, and this issue need not be repeated here. Jakeman and Taylor (1989) summarise a substantial literature, which indicates that many probability distribution functions have proven to be good representations of the frequency distribution of air pollutant data. Among these distributions are the six analysed in this thesis. The gamma, Weibull and lognormal distributions are the most well known and frequently used, particularly the two-parameter distributions and sometimes their three-parameter counterparts.

These distributions are not only parsimonious but are also sufficiently flexible in fitting real data. Importantly, they cover a wide range of situations. For example, the gamma distribution will tend to the normal as the shape parameter becomes very large (Pearson (1959)). The lognormal distribution can also approximate the normal distribution when the value of the shape parameter is zero (Aitchison and Brown (1957)). When the shape parameter is equal to or less than unity, both the gamma and Weibull distributions approach the exponential or become inverse "J" shaped. As will be shown in this chapter, air pollution data can conform to both the "bell" shaped distribution with a wide range of skewness and the inverse "J" shaped distribution. Unlike some

other skewed distributions, the gamma and Weibull distributions can fit both kinds of data reasonably well. Moreover, the lognormal distribution, particularly the two-parameter version, has been considered as "the most appropriate for characterizing both reactive and inert pollutant concentrations for a wide range of averaging times" (Benarie (1980, p. 305)). Therefore, the gamma, Weibull and lognormal distributions are the three candidates used in the empirical application in this chapter.

As discussed in Chapter 3, there have traditionally been problems with employing the maximum likelihood (ML) method to estimate the parameters of the gamma and Weibull distributions. A theoretical difficulty occurs when the shape parameter is less than or equal to unity, since one of the three first-derivative equations often used to maximize the likelihood function is not valid in these cases. Chapter 3 proposed a new algorithm which can provide computationally efficient ML estimates for both the three-parameter gamma and Weibull distributions. When the sampling distribution is quite skewed, it was found that the classical ML method for estimation of the three-parameter lognormal distribution, which involves solving the first-derivative equations, has difficulty in converging. Using the new approach in Chapter 3, this problem can also be avoided. Therefore, in this chapter, the new ML method is used for fitting air pollution data.

In Chapters 6 and 8, extensive investigations were undertaken for discrimination procedures among a number of alternative distributions. In Chapter 8, it was argued that the application of standard hypothesis tests will not strictly be valid if there are more than two models. Monte Carlo simulation experiments also revealed that existing discrimination criteria, such as Akaike's information criterion (AIC), Schwarz's information criterion (SIC), the Kullback-Leibler (KL) information criterion, the chi-square (CHI) test and the Kolmogorov-Smirnov (KS) test, are not sufficiently reliable for the intended use of a model. Two performance criteria were proposed to investigate their discrimination ability in the same chapter. The criteria relate to the upper percentile error (UPE) and the entire or full percentile error (FPE) (see Chapter 8 for details). When applied in isolation, both UPE and FPE proved inadequate for selecting among

the different underlying distributions.

To address this problem, a new generalised information criterion (GIC) was proposed. From the Monte Carlo simulation experiments, the GIC was shown in Chapters 6 and 8 to perform reliably within the parameter spaces and sample sizes investigated. In this chapter, the GIC and other criteria are applied to discriminate among six distributions for annual sets of air pollutant concentration data. An interesting result from the experiments of Chapters 6 and 8 is that when the true underlying distribution is one of the three-parameter gamma, Weibull or lognormal distributions, but discrimination is restricted among the two-parameter distributions, the two-parameter lognormal distribution will frequently be chosen if the sample distribution is skewed. This means that the two-parameter lognormal distribution will often be a better approximation to one of the three-parameter distributions than the nested member (with location parameter equal to zero) of the underlying true distribution. This result is consistent with the use of the two-parameter lognormal distribution being most frequently used for fitting probability distributions of air pollution data. Therefore, care should be exercised in such situations because a more appropriate underlying distribution may be one of the three-parameter models. This gives a strong reason for including the three-parameter distributions in the discrimination procedure.

In the development of the research for this thesis, a computer package named "Probability-distribution Fitting" (PROFIT) has been designed for the general fitting of probability distributions and, in particular, for modelling problems in air quality management. One feature of the package is that it involves the new estimation and discrimination procedures developed in the thesis for the six distributions (initially), as well as traditional moment estimation and discrimination procedures. PROFIT requires only free format for input data. It can analyse observational data from a number of years and sites independently or simultaneously. The output will provide detailed statistical information in a form similar to the results reported in this chapter. In addition, this package can also perform Monte Carlo simulation experiments. This is a very convenient facility for understanding the performance of the estimation and discrim-

ination procedures applied to observed data by comparing the empirical results with synthetic results from parent distributions which closely fit the real data of interest.

In this chapter, PROFIT is applied to annual sets of air pollutant concentrations recorded at several monitoring sites in Melbourne, Australia. Six types of pollutants from up to five monitoring stations are analysed using a comprehensive model selection procedure. This procedure incorporates use of discrimination and performance criteria, together with analyzing the effects of misspecifying the distribution and of errors in estimation of observed upper percentile values. The results contained in this chapter are useful for general purposes, such as summarising or smoothing the data, particularly the upper percentiles, as well as providing statistical information to construct hybrid models. The last feature will be demonstrated in Chapter 11.

10.2 Considerations for a Comprehensive Model Selection Procedure

A brief review of existing discrimination criteria has been given in Chapter 8. These criteria can be categorised in terms of functions of maximized log-likelihood values of distributions, such as AIC and SIC, or as a deviation between properties of the fitted and observed distributions, such as the CHI and KS tests. When the sample size tends to infinity, all of these criteria are expected to provide similar results since the sampling distribution will converge to the underlying population distribution. However, the performance of these criteria will differ in finite samples, as occurs when fitting empirical data or data from simulation experiments. In isolation, these criteria may not provide adequate discrimination for the types of probability distribution applications that arise in air quality management.

The choice of an appropriate distribution should depend on the modelling objectives. Application of probability distributions to air quality management problems requires adequate fitting of both the upper percentiles and overall fitting of the distribution to observed data, particularly those parts such as the mean which are used as

calibration points for hybrid modelling. Distributions which fit the upper percentiles well may not fit other parts of the data well. Criteria based upon deviations from the upper percentiles of the distribution should not be the sole criteria for discrimination. Such a criterion to select a distribution is the UPE, defined in Chapter 8, which is based on relative mean square fitting errors averaged over the upper two percentiles. From the results given in Chapter 8, UPE can perform quite well in discriminating among certain distributions but, as with other criteria, it frequently does not perform adequately when used in isolation. Thus, the selection of an appropriate distribution may be a compromise among various discrimination criteria if they yield conflicting results. In addition, an objective for hybrid and other modelling is to select a suitable distribution for a given pollutant at a particular site for separate observation periods (e.g. years), or over varying sites and periods. The weighting of these factors in a comprehensive model selection procedure is the motivation of the present chapter.

There are four basic considerations in this chapter. The first is to use GIC and traditional discrimination criteria to examine which parametric distribution best describes the data most frequently. The GIC procedure is regarded as the most important of these criteria because it does not reject a candidate if, after adjusting for the number of parameters and/or observations, its maximised log-likelihood value is not significantly inferior to the distribution with the highest maximised log-likelihood value. The second consideration involves the level of errors in fitting the observed percentiles of interest, which has generally been taken to be the relative root mean square error of the maximum percentile and a weighted average of the upper percentiles (i.e. the UPE based upon deviations from sample percentiles). However, other error functions, for example, the second highest concentration value, may be of more interest for specific applications. One of the purposes served by these functions is to allow rejection of alternative distributions which lead to larger errors than some practical level designated as being tolerable.

The other two considerations in the comprehensive model selection procedure can be viewed as of secondary importance. The third consideration involves use of the

misspecification results of Chapters 7 and 9 for nested and non-nested distributions, respectively. These results are of value especially when it is difficult to select among alternatives with comparable performances on the basis of the two primary considerations. They allow selection of the distribution which, if the alternative were true, leads to the most tolerable percentile deviations from the true distribution. The fourth consideration involves misspecification when the associated performance measures are based upon deviations from the empirical rather than from the true distribution. The analysis of this type of misspecification and the rationale for its use are addressed in the next section, followed by the results from fitting real air pollutant data from the urban Melbourne airshed.

10.3 Evaluating Misspecification Errors

The effects of misspecification in estimating the upper percentiles of nested and non-nested distributions have been discussed in Chapters 7 and 9, respectively. The results obtained from the Monte Carlo experiments are based on comparisons of sample errors with the percentiles of the known true distribution. These results are useful for determining the errors in the percentiles of interest of the true distribution as well as other alternatives. Although the true distribution is unknown in practice, the Monte Carlo results provide some guidance as to the adoption of low risk strategies for model selection. Thus, if it is difficult to choose between two models on the basis of GIC and other criteria, the errors of misspecification can then be used to infer the consequences of selecting one model when the other is true. For example, using historical data sets, three-parameter gamma and two-parameter lognormal distributions might be preferred over other distributions with approximately the same frequency of acceptance. From Table 9.1, a two-parameter lognormal estimated using data from a true three-parameter gamma distribution may estimate the percentiles as successfully as the true distribution estimated for the same samples in terms of RRMSE when the value of the shape parameter is 2. However, when using a three-parameter gamma to fit sam-

ples taken from a two-parameter lognormal distribution in Table 9.3, a substantially inferior reproduction of the percentiles is obtained in comparison with those obtained using the true distribution when the shape parameter is 0.9. Thus, the two-parameter lognormal would be selected as a low risk option relative to the three-parameter gamma distribution.

As an additional aid in model selection, it is also useful to re-examine the effects of misspecification in estimating the upper percentiles of nested and non-nested distributions using analogous goodness-of-fit measures based on the observed data rather than the underlying true values. This allows errors of fit with real data to be compared with the errors obtained from simulations where the distribution and true parameters are similar to those estimated using real data.

Corresponding to this need, the associated performance measure given importance here is a relative root mean square error of the upper percentiles averaged over all observations in the 98 to maximum percentile range, and then averaged over all N Monte Carlo simulations from the same parent distribution. The performance criterion (PC) can be written as

$$PC = \frac{1}{N}\sum_{i=1}^{N}[\frac{1}{1+(1-p)n}\sum_{j=pn}^{n}(\frac{\hat{q}_{ij}-q_{ij}}{q_{ij}})^2]^{0.5} \qquad (10.1)$$

where p is the ordinate of the p-quantile which is generally chosen at 0.98, while the range between 98 per cent and maximum quantities of the fitted $\hat{q}_{ij}$ and observed q_{ij} is of most concern in the management of air quality standards. Each percentile error is a relative mean square error that is weighted from 98 to the maximum percentiles by the observed sample value. The errors are averaged and transformed to derive a performance criterion for each experiment. The mean value (MEAN) of this performance criterion, its standard deviation(SD), and maximum (MAX) and minimum (MIN) values over all experiments are used to examine goodness-of-fit for both nested and non-nested distributions.

The results from the experiments are reported in Tables 10.1 to 10.3. Experiments are replicated 1000 times for each set of parameters. The sample size used in each

experiment is 365 which is a common case, representing a period of one year of 24-hour average concentrations or one year of the daily maxima of any shorter time average. Results for smaller sample sizes would be useful where there are many missing data points or where shorter (e.g. seasonal rather than annual) periods are of interest. Two values of the shape parameter have been used to accommodate quite skewed and near-symmetric density functions. The following points should be noted in the results obtained with a sample size of $n = 365$ and the performance criterion in equation (10.1).

(i) When either the two- or three-parameter gamma distribution is true, the three-parameter Weibull distribution performs almost as well as the true model in terms of all four statistics, irrespective of the values of the shape parameter. For the case of the true three-parameter gamma distribution where the value of the shape parameter is 2, the two-parameter lognormal distribution also performs very well overall. However, the two-parameter lognormal performs poorly when the underlying distribution is the two-parameter gamma, and the former is inferior in the more symmetric case where the value of the shape parameter is 6 when the underlying distribution is the three-parameter gamma. The two-parameter gamma distribution also performs well for the more symmetric version of the underlying three-parameter gamma. As expected, three-parameter gamma distribution performs very well when the underlying distribution is the two-parameter gamma. The three-parameter lognormal also performs well when either the two- or three-parameter gamma is the underlying distribution and the value of the shape parameter is 6.

Based on these results, if the true distribution is the three-parameter gamma and the probability density function is quite skewed, the three-parameter Weibull and two-parameter lognormal distributions will have similar performances to the true model in fitting the upper percentiles. However, the two-parameter lognormal distribution will worsen and the two-parameter Weibull will improve if the true distribution is the two-parameter gamma. When the probability density function of the data tends to symmetry, the three-parameter lognormal and Weibull distributions are very good in

terms of the performance criterion defined in equation (10.1).

(ii) When the true distribution is either the two- or three-parameter Weibull, the true model almost invariably fits substantially better than all other alternatives (see Table 10.2). When the value of the shape parameter is 2 for the sample from the three-parameter Weibull, the two-parameter gamma distribution performs as well as the true distribution. The three-parameter gamma and the two- and three-parameter lognormal distributions perform poorly for both values of the shape parameter when the data come from the true two- or three-parameter Weibull distribution.

(iii) When either the two- or three-parameter lognormal is correct, the true model almost invariably fits substantially better than the other alternatives (see Table 10.3), especially if the distribution is very skewed. However, when the distribution becomes quite symmetric, the three-parameter gamma distribution yields relatively lower errors than those of other false distributions. An interesting point arises if the true distribution is the two-parameter lognormal, in which case all non-nested distributions have relatively large errors, facilitating recognition of this distribution.

In summary, the use of the performance criterion given in (10.1) in isolation can lead to the following conclusions. The three-parameter Weibull or two-parameter lognormal could be selected incorrectly, especially for more skewed probability density functions, even though the data come from the three-parameter gamma. Similarly, the two- and (especially) three-parameter Weibull could be selected for samples taken from the two-parameter gamma distribution. When the probability density function tends to symmetry, the errors in fitting the upper percentiles become relatively lower for different model alternatives, implying that wrong distributions will have higher probabilities of being selected. Generally, discrimination using this performance criterion can be confusing if the sample comes from either the two- or three-parameter gamma distribution. When the sample is taken from the two-parameter lognormal, the true model and its three-parameter counterpart should perform similarly, and differ from other non-nested alternatives, especially for the skewed sample case. If the sample

comes from the three-parameter lognormal, the three-parameter gamma distribution may perform reasonably well if the sample becomes quite symmetric. When the sample is taken from the Weibull distribution, the true distribution fits the data very well. However, when the sample comes from the two-parameter Weibull, its three-parameter counterpart may perform slightly better in terms of fitting the upper percentiles. The other remaining distributions produce substantially larger errors than the true model. Hence, in most cases discrimination is generally much clearer when the samples are taken from the Weibull distribution.

The errors reported in Tables 10.1 - 10.3 for the true underlying distribution represent the minimum errors that could be expected in fitting the upper percentiles. Thus, for the three-parameter gamma distribution, this error is on average between 0.052 and 0.074 as the value of the shape parameter changes from a value of 2 (quite skewed) to 6.0 (nearly symmetric). The standard deviation of this error varies between 0.022 and 0.033, while the maximum error found over 1000 sets of random samples correspondingly varies from 0.144 to 0.206. For the underlying three-parameter Weibull distribution, this error is on average between 0.014 and 0.033 when the value of the shape parameter changes from a value of 2 to 4. The standard deviation of this error varies between 0.006 and 0.015, while the maximum error found over 1000 sets of random samples correspondingly varies from 0.043 to 0.097. The skewed sample case yields worse errors. When the three-parameter lognormal distribution is correct, the error is on average between 0.083 and 0.167 as the value of the shape parameter changes from 0.9 to 0.5. The standard deviation of this error varies between .037 and .077, while the maximum error found over 1000 sets of random samples correspondingly varies from 0.263 to 0.500. These figures provide an indication of the errors to be expected in fitting the observed percentiles of the six distributions.

10.4 Fitting Real Data

To demonstrate how the preceding work in the thesis can be combined to form a useful comprehensive model selection procedure, six air pollutants over 64 sites and a ten-year period from Melbourne, Australia, have been used. They are carbon monoxide (CO), nitrogen monoxide (NO), nitrogen dioxide (NO_2), nitrogen oxides (NO_x), sulphur dioxide (SO_2), and β-scattering. To simplify the presentation, results are provided for annual sets of pollutant data from five stations, namely the Museum (site number 11), Alphington (27), Dandenong (34), Camberwell (81) and Footscray (91) stations. The time period considered in this study is the 24-hour average since, for almost all pollutants, there is an international air quality standard expressed in terms of this time period.

As an illustrative example, full details of the estimation and discrimination procedure for CO and β-scattering at the Museum station are given in Subsections 10.4.1 and 10.4.2, respectively, while the remaining results for other pollutants at this site are given in Appendix 2. Following these two subsections, results are provided for CO over three sites and for β-scattering over five sites in the ten-year period reported in Subsections 10.4.3 and 10.4.4, respectively. Then a summary is given of results for NO, NO_2, NO_x and SO_2 in Section 10.5, with more detailed tables being available in Appendix 2. Some concluding remarks are given in Section 10.6.

10.4.1 Detailed CO Results for the Museum Site

Table 10.4 includes three major pieces of information for estimating distributions of CO over different years at the Museum station, namely the estimated parameter values, maximized values of the log-likelihood functions and RRMSE values between predicted and observed percentiles. Note that, for the three-parameter distributions, the estimate of the location parameter may be negative, which is not consistent with air quality data being positive. In such cases, these distributions are automatically deleted from the discrimination procedure. There are three years, namely 1975, 1978 and 1984, in which

all three-parameter distributions are eliminated. An initial reaction might be that the three-parameter distributions might be inadequate for describing CO at this site. The estimated values of the shape and scale parameters suggest that the samples are heavily skewed. There is only one exception, in 1976, where the estimated distributions are quite symmetric.

In Table 10.4, the maximized values of the log-likelihood functions show that, in many cases, the differences between some non-nested distributions are very small. This demonstrates the importance of using GIC, as recommended in Chapter 8, because it does not reject alternatives which are not significantly inferior to the model with the highest maximized log-likelihood value. The estimated RRMSE values for different percentiles fitted to CO data are also reported in Table 10.4. MAX1, MAX2 and 98 denote the maximum, second highest and 98 percentiles, respectively. MEAN-U is the weighted mean of RRMSE from 98 to the maximum percentiles, as defined in (10.1). MEAN-F has a similar definition when the parent percentile is replaced by the observed or sample value. Generally, the estimated three-parameter distributions have lower values of RRMSE than the two-parameter distributions but, in cases such as 1980, the two-parameter gamma distribution performs very well at the upper percentiles.

A summary of eight discrimination and performance criteria applied to samples of CO over the ten-year period is given in Table 10.5. Of these criteria, the first four are based on the maximized values of the log-likelihood function and the next four are specialised performance criteria adopted for the fitting of distributions. Basically, the comprehensive model selection procedure proceeds in one of two ways: among one nested and five non-nested distributions, or among two sets of non-nested distributions, namely the three- and two-parameter distributions. The procedure considered here presumes that, in practice, discrimination arises only for non-nested two- or three-parameter distributions. The results for CO are discussed below.

In discriminating among the six distributions, GIC selects the two-parameter Weibull distribution most frequently for the CO data sets. GIC yields the highest acceptance

rate, namely 8 acceptances from 10 samples. The two-parameter gamma distribution is ranked as second best, with a slightly lower acceptance rate, namely 7 from 10. CHI and KS are also marginally in favour of the two-parameter Weibull over the two-parameter gamma distribution. For KL, AIC, SIC, and UPE, the frequency of acceptance of the preferred distribution is relatively lower, but the two-parameter Weibull distribution still has the highest or equal highest frequency of acceptance. FPE favours the two-parameter Weibull in 9 cases, while the three-parameter Weibull is accepted once. In discriminating among only three-parameter distributions, the three-parameter Weibull is clearly favoured by all criteria. When the discrimination procedure is undertaken among only the two-parameter distributions, the Weibull is preferred most often but the gamma is also quite competitive. Therefore, based solely on the results of the various discrimination criteria in Table 10.5, the two-parameter Weibull is the appropriate distribution for the ten annual sets of CO data from the Museum station.

Although the discrimination criteria indicate the two-parameter Weibull distribution to be the appropriate distribution for the annual data sets, it is necessary to analyse the errors associated with percentile predictions. Table 10.6 provides a summary of the RRMSE performance for different percentile predictions including the maximum, upper and full percentiles. The maximum percentile is especially important in assessing CO in terms of air quality standards. In considering the maximum percentile, the best distribution is the two-parameter gamma, but the two-parameter Weibull also works reasonably well. For fitting the full percentiles, the two-parameter Weibull is much better than the other distributions. However, FPE is not necessarily a critical statistic since it can be dominated by errors at very low percentiles, which need not be predicted accurately for air quality management purposes. For upper percentile performance, the two-parameter gamma is the best among the two-parameter distributions.

From the simulation results in this chapter, if a two-parameter Weibull were the underlying distribution, then the two-parameter Weibull would easily outperform the two-parameter gamma in terms of the errors from the observed upper percentiles. Sometimes, a three-parameter distribution provides the best fit of the data. However,

the results using the discrimination criteria might suggest that the data are not necessarily from a three-parameter distribution. Such a result is made very clear when the three-parameter distribution fits the data better on account of an unrealistic negative estimate of the location parameter. The RRMSE values over the observed upper percentiles in Table 10.6 for the two-parameter gamma and Weibull distributions are quite competitive, which suggests a two-parameter gamma distribution is the appropriate distribution, according to the analysis of the errors in Section 10.3. The results of both the discrimination and performance criteria suggest that either the two-parameter gamma or Weibull distribution are adequate for representing the historical frequency distributions of 24-hour average concentrations of CO annually at this site.

10.4.2 Detailed β-Scattering Results for the Museum Site

As another example to illustrate the comprehensive model selection procedure, the results for β-scattering data are reported in Tables 10.7 to 10.9. Table 10.7 lists the estimates of the parameters, the maximized values of the log-likelihood functions and RRMSE values for predictions of different percentiles. In terms of the estimated shape and scale parameters, the sampling distributions can be skewed or nearly symmetric, and all of the estimated location parameters for the three-parameter distributions over the ten data sets are positive. Unlike the CO data sets, the differences between the three- and two-parameter distributions in terms of both the maximized log-likelihood and RRMSE values for different percentiles are relatively large.

Table 10.8 provides a summary of the eight discrimination and performance criteria applied to the β-scattering data. GIC chooses the three-parameter lognormal as the appropriate distribution from the alternatives, and its acceptance rate is 8 from 10. It is interesting to note that, for the 1977 data, the results and the maximized log-likelihood values for the lognormal distributions are typical of an underlying two-parameter distribution. Since the two-parameter distribution is a special case of the three-parameter model, this is not especially surprising. There are two cases, for the 1981 and 1982 data, where the maximized log-likelihood values of the three-parameter

lognormal distribution are lower than for the three-parameter gamma and Weibull distributions. In these cases, the three-parameter lognormal cannot fit a heavily skewed distribution as well as the three-parameter gamma and Weibull distributions. Nevertheless, the three-parameter lognormal distribution is still generally the appropriate distribution.

All other criteria, and especially the CHI and KS tests, also favour the three-parameter lognormal distribution. When discriminating separately among two- and three-parameter distributions, the two- and three-parameter versions of the lognormal distribution are also strongly preferred. Table 10.9 provides further evidence that the three-parameter lognormal is the most suitable distribution overall. The errors in fitting the upper and maximum observed percentiles are substantially lower for the three-parameter lognormal distribution than for the others.

The misspecification results of Chapter 9 can also be recalled to assist in the selection of the appropriate distribution for β-scattering at the Museum site. The estimated values of the parameters of all distributions for the Museum site suggest a skewed distribution. For data from the lognormal distribution and parameter values in this range, the errors in the upper percentiles of fitting a three-parameter gamma or Weibull distribution are generally substantially higher than the percentile errors in fitting the three-parameter lognormal distribution to data from either the gamma or Weibull distribution. A similar result can also be observed for the two-parameter gamma and Weibull distributions for the CO data. However, for the β-scattering data, the values of the errors and the risks of misspecification are much greater. Indeed, the error in fitting a three-parameter gamma or Weibull distribution to data from a three-parameter lognormal can be as high as 40 per cent for the maximum percentile (see Table 10.3), whereas fitting a lognormal incorrectly to either the three-parameter gamma or Weibull distribution is less than 30 per cent.

Considering the RRMSE values over the upper percentile range in Table 10.9 and the results of the discrimination criteria, the three-parameter lognormal distribution

is selected as the appropriate distribution for representing the historical frequency distribution of β-scattering data at the Museum station.

10.4.3 Summary of Air Pollutant Distributions for Multiple Sites

For a particular pollutant, it is ideal if one distribution fits the data consistently well, not only over various years but also over various sites. Such a result will lead to a useful simplification when applying statistical and hybrid modelling approaches to predict probability distributions of pollutant concentrations.

Following the first example, a total of twenty annual CO data sets from three sites (namely 11, 27 and 81) are used for discrimination, and the associated results are listed in Tables 10.10 and 10.11. The two-parameter Weibull is selected by GIC, CHI and KS as the best distribution, with a high acceptance rate when discriminating among the six distributions. The UPE statistic also yields a high acceptance of the two-parameter Weibull distribution, which fits the upper range of sample percentiles best for the data sets. The AIC and SIC procedures also support the dominance of the two-parameter Weibull distribution, although less strongly. When considering discrimination among non-nested two- (three-) parameter distributions, the two- (three-) parameter Weibull distribution is preferred.

On the basis of RRMSE values for different percentiles in Table 10.11, the two-parameter Weibull distribution performs most strongly. The analysis of misspecification of this chapter also suggests that, if a two- or three-parameter Weibull distribution fits the observed values significantly better than non-nested alternatives, then the underlying distribution is Weibull. Therefore, for CO data sets over these sites and years, the two-parameter Weibull can be regarded as the appropriate distribution.

For β-scattering, there are annual data sets from five sites available for discrimination purposes, yielding twenty-eight years of data. The results are reported in Tables 10.12 and 10.13. From Table 10.12, the three-parameter lognormal alternative may be regarded as the best distribution. GIC has an acceptance rate of 86.2 per cent for the

three-parameter lognormal, and the CHI and KS tests perform similarly. AIC and SIC also yield high acceptance frequencies for the three-parameter lognormal distribution. The fact that the UPE and FPE statistics also yield high acceptance rates is an added bonus, as they show that the three-parameter lognormal almost always fits the maximum and upper percentiles best. When evaluating the results from discrimination among only two- and only three-parameter distributions, the lognormal distribution is again preferred.

In Table 10.13, the RRMSE values for different percentiles indicate that the three-parameter lognormal distribution fits the data best in most cases. Occasionally, the three-parameter gamma and Weibull distributions have slightly lower RRMSE statistics; for example, both have lower minimum errors than the three-parameter lognormal distribution for the mean of the upper percentiles. For the two-parameter Weibull distribution, both the minimum and maximum values of the mean of the upper percentiles are slightly smaller than for the three-parameter lognormal. However, the mean and standard deviation for all percentiles of the three-parameter lognormal distribution are consistently smaller than the others. Based on the results from both discrimination and performance criteria, the three-parameter lognormal can be regarded as the appropriate distribution.

In conclusion, for the two pollutants illustrated above, certain types of distributions were found to perform consistently well over different sites and years. The discrimination and performance criteria can be used jointly to assist in the selection of the appropriate distribution for a given air pollutant. Based on these results, a hybrid deterministic-statistical model could be constructed for these pollutants to link the causal variables to properties of the distribution, such as using the deterministic model to predict the mean or other mid-percentile values of the probability distribution. The parameters of the distribution can be estimated based on these properties and the probability distribution of pollutant concentrations can then be determined. The distributional assumption could be applied at all sites investigated here and, with reasonable confidence, also at intervening locations for which records of historical concentrations

are not available.

10.5 A Brief Discussion of the Results for Pollutants NO, NO_2, NO_x and SO_2

The results of estimation and discrimination for the pollutants NO, NO_2, NO_x and SO_2 over different years and sites are given in Appendix 2. Tables A2.1 to A2.12 in Appendix 2 provide the details of fitting the pollutant data at the Museum site. The remaining tables there give results for these pollutants at multiple sites. Results for the Museum site are discussed before the other sites.

Based on the results in Tables A2.1 to A2.3, the two-parameter lognormal distribution is appropriate for the NO data, which has the highest acceptance rate by the GIC criterion, namely 8 from 10. The RRMSE for the two-parameter lognormal distribution is lowest with respect to the upper percentile range, although the three-parameter lognormal has lower errors for the maximum percentile. Of the remaining candidates, the three-parameter gamma distribution is very competitive as compared with the two-parameter lognormal distribution, having an acceptance rate of 7 from 10.

If the secondary factors in the model selection procedure are considered, then selection of the two-parameter lognormal distribution is reinforced. First, the misspecification results of Tables 9.1 and 9.3 indicate that RRMSE values in fitting the upper percentiles of a true three-parameter gamma distribution using either a two-parameter lognormal or a three-parameter gamma distribution are not substantially different in absolute terms. However, the errors in fitting the percentiles of a true two-parameter Weibull distribution with a three-parameter gamma can be very large. Second, misspecification results in this chapter calculated for deviations from observed values suggest that the errors for the upper percentiles for the two-parameter lognormal distribution in Table A2.3 are as expected if the NO data were taken from a lognormal distribution.

Results for the NO_2 data collected at the Museum site are reported in Tables A2.4

to A2.6. All criteria except FPE choose the two-parameter gamma as being superior to the other distributions, the acceptance rate by GIC, for example, being 6 from 8. The RRMSE values over different upper percentiles are also the lowest for this distribution. Unfortunately, the errors for the upper percentiles are not as low as those obtained if the data were taken from a two-parameter gamma distribution.

For NO_x data at the Museum site, the two-parameter gamma and lognormal distributions perform quite well, both having the same acceptance rate (8 from 10) using GIC, as shown in Table A2.8. GIC also accepts the three-parameter gamma distribution with high frequency. Comparing the RRMSE values in Table A2.9, the two-parameter lognormal has much lower errors than the two- and three-parameter gamma distributions. Misspecification results for the level of skewness encountered in these data sets can also be used to assist in the selection of the appropriate model. The comments for NO also apply in this case. Thus, it is safest to specify the two-parameter lognormal distribution. The two-parameter lognormal seems to be, therefore, the appropriate distribution for the NO_x data at the Museum site.

Tables A2.10 to A2.12 list the results for annual SO_2 data sets at the Museum site. From Table A2.10, the sample sizes over different years can be quite low, so the results should be interpreted with caution. Of the six distributions, the three-parameter gamma distribution performs the best according to GIC, KL, AIC and SIC. However, the two-parameter gamma distribution is preferable if choosing on the basis of the lowest RRMSE values.

Tables A2.13 to A2.18 provide a summary of the NO, NO_2 and NO_x pollutants at multiple sites. For NO, the GIC criterion favours the three-parameter gamma distribution most frequently. However, the two-parameter lognormal, which was selected to represent this pollutant at the Museum site, is favoured by most of the other discrimination criteria. Consideration of the RRMSE values in Table A2.14 indicate that no distribution is entirely satisfactory, the two-parameter Weibull possibly being preferred in that its errors for the maximum percentile are the lowest and are second

lowest for the upper percentiles. At this stage, therefore, none of these distributions can be considered as adequate for NO concentrations at all sites.

For the NO_2 data sets at the multiple sites, the two-parameter gamma distribution is clearly favoured by GIC and all discrimination criteria. The UPE and FPE are inconsistent with respect to the other tests. In terms of RRMSE values, the two-parameter gamma yields acceptable errors in fitting the upper and maximum percentiles, but the two-parameter gamma is generally outperformed by the three-parameter distributions. This is to be expected since the results of the misspecification analysis suggest that, if the underlying distribution is the two-parameter gamma, then all the three-parameter distributions can fit the observed upper percentiles well. Thus, for annual NO_2 data sets, the distribution previously selected as appropriate for the Museum site is also suitable for the other sites.

For the NO_x data sets at multiple sites, the GIC procedure marginally favours the three-parameter gamma over the two-parameter lognormal distribution, the latter being the distribution selected most frequently by the other tests. The GIC criterion also accepts the two-parameter gamma distribution with high frequency. The RRMSE values for the upper percentiles also highlight the superiority of the two-parameter lognormal distribution. However, the RRMSE values for the maximum percentile indicate that large errors can occur for some data sets, ranging up to 90 per cent. If required to select a single distribution to represent all the sites, the two-parameter gamma distribution would seem to be the most suitable. The two-parameter gamma distribution is similar to the two-parameter lognormal at the Museum site, its frequency of selection by GIC is high, and historically it has not led to errors of more than 40 per cent in reproducing the upper and maximum percentiles.

10.6 Concluding Remarks

In this chapter, new estimation and discrimination procedures developed in the thesis have been applied to air pollution data from Melbourne, Australia. The emphasis was

placed on how to select an appropriate distribution for a given pollutant over different years and, whenever possible, over different sites. An important issue addressed here is the need for compromise between using discrimination criteria and analysing RRMSE values in fitting percentiles.

These two factors do not always lead to identical choices. Sometimes it may be helpful to consider secondary factors based upon errors of misspecification. If the outcome of considering all factors is not clear, the final decision has to depend on the modelling objectives. For air quality management, the prediction of certain upper percentiles is the major concern. However, in applying the hybrid deterministic-statistical modelling approach, the overall distribution and particular calibration points, usually the high frequency percentiles, require reasonable prediction. Hence, likelihood-based discrimination is also important. If a distribution performs quite well in both cases, as with the three-parameter lognormal distribution for the β-scattering data, the appropriate distribution will be selected without any difficulty. Otherwise, the first consideration is the prediction of the relevant upper percentiles which should not be allowed to yield deviations outside a tolerable range. The RRMSE values can suggest the appropriate underlying distribution. Through using this selection rule, certain distributions can generally be determined as being optimal for a certain pollutant over different years and/or sites.

However, the results for some of the pollutants investigated in this chapter were not sufficiently convincing to lead to selection of a particular empirical distribution with reasonable confidence. Small samples may lead to such difficulties. In many cases, however, it is possible that the conditions for fitting probability distributions to data are violated. For example, it is well known that NO_2 time series of concentrations show strong seasonal trends and, hence, non-stationarity. It should always be checked whether samples are stationary and serially uncorrelated before a particular distributional form is accepted as an adequate description of air pollutant concentrations. Chapter 11 illustrates methods of investigating these assumptions.

TABLE 10.1

Descriptive statistics for RRMSE values from the 98 to maximum percentiles over 1000 Monte Carlo simulations with sample size $n = 365$ for the gamma distribution ($\beta = 1$)

GAMMA DISTRIBUTION							
True Values	Statistics for RRMSE	Six Estimated Models G3	G2	LN3	LN2	W3	W2
G3 α=2.0 γ=1.0	MEAN	0.074	0.114	0.159	0.078	0.072	0.139
	SD	0.033	0.045	0.066	0.033	0.032	0.037
	MIN	0.013	0.020	0.025	0.013	0.015	0.033
	MAX	0.206	0.271	0.401	0.209	0.202	0.258
G2 α=2.0 γ=0.0	MEAN	0.085	0.084	0.181	0.610	0.081	0.090
	SD	0.038	0.037	0.076	0.169	0.035	0.039
	MIN	0.015	0.014	0.028	0.172	0.016	0.013
	MAX	0.251	0.228	0.468	1.197	0.221	0.231
G3 α=6.0 γ=1.0	MEAN	0.052	0.054	0.059	0.093	0.059	0.091
	SD	0.022	0.024	0.026	0.040	0.027	0.028
	MIN	0.010	0.008	0.010	0.016	0.009	0.010
	MAX	0.144	0.155	0.211	0.234	0.160	0.172
G2 α=6.0 γ=0.0	MEAN	0.056	0.057	0.063	0.168	0.063	0.087
	SD	0.023	0.025	0.028	0.058	0.028	0.031
	MIN	0.011	0.008	0.010	0.029	0.010	0.014
	MAX	0.156	0.154	0.227	0.353	0.168	0.176

Note: α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the three- and two-parameter gamma distributions, W3 and W2 the three- and two-parameter Weibull distributions, and LN3 and LN2 the three- and two-parameter lognormal distributions. SD is the standard deviation of RRMSE, and MIN and MAX are the minimum and maximum values of RRMSE over 1000 simulations

TABLE 10.2
Descriptive statistics for RRMSE values from the 98 to maximum percentiles over 1000 Monte Carlo simulations with sample size $n = 365$ for the Weibull distribution ($\beta = 1$)

WEIBULL DISTRIBUTION							
True Values	Statistics for RRMSE	Six Estimated Models G3	G2	LN3	LN2	W3	W2
W3 α=2.0 γ=1.0	MEAN	0.063	0.037	0.069	0.052	0.033	0.064
	SD	0.026	0.016	0.027	0.024	0.015	0.020
	MIN	0.009	0.006	0.009	0.012	0.007	0.006
	MAX	0.169	0.113	0.178	0.137	0.097	0.132
W2 α=2.0 γ=0.0	MEAN	0.092	0.171	0.100	0.668	0.049	0.049
	SD	0.039	0.056	0.040	0.140	0.021	0.022
	MIN	0.014	0.026	0.013	0.325	0.010	0.009
	MAX	0.254	0.347	0.263	1.316	0.142	0.137
W3 α=4.0 γ=1.0	MEAN	0.569	0.047	0.032	0.066	0.014	0.020
	SD	1.844	0.015	0.247	0.017	0.006	0.009
	MIN	0.006	0.008	0.005	0.015	0.002	0.003
	MAX	14.330	0.091	7.808	0.113	0.043	0.053
W3 α=4.0 γ=0.0	MEAN	0.956	0.155	0.054	0.286	0.023	0.024
	SD	3.072	0.034	0.419	0.053	0.010	0.011
	MIN	0.009	0.060	0.008	0.150	0.003	0.005
	MAX	24.160	0.266	13.263	0.516	0.072	0.068

Note: α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the three- and two-parameter gamma distributions, W3 and W2 the three- and two-parameter Weibull distributions, and LN3 and LN2 the three- and two-parameter lognormal distributions. SD is the standard deviation of RRMSE, and MIN and MAX are the minimum and maximum values of RRMSE over 1000 simulations

TABLE 10.3

Descriptive statistics for RRMSE values from the 98 to maximum percentiles over 1000 Monte Carlo simulations with sample size $n = 365$ for the lognormal distribution ($\beta = 1$)

LOGNORMAL DISTRIBUTION							
True Values	Statistics for RRMSE	Six Estimated Models G3	G2	LN3	LN2	W3	W2
LN3 α=0.9 γ=1.0	MEAN	0.258	0.338	0.167	0.246	0.233	0.301
	SD	0.078	0.067	0.077	0.086	0.069	0.055
	MIN	0.038	0.129	0.033	0.044	0.035	0.125
	MAX	0.492	0.536	0.500	0.533	0.463	0.472
LN2 α=0.9 γ=0.0	MEAN	0.267	0.301	0.174	0.175	0.240	0.275
	SD	0.079	0.075	0.081	0.082	0.070	0.065
	MIN	0.040	0.069	0.035	0.023	0.037	0.070
	MAX	0.497	0.522	0.521	0.533	0.468	0.476
LN3 α=0.5 γ=1.0	MEAN	0.097	0.160	0.083	0.110	0.119	0.168
	SD	0.048	0.052	0.037	0.051	0.046	0.037
	MIN	0.018	0.023	0.018	0.019	0.015	0.058
	MAX	0.319	0.358	0.263	0.335	0.306	0.313
LN2 α=0.5 γ=0.0	MEAN	0.106	0.144	0.091	0.096	0.131	0.168
	SD	0.051	0.057	0.040	0.044	0.049	0.044
	MIN	0.020	0.023	0.020	0.013	0.018	0.040
	MAX	0.337	0.361	0.277	0.297	0.323	0.331

Note: α is the shape parameter, β the scale parameter and γ the location parameter. G3 and G2 denote the three- and two-parameter gamma distributions, W3 and W2 the three- and two-parameter Weibull distributions, and LN3 and LN2 the three- and two-parameter lognormal distributions. SD is the standard deviation of RRMSE, and MIN and MAX are the minimum and maximum values of RRMSE over 1000 simulations

TABLE 10.4

Estimated parameter values, maximized log-likelihood values and RRMSE values for different percentiles fitted to CO samples over different years at the Museum site

Year	n	Estimated Model	Shape α	Scale β	Location γ	Max(log L)	RRMSE for different percentiles				
							MAX1	MAX2	98	MEAN-U	MEAN-F
1975	102	G3	3.61	1.39	-0.55	-233.37	0.153	0.008	0.040	0.058	0.801
		G2	2.46	1.81	0.00	-235.78	0.258	0.066	0.012	0.101	1.091
		W3	1.76	5.09	-0.08	-234.85	0.096	0.034	0.087	0.038	0.623
		W2	1.71	4.98	0.00	-235.14	0.110	0.024	0.079	0.043	0.733
		LN3	0.41	1.76	-1.85	-232.83	0.240	0.034	0.047	0.092	0.843
		LN2	0.78	1.28	0.00	-249.55	1.246	0.649	0.426	0.531	1.361
1976	280	G3	37.12	0.19	-3.06	-428.51	0.153	0.095	0.009	0.100	0.030
		G2	10.68	0.36	0.00	-431.22	0.104	0.055	0.026	0.069	0.049
		W3	2.88	3.42	0.76	-430.77	0.178	0.109	0.018	0.109	0.034
		W2	3.59	4.22	0.00	-433.89	0.195	0.121	0.029	0.119	0.041
		LN3	0.11	2.29	-6.16	-428.44	0.150	0.093	0.008	0.099	0.030
		LN2	0.32	1.29	0.00	-439.02	0.007	0.016	0.089	0.024	0.069
1977	221	G3	3.65	0.94	-0.21	-421.40	0.139	0.003	0.022	0.072	0.085
		G2	3.09	1.04	0.00	-421.92	0.172	0.020	0.002	0.077	0.121
		W3	1.85	3.55	0.08	-422.26	0.050	0.053	0.058	0.070	0.082
		W2	1.91	3.65	0.00	-422.63	0.039	0.061	0.064	0.073	0.071
		LN3	0.39	1.44	-1.30	-422.13	0.208	0.030	0.000	0.089	0.063
		LN2	0.63	1.00	0.00	-432.83	0.646	0.302	0.217	0.277	0.227
1978	160	G3	3.67	0.72	-0.41	-262.48	0.216	0.199	0.030	0.114	0.232
		G2	2.15	1.04	0.00	-266.08	0.114	0.114	0.060	0.061	0.422
		W3	1.67	2.54	-0.04	-263.42	0.240	0.208	0.033	0.123	0.239
		W2	1.61	2.48	0.00	-263.65	0.229	0.199	0.024	0.117	0.297
		LN3	0.37	1.21	-1.36	-262.06	0.184	0.184	0.021	0.101	0.261
		LN2	0.85	0.55	0.00	-288.13	0.699	0.447	0.600	0.373	0.598
1979	223	G3	3.38	0.72	-0.05	-354.98	0.090	0.058	0.043	0.062	0.073
		G2	3.20	0.74	0.00	-355.03	0.101	0.051	0.037	0.061	0.080
		W3	1.76	2.50	0.16	-355.76	0.015	0.099	0.073	0.070	0.073
		W2	1.92	2.69	0.00	-357.39	0.013	0.117	0.088	0.079	0.070
		LN3	0.41	1.07	-0.78	-355.64	0.166	0.021	0.018	0.073	0.068
		LN2	0.61	0.70	0.00	-362.67	0.495	0.179	0.147	0.205	0.158

TABLE 10.4 continued

Year	n	Estimated Model	Shape α	Scale β	Location γ	Max(log L)	RRMSE for different percentiles				
							MAX1	MAX2	98	MEAN-U	MEAN-F
1980	145	G3	2.34	0.56	-0.03	-159.55	0.014	0.155	0.024	0.059	0.112
		G2	2.19	0.58	0.00	-159.65	0.002	0.144	0.036	0.054	0.133
		W3	1.54	1.39	0.02	-159.56	0.086	0.196	0.012	0.082	0.094
		W2	1.58	1.42	0.00	-159.82	0.099	0.204	0.020	0.088	0.068
		LN3	0.49	0.41	-0.42	-160.74	0.071	0.117	0.053	0.052	0.080
		LN2	0.78	0.00	0.00	-168.55	0.625	0.214	0.378	0.280	0.288
1981	290	G3	1.46	0.74	0.02	-302.36	0.246	0.139	0.054	0.183	0.117
		G2	1.61	0.69	0.00	-303.06	0.269	0.162	0.029	0.200	0.094
		W3	1.24	1.15	0.03	-301.56	0.290	0.174	0.021	0.207	0.101
		W2	1.32	1.21	0.00	-303.89	0.319	0.200	0.006	0.226	0.057
		LN3	0.66	0.06	-0.20	-308.58	0.104	0.050	0.126	0.130	0.117
		LN2	0.92	-0.24	0.00	-318.27	0.347	0.299	0.468	0.180	0.274
1982	301	G3	3.21	0.60	-0.22	-414.93	0.118	0.218	0.038	0.178	0.104
		G2	2.21	0.77	0.00	-417.51	0.041	0.163	0.102	0.135	0.198
		W3	1.65	1.90	0.00	-413.32	0.168	0.245	0.012	0.196	0.118
		W2	1.64	1.90	0.00	-413.32	0.168	0.245	0.013	0.196	0.118
		LN3	0.37	0.95	-1.06	-416.35	0.095	0.212	0.038	0.177	0.173
		LN2	0.80	0.29	0.00	-447.71	0.674	0.265	0.552	0.289	0.358
1983	337	G3	5.16	0.44	-0.43	-454.87	0.192	0.062	0.000	0.143	0.073
		G2	3.00	0.61	0.00	-458.32	0.119	0.006	0.063	0.097	0.120
		W3	1.90	2.04	0.03	-453.45	0.237	0.093	0.022	0.160	0.058
		W2	1.94	2.08	0.00	-453.54	0.241	0.097	0.026	0.163	0.047
		LN3	0.29	1.17	-1.51	-455.76	0.181	0.058	0.000	0.142	0.118
		LN2	0.66	0.43	0.00	-484.26	0.308	0.344	0.349	0.199	0.227
1984	85	G3	6.53	0.29	-0.48	-91.39	0.110	0.058	0.011	0.047	0.247
		G2	3.14	0.46	0.00	-92.83	0.238	0.155	0.090	0.107	0.293
		W3	2.04	1.63	-0.01	-90.96	0.047	0.021	0.014	0.019	0.230
		W2	2.02	1.62	0.00	-90.96	0.050	0.023	0.012	0.021	0.234
		LN3	0.26	0.99	-1.36	-91.60	0.127	0.065	0.014	0.054	0.241
		LN2	0.65	0.19	0.00	-100.16	0.839	0.546	0.389	0.378	0.351

TABLE 10.5

Summary of tests and criteria applied to CO samples by six distributions over different years at the Museum site

Selection Set	Estimated Model	Tests and Criteria							
		GIC	KL	AIC	SIC	CHI	KS	UPE	FPE
2- and 3-Parameter Models	G3	1	0	0	0	1	1	0	0
	G2	7	2	4	5	7	9	3	0
	W3	6	4	1	0	4	6	2	1
	W2	8	4	5	5	8	10	5	9
	LN3	0	0	0	0	0	0	0	0
	LN2	0	0	0	0	3	6	0	0
3-Parameter Models	G3	1	0	0	0	1	1	0	0
	W3	6	6	6	6	4	6	6	6
	LN3	0	0	0	0	0	0	0	0
2-Parameter Models	G2	7	5	5	5	7	9	4	0
	W2	8	5	5	5	8	10	6	10
	LN2	0	0	0	0	3	6	0	0

Table 10.6

Summary of RRMSE values for different percentiles fitted to CO samples over different years at the Museum site

Estimated Model	Upper Percentiles				Full Percentiles				Maximum Percentile			
	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX
G3	0.101	0.048	0.047	0.183	0.187	0.215	0.030	0.801	0.143	0.063	0.014	0.246
G2	0.096	0.042	0.054	0.200	0.260	0.297	0.049	1.091	0.142	0.086	0.002	0.269
W3	0.108	0.061	0.019	0.207	0.165	0.166	0.034	0.623	0.141	0.090	0.015	0.290
W2	0.112	0.063	0.021	0.226	0.174	0.204	0.041	0.733	0.146	0.095	0.013	0.319
LN3	0.101	0.037	0.052	0.177	0.199	0.226	0.030	0.843	0.153	0.051	0.071	0.240
LN2	0.274	0.130	0.024	0.531	0.391	0.350	0.069	1.361	0.588	0.317	0.007	1.246

TABLE 10.7

Estimated parameter values, maximized log-likelihood values and RRMSE values for different percentiles fitted to β-scattering samples over different years at the Museum site

Year	n	Estimated Model	Shape α	Scale β	Location γ	Max(log L)	RRMSE for different percentiles				
							MAX1	MAX2	98	MEAN-U	MEAN-F
1975	156	G3	1.62	4.88	2.32	-468.58	0.245	0.246	0.278	0.162	0.125
		G2	2.90	3.52	0.00	-481.07	0.311	0.298	0.292	0.197	0.178
		W3	1.20	8.38	2.37	-472.66	0.217	0.214	0.213	0.141	0.147
		W2	1.53	11.47	0.00	-494.82	0.272	0.249	0.237	0.166	0.268
		LN3	0.73	1.84	1.84	-458.89	0.030	0.121	0.160	0.078	0.074
		LN2	0.56	2.14	0.00	-464.68	0.206	0.241	0.256	0.155	0.104
1976	311	G3	2.22	3.09	2.25	-864.30	0.392	0.422	0.019	0.295	0.066
		G2	4.26	2.14	0.00	-877.81	0.444	0.462	0.077	0.330	0.101
		W3	1.41	7.44	2.40	-874.19	0.394	0.416	0.002	0.287	0.094
		W2	1.86	10.33	0.00	-910.51	0.427	0.435	0.024	0.302	0.206
		LN3	0.57	1.89	1.25	-855.78	0.292	0.366	0.048	0.249	0.042
		LN2	0.47	2.09	0.00	-858.80	0.364	0.414	0.018	0.291	0.053
1977	251	G3	3.54	1.88	1.42	-647.61	0.203	0.110	0.049	0.149	0.049
		G2	5.50	1.47	0.00	-650.48	0.243	0.145	0.082	0.175	0.057
		W3	1.74	6.94	1.91	-654.76	0.234	0.126	0.057	0.158	0.075
		W2	2.26	9.13	0.00	-670.21	0.271	0.155	0.081	0.178	0.145
		LN3	0.43	1.99	0.01	-644.62	0.127	0.058	0.009	0.117	0.039
		LN2	0.43	1.99	0.00	-644.62	0.128	0.058	0.009	0.117	0.039
1978	215	G3	1.24	6.51	2.35	-661.40	0.310	0.425	0.260	0.283	0.138
		G2	2.24	4.65	0.00	-686.77	0.379	0.471	0.311	0.320	0.223
		W3	1.04	8.24	2.36	-663.67	0.265	0.391	0.219	0.256	0.141
		W2	1.35	11.54	0.00	-701.96	0.331	0.425	0.250	0.283	0.295
		LN3	0.87	1.71	2.10	-644.42	0.015	0.283	0.134	0.187	0.055
		LN2	0.63	2.11	0.00	-659.12	0.270	0.425	0.277	0.285	0.125
1979	257	G3	2.46	2.39	2.59	-666.28	0.372	0.340	0.111	0.261	0.076
		G2	5.32	1.59	0.00	-682.42	0.421	0.379	0.155	0.293	0.114
		W3	1.47	6.45	2.69	-678.33	0.367	0.325	0.085	0.247	0.105
		W2	2.02	9.61	0.00	-716.00	0.395	0.341	0.097	0.258	0.218
		LN3	0.57	1.70	1.99	-655.32	0.269	0.276	0.050	0.212	0.046
		LN2	0.42	2.04	0.00	-663.19	0.368	0.348	0.128	0.269	0.072

TABLE 10.7 continued

Year	n	Estimated Model	Shape α	Scale β	Location γ	Max(log L)	RRMSE for different percentiles				
							MAX1	MAX2	98	MEAN-U	MEAN-F
1980	199	G3	1.68	2.72	2.28	-487.32	0.457	0.301	0.202	0.283	0.082
		G2	4.12	1.66	0.00	-507.07	0.519	0.364	0.264	0.324	0.132
		W3	1.22	4.87	2.31	-492.49	0.441	0.276	0.172	0.267	0.102
		W2	1.73	7.73	0.00	-533.43	0.477	0.300	0.186	0.281	0.247
		LN3	0.68	1.38	1.79	-478.81	0.338	0.213	0.134	0.226	0.053
		LN2	0.46	1.80	0.00	-487.36	0.473	0.334	0.244	0.304	0.079
1981	277	G3	0.76	3.94	2.00	-573.93	0.268	0.220	0.150	0.140	0.101
		G2	3.90	1.28	0.00	-625.11	0.518	0.449	0.156	0.338	0.113
		W3	0.89	2.87	2.00	-578.40	0.262	0.219	0.147	0.140	0.087
		W2	1.75	5.66	0.00	-655.60	0.493	0.411	0.092	0.299	0.210
		LN3	0.81	0.92	1.57	-587.44	0.203	0.217	0.118	0.141	0.039
		LN2	0.49	1.48	0.00	-602.63	0.456	0.408	0.114	0.307	0.064
1982	272	G3	0.97	2.98	2.00	-559.43	0.005	0.121	0.045	0.133	0.039
		G2	3.88	1.26	0.00	-607.69	0.246	0.298	0.141	0.256	0.140
		W3	0.99	2.86	2.00	-559.50	0.008	0.123	0.042	0.134	0.038
		W2	1.82	5.53	0.00	-632.68	0.237	0.275	0.105	0.234	0.216
		LN3	0.91	0.75	1.77	-563.11	0.423	0.100	0.231	0.172	0.058
		LN2	0.49	1.45	0.00	-586.93	0.142	0.241	0.094	0.220	0.089
1983	324	G3	1.78	2.24	1.95	-745.70	0.400	0.475	0.178	0.349	0.080
		G2	4.44	1.34	0.00	-770.22	0.473	0.526	0.247	0.397	0.118
		W3	1.25	4.23	2.03	-752.31	0.390	0.462	0.155	0.334	0.100
		W2	1.82	6.73	0.00	-812.05	0.437	0.486	0.179	0.352	0.233
		LN3	0.61	1.35	1.25	-735.19	0.316	0.436	0.140	0.319	0.054
		LN2	0.45	1.67	0.00	-742.98	0.418	0.499	0.220	0.376	0.071
1984	89	G3	1.64	1.07	3.09	-133.65	0.167	0.053	0.011	0.066	0.049
		G2	13.85	0.35	0.00	-147.64	0.012	0.070	0.107	0.027	0.088
		W3	1.25	1.86	3.12	-134.20	0.147	0.046	0.013	0.058	0.052
		W2	3.31	5.38	0.00	-161.24	0.015	0.053	0.083	0.021	0.155
		LN3	0.64	0.51	2.79	-132.64	0.341	0.141	0.045	0.140	0.047
		LN2	0.26	1.54	0.00	-143.18	0.022	0.053	0.097	0.022	0.075

TABLE 10.8

Summary of tests and criteria applied to β-scattering samples by six distributions over different years at the Museum site

Selection Set	Estimated Model	Tests and Criteria GIC	KL	AIC	SIC	CHI	KS	UPE	FPE
2- and 3-Parameter Models	G3	3	2	2	2	4	5	1	0
	G2	0	0	0	0	1	1	0	0
	W3	2	0	0	0	1	4	0	1
	W2	0	0	0	0	0	0	0	0
	LN3	8	8	7	7	10	10	8	8
	LN2	1	0	1	1	6	6	1	1
3-Parameter Models	G3	3	2	2	2	4	5	1	0
	W3	2	0	0	0	1	4	1	1
	LN3	8	8	8	8	10	10	8	9
2-Parameter Models	G2	0	0	0	0	1	1	0	0
	W2	0	0	0	0	0	0	1	0
	LN2	10	10	10	10	6	6	9	10

Table 10.9

Summary of RRMSE values for different percentiles fitted to β-scattering samples over different years at the Museum site

Estimated Model	Upper Percentile MEAN	SD	MIN	MAX	Full Percentile MEAN	SD	MIN	MAX	Maximum Percentile MEAN	SD	MIN	MAX
G3	0.212	0.088	0.066	0.349	0.081	0.031	0.039	0.138	0.282	0.128	0.005	0.457
G2	0.266	0.102	0.027	0.397	0.127	0.044	0.057	0.223	0.357	0.150	0.012	0.519
W3	0.202	0.083	0.058	0.334	0.094	0.032	0.038	0.147	0.273	0.125	0.008	0.441
W2	0.237	0.090	0.021	0.352	0.219	0.044	0.145	0.295	0.335	0.137	0.015	0.493
LN3	0.184	0.067	0.078	0.319	0.051	0.010	0.039	0.074	0.236	0.131	0.015	0.423
LN2	0.235	0.101	0.022	0.376	0.077	0.023	0.039	0.125	0.285	0.147	0.022	0.473

TABLE 10.10

Summary of tests and criteria applied to CO samples by six distributions over different years and sites with 20 annual data sets

Selection Set	Estimated Model	Tests and Criteria							
		GIC	KL	AIC	SIC	CHI	KS	UPE	FPE
2- and 3-Parameter Models	G3	3	2	1	1	1	1	0	0
	G2	13	4	6	7	13	18	4	2
	W3	10	6	1	0	7	10	2	1
	W2	16	8	12	12	15	19	14	17
	LN3	0	0	0	0	0	0	0	0
	LN2	0	0	0	0	4	8	0	0
3-Parameter Models	G3	3	2	2	2	1	1	0	1
	W3	11	10	10	10	7	10	12	11
	LN3	0	0	0	0	0	0	0	0
2-Parameter Models	G2	14	8	8	8	13	18	5	2
	W2	17	12	12	12	15	19	15	18
	LN2	0	0	0	0	4	8	0	0

TABLE 10.11

Summary of RRMSE values for different percentiles fitted to CO samples over different years and sites with 20 annual data sets

Estimated Model	Upper Percentile				Full Percentile				Maximum Percentile			
	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX
G3	0.145	0.115	0.047	0.580	0.214	0.188	0.030	0.801	0.256	0.274	0.009	1.277
G2	0.132	0.075	0.054	0.361	0.265	0.245	0.049	1.091	0.253	0.202	0.002	0.805
W3	0.129	0.075	0.019	0.300	0.168	0.132	0.034	0.623	0.207	0.178	0.015	0.671
W2	0.121	0.067	0.021	0.272	0.174	0.156	0.041	0.733	0.188	0.155	0.013	0.609
LN3	0.145	0.120	0.052	0.584	0.220	0.185	0.030	0.843	0.264	0.298	0.001	1.333
LN2	0.439	0.326	0.024	1.517	0.396	0.262	0.069	1.361	0.979	0.786	0.007	3.499

TABLE 10.12

Summary of tests and criteria applied to β-scattering samples by six distributions over different years and sites with 28 annual data sets

Selection Set	Estimated Model	Tests and Criteria							
		GIC	KL	AIC	SIC	CHI	KS	UPE	FPE
2- and 3-Parameter Models	G3	9	5	5	4	11	17	4	2
	G2	1	0	0	0	4	5	0	1
	W3	6	1	0	0	5	14	2	1
	W2	0	0	0	0	1	3	1	0
	LN3	25	22	21	21	23	25	19	23
	LN2	2	0	2	3	14	16	2	1
3-Parameter Models	G3	9	5	5	5	11	17	4	2
	W3	6	1	1	1	5	14	5	2
	LN3	25	22	22	22	23	25	19	24
2-Parameter Models	G2	1	0	0	0	4	5	0	1
	W2	0	0	0	0	1	3	8	0
	LN2	28	28	28	28	14	16	20	27

TABLE 10.13

Summary of RRMSE values for different percentiles fitted to β-scattering samples over different years and sites with 28 annual data sets

Estimated Model	Upper Percentile				Full Percentile				Maximum Percentile			
	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX
G3	0.209	0.122	0.006	0.437	0.077	0.033	0.027	0.138	0.266	0.164	0.005	0.618
G2	0.251	0.132	0.027	0.462	0.126	0.046	0.031	0.223	0.330	0.173	0.012	0.661
W3	0.197	0.117	0.002	0.401	0.088	0.035	0.031	0.150	0.254	0.160	0.002	0.597
W2	0.222	0.116	0.014	0.399	0.213	0.056	0.056	0.316	0.307	0.160	0.015	0.605
LN3	0.193	0.088	0.076	0.405	0.050	0.017	0.026	0.102	0.248	0.147	0.004	0.602
LN2	0.228	0.132	0.015	0.469	0.080	0.027	0.034	0.140	0.273	0.175	0.019	0.654

Chapter 11

Prediction of Non-stationary Seasonal Extremes of One-hour Average Urban CO Concentrations

11.1 Introduction

In urban planning, it is desirable to assess the air quality implications of land use options by linking the emissions that result from specific land uses to ensuing ambient concentrations of pollutants. Where motor vehicles are a dominant source of air pollutant concentrations, model predictions of the dispersion of emissions from road networks are needed. The resultant concentrations can be compared with air quality goals. This requires predictions of potential annual average levels or short-term extremes or both. In the case of carbon monoxide, which is the prime consideration of this chapter and a significant pollutant in the city of our case study, Canberra, it is the 1-hour average and 8-hour average annual maxima which are of interest. These averaging times and maxima relate to air quality goals used by the planning authorities in Canberra and to air quality standards used by other agencies such as the United States Environmental Protection Agency (EPA). The prediction method developed in this chapter is illustrated for the 1-hour averaging time and the maximum extreme value only. However, it applies equally to the 8-hour case and other extremes which can be related to high percentile properties of the frequency distribution of concentrations. Air pollutant data sets from 1982-1987 are available at Canberra's central

monitoring site, Civic.

To provide the air quality advice required for the planning of urban areas, a hybrid modelling approach is described which combines a deterministic model with statistical techniques for estimating extremes from the deterministic model output and analysis of historical pollutant concentrations. The particular method applied is a very simple but useful implementation of the general hybrid approach of Jakeman et al. (1988). It uses only a basic deterministic model with static emission and meteorological input. While its predictive performance is considered satisfactory in this chapter, the deterministic model may be inadequate for some applications. However, the general hybrid approach does allow for the adoption of any suitable deterministic model.

In order to make the prediction problem above more tractable, the objective of the modelling exercise in this chapter is narrowed from prediction of the annual frequency distribution of all 1-hour concentrations of CO to the seasonal frequency distribution of 1-hour concentrations for separate hours of the day within which the maximum annual concentration has been reported. For our case study in Canberra, this maximum has always occurred in the evening, and from 1800-1900 predominantly. While this is the hour for which our method is illustrated in this chapter, the annual maximum did occur between 1900-2000 hours in 1985 and 1987. In practice the method should be applied to this time of the day as well although this was not carried out for the chapter. The frequencies are estimated over each winter period, taken as April to September inclusive for the southern hemisphere. It is within the winter months which characteristically contain more stable meteorological conditions that the maximum has always been observed to occur. Figure 11.1 shows a plot of the diurnal variation of 1-hour CO concentration at the Civic site averaged over the 1987 winter season. Other years show the same pattern with early evening peaks in CO concentrations.

This simplified treatment increases the validity of one of the main assumptions of our approach delineated in the next section: within a period of predictive interest (taken here as the winter season), the probability density function (pdf) from which

the pollutant observations are derived can be characterised by an efficient parameterisation. A sufficient condition for this to apply is that the ambient observations are independent and identically distributed according to some pdf with a small number of parameters. By systematically isolating particular hours of the day instead of analysing all hours of the day together, problems of autocorrelation among observations and short-term cyclical trends caused by diurnal variation are minimised. By selecting a seasonal subset of the annual period, problems of nonstationarity in the observations caused by substantial variations in meteorology between summer and winter are minimised. Before outlining the plan of the remainder of the chapter, some discussion and quantification of these key assumptions is appropriate.

Figure 11.2 shows the time series of evening peak hour concentrations of CO measured in 1987 at the Civic site. The time series plot is typical of other years in that concentrations are generally higher in winter than summer. However, within season, the data can be more strongly regarded as emanating from a stationary process. Note that the time series for 1987 represents samples which, of all years, conform least to the independent and identically distributed assumption. Table 11.1 provides, for each year of available data from the Civic site, a summary of the autocorrelation of observations obtained during the evening peak hour of each winter. The maximum first lag autocorrelation which occurs in 1987 is 0.452 but the median value over all years is 0.290. It has been demonstrated by Taylor (1985), using samples which are lognormally distributed, that the additional effect of such levels of autocorrelation on pdf fitting errors will be small. Table 11.2 provides for each year an indication of the stationarity of the underlying processes generating the time series data. For both the entire year and the winter season of evening peak hour values, the table gives the ratio of two variances. The denominator is the variance about the arithmetic mean while the numerator is the variance about a simple uniformly weighted moving average over 20 time steps (the results over 40 time steps are very similar). The ratio is substantially larger for each winter period than for the entire corresponding year, reflecting stronger stationarity for the winter period. The difference in variance about the constant mean and moving

average mean is highest in 1985 and this difference is closely followed in 1987. While the effect of discrepancies of this level on ensuing predictions of extremes needs to be investigated, this analysis of the modest amount of historical data available suggests that the stationarity assumption can be supported in most years.

An outline of the chapter is as follows. In the next section, the steps and assumptions in the general hybrid deterministic-statistical distribution approach of Jakeman et al. (1988) are reviewed. In the following section, details are provided of the deterministic model and data inputs used to predict mean seasonal concentrations of CO. Then the relevance of assuming a particular parametric form for the seasonal distribution of evening peak hour CO concentrations is demonstrated. Goodness-of-fit tests and error performance criteria are applied to the six available winter data sets in order to discriminate among candidate parametric forms. This is followed by a simple implementation of the hybrid approach and an evaluation of the prediction results. The conclusions complete the chapter.

11.2 The Hybrid Approach

As outlined in Chapter 2, the hybrid approach described in Jakeman et al. (1988) was developed to combine the different advantages of deterministic and statistical distribution modelling techniques and to overcome their disadvantages. The hybrid approach achieves this by using a deterministic model to predict those concentrations which occur frequently (eg median or arithmetic mean), in conjunction with statistical techniques for identifying the parametric distributional form of the air pollutant data in question, and estimating extreme values of the appropriate distribution (eg maximum). A suitable deterministic model can make causal links between emissions, meteorology and mean ground level concentrations. Statistical distribution models can predict the frequency distribution of all events about the mean, once the distributional form appropriate can be identified or assumed from historical observations and its parameter values inferred from predictions of some of its properties by the deterministic model.

The hybrid approach can be summarised in the following four steps:

(1) identify, from a range of alternatives, the parametric form of the probability distribution function (pdf) of historical observations recorded within individual periods of predictive interest and assess its suitability over all periods (e.g. gamma, Weibull or lognormal distribution for each winter season of evening peak hour concentrations);

(2) select an appropriate deterministic model to predict properties of the pdf over each of the desired periods (e.g. seasonal means of 1-hour concentrations);

(3) fit, for each prediction period, the predictions of the deterministic model in (2) to the parametric form identified in (1) so that its parameters can be estimated, and;

(4) calculate the desired extreme(s) of the pdf for each period (e.g. maximum of evening peak hour concentrations each winter) using the estimated parameter values from (3), and if possible, estimate the variability of the extreme(s).

The major assumptions of the approach are:

(i) within each period the pdf of concentrations can be adequately and efficiently parameterised,

(ii) the parametric form of the pdf at a site of interest remains suitable from one period to another, and

(iii) the deterministic model yields sufficiently accurate estimates of the properties of the pdf to calibrate its parameters.

Assumptions (i) and (ii) can often be investigated by analysis of historical concentrations. Clearly, no guarantee can be made that the parametric form most applicable historically will not change. Therefore, the assumption (ii) means that the parametric form remains invariant to future changes in emissions and meteorology, only the parameter values of the assumed form vary.

Additional assumptions related to some of the pdf parameters of the statistical distribution model must be made if it has two or more parameters and a deterministic model cannot predict variability about the mean. This is the case with the deterministic model applied to the Canberra data in this chapter. Sometimes, however, there may

be sufficient model input information available to drive a deterministic model that some variability of the mean concentration can be adequately predicted. For example, Taylor et al. (1985) use the GM line source model (Chock, 1978) to produce estimates of the 30 to 70 percentile values of hourly average CO data recorded in Melbourne, Australia. This range of percentile values is then used to calibrate a two-parameter Weibull distribution which was selected as an appropriate distribution to represent the entire percentile range of concentrations.

Enhancement of the analytical tools required for construction of the statistical component of the hybrid approach have been the main concern of this thesis. For the estimation of parameters the new maximum likelihood algorithms constructed in Chapter 3 for the three-parameter gamma, Weibull and lognormal distributions are used. These methods are computationally simpler than traditional alternatives and allow fitting when the value of the parent shape parameter leads to exponential shaped distributions. For the identification exercise, this chapter will demonstrate the basic principles of the procedure developed in Chapter 12 on the basis of work on discrimination and misspecification in Chapters 6 to 9.

11.3 The Deterministic Model

The deterministic model applied in this chapter is POLDIF, which is a model that attempts to predict the dispersion of air and noise pollution from urban road networks. It has been developed and made commercially available by the CSIRO Division of Building Research (Anderson, 1985). This deterministic model uses a combination of traffic volumes, emission rates and meteorological conditions to determine ground level concentrations of pollutants for a given road network. The specific data required by the model include: date, latitude, time, cloud cover, wind speed and direction, inversion height, vehicle fleet compositions, link volumes and speeds, point source strengths and grid size for the receptor location. Table 11.3 contains the primary meteorological data used to produce the POLDIF estimates of CO.

With these data the model calculates the ground level concentration for air pollutants using principles of Gaussian plume dispersion (Turner, 1969). Thus, the concentration of a gas $\chi(g/m^3)$ at a point (x, y, z), emitted at a rate of $Q(g/s)$ and at a location and height (x', y', z'), is dispersed according to vertical (m) and horizontal (m) dispersion parameters and $U(m/s)$, the mean wind speed. For vehicle exhausts z can be taken as $1.2m$, the effective height for the mixing area in the wake of the vehicle. To reduce the computing time significantly, this model replaces the integrals of distances between the points (x, y, z) and (x', y', z') into exponential functions and error functions, which can be expressed as (Taylor and Anderson, 1982)

$$\chi = \frac{Q}{\sqrt{2\pi}\sigma_z U}\frac{(1+g^2)^{\frac{1}{2}}}{g} \quad exp \quad [\frac{-(z-z')^2}{2\sigma_z^2} - \frac{V(z-z')(x-x')}{U\sigma_z^2} - \frac{V^2(x-x')^2}{2U^2\sigma_z^2}] \cdot[erf\frac{(y-y')}{\sqrt{2}\sigma_y} - erf\frac{(y-y')}{\sqrt{2}\sigma_y}] \tag{11.1}$$

where g is the gradiant in the horizontal place of the line between the end points of integration and σ_y and σ_z are the standard deviations of pollutant concentration in the horizontal and vertical directions, respectively, and V is the terminal settling velocity of the pollutant at a specific location.

Each link within the network is treated as a line source, similar to the HIWAY model (Zimmerman and Thompson, 1975), where the concentration at any point is found by integrating over the line upstream of the receptor. The contribution of each link to each receptor point is calculated and the contributions of all links to each receptor are then added. The contributions of links at a distance greater than $3\sigma_y$ from any receptor are neglected.

The vehicle emissions are determined by link volumes, average link speeds and emission factors associated with vehicle fleet composition. The vehicle fleet composition has been determined for Canberra based on Australian Bureau of Statistics data for vehicle registrations. The fleet composition is divided into petrol, diesel, 6 cylinder, 4 cylinder, unleaded petrol and gas-fired vehicles. Although POLDIF is not particularly sensitive to small changes in fleet composition (Taylor and Anderson, 1984), it should

be noted that recent changes in Australian emission control regulations for new vehicles have not yet been included in the emission factors released with POLDIF in 1985.

Link volumes and speeds are determined by a local area traffic model, MULATM (Anderson and Taylor, 1988). MULATM predicts traffic movements based primarily on origin-destination data and intersection controls. The predicted link volumes and speeds have then been calibrated to reflect observed data within the Civic network during 1985-86. It must therefore be assumed that 1985-1986 volumes are not significantly different to other years for which the model has been run here (1985-1987).

The dispersion parameters are calculated in POLDIF using the angle of the sun and average cloud cover, to determine incident solar radiation, and average wind speed. Miles et al. (1989) calculate mean annual concentrations of CO in Canberra assuming that neutral atmospheric conditions (i.e. adiabatic lapse rate equals environmental lapse rate) best represents the average hour for any given year in Canberra. This is achieved by selecting an average length day (22 March or 22 August) to calculate incident solar radiation. For the calculation of mean seasonal concentrations of pollutants, the shortest day (22 June) and the longest day (22 December) could be used to produce more stable and unstable conditions, respectively. In this chapter, we use the former to predict mean winter concentrations for the evening peak hours.

Temperature inversions, which restrict the vertical distribution of air pollutants, are also taken into account by POLDIF. The model allows the vertical dispersion parameter to vary downwind until $\sigma_z = h_i/2.15$ where h_i is the inversion height (m). It is held constant from then on. The average inversion height varies seasonally in Canberra. From the limited data available it can be assumed that the mixing depth for Canberra in summer varies between 1500-2000 m (MacNicol, 1982). In winter, however, much lower, stronger, surface temperature inversions can occur during the night and early morning (Ferrari et al., 1986). The average winter inversion height has been assumed to be 80 m.

The results of applying POLDIF to the prediction of seasonal mean CO concen-

trations in Canberra for 1985-87 are reported in Table 11.4. The seasonal mean concentrations predicted by the POLDIF model for the peak hour are generally within a factor of two of the observed concentrations. The factor of two is often regarded as an acceptable level of accuracy for modelling air pollution given the limitations of modelling illustrated by Hanna (1982). However, there is a bias in that the model underpredicts the observed means by around 30 per cent in 1985 and 1987 and about 6 per cent in 1986. This is to be expected since no attempt has been made so far to include background concentrations before the 1800-1900 evening peak hour.

In order to estimate background concentrations, an approach used by Tiao et al. (1975) could be applied. It can be argued that CO concentration at time (hour) t is likely to be a function of immediate traffic emissions and traffic emissions in the recent past exponentially discounted because of dispersion with time. Thus, the effective emission factor Q_t^e at time t could be written

$$Q_t^e = Q_t + \delta Q_{t-1} + \delta^2 Q_{t-2} + \delta^3 Q_{t-3} + \cdots \tag{11.2}$$

where $0 \leq \delta \leq 1$.

In this way the effective concentration χ^e, which is background plus concentration due to emissions at time t only, could be predicted as a seasonal mean by replacing Q terms in the deterministic model (11.1) by Q_t^e to obtain

$$\chi_t^e = \chi_t + \delta \chi_{t-1} + \delta^2 \chi_{t-2} + \delta^3 \chi_{t-3} + \cdots \tag{11.3}$$

Use of equation (11.3) requires running POLDIF for hours prior to the evening peak hour (where $t = 1800 - 1900$) to obtain χ_{t-j}^e $(j = 1, 2, \ldots)$ and calibrating δ in (2) with the latter predictions of concentration and the observed concentrations which should be χ_t^e.

Lack of meteorological data prevents this exercise from being undertaken comprehensively here. Wind speed data were recorded only as 10 minute averages at 3 hourly intervals and the wind speed used for the 1800-1900 hour was recorded from 1755-

1805. Further evaluation of the deterministic model component would also be useful to appreciate the errors involved in predicting χ_{t-j} in (2) and hence in obtaining uncertainty bounds for the decay factor δ. We do not wish to dwell on these aspects here. The main purpose of the chapter is to show how a hybrid modelling approach can be used to infer extreme concentrations. The selection of a suitable deterministic model and its method of calibration will be undertaken in the same way, whether means or extremes are to be predicted. For the present purposes, we are satisfied with the POLDIF model performance as it is well within the limits of accuracy that could be achieved and as it underpredicts when background concentrations are not considered. However, we conclude this section with some assessment of the credibility of the level of underprediction.

Therefore, we assume for the moment that POLDIF accurately predicts the non-background concentration. Later in the chapter we consider the sensitivity of our results to this assumption when predicting the seasonal maximum concentration of evening peak hour CO. If this assumption were true, then background in 1985 and 1987 for the peak hour is 1.0 mg/m^3 on average. Note that it is lower in 1986 because of better dispersive conditions as indicated by the higher wind speeds in Table 11.3. Also, the average predicted χ_t $(t = 1800 - 1900)$ for these years is 2.3 mg/m^3. Further χ_t is our best estimate of χ_{t-1} since the major meteorological variable, wind speed, is measured during the last five minutes of the earlier interval $t-1$ and the first five minutes of the following interval t and traffic levels are very similar for both intervals. Hence from (11.3) where

$$\chi_t^e - \chi_t = \delta\chi_{t-1} + \delta^2\chi_{t-2} + \delta^3\chi_{t-3} + \cdots \tag{11.4}$$

the approximation

$$\delta = \frac{\chi_t^e - \chi_t}{\chi_t} \tag{11.5}$$

can be applied to obtain an upper bound for δ by neglecting higher order terms in δ

and setting $\chi_{t-1} = \chi_t$. The value of δ then for 1985 and 1987 is $1.0/2.3 = 0.43$ which can be compared with the value of 0.34 obtained by Tiao et al. (1975) for downtown Los Angeles using non-linear least squares estimation methods and a causal statistical model of carbon monoxide generation.

Note that the value of δ computed by (11.5) instead of (11.4) can be regarded as a background factor calculated by reference to the single previous hourly concentration prediction.

11.4 Parametric Form of Annual and Seasonal Distributions of Historical Concentrations

Since the POLDIF model and available input data do not directly allow generation of the variability of concentrations within a seasonal period, we now consider the parametric form of the distribution of evening peak hour CO concentrations over winter seasonal periods is now considered. For each of the calendar years 1982-1987, two- and three-parameter versions of the Weibull, gamma and lognormal distributions were fitted by maximum likelihood estimation to these historical pollutant data sets. As noted in previous chapters, these shape-scale-location parameter distributions are felt to cover a large range of possible forms from which air pollutant concentration data could be assumed to be derived.

The procedure followed to identify the most appropriate candidate from the selected set of distributional forms can be considered in two principal steps as given in Chapter 10. First, a parametric candidate can be eliminated from consideration if it leads to substantially inferior fitting of the quantities of predictive interest in historical data sets. Here these are the higher percentiles and especially the maximum. The errors of model fits can be calculated on historical data sets as deviations from observed or recorded values. Table 11.5 shows the relative root mean squared errors in predicting the observed evening peak hour maximum in each of the winter periods from 1982-1987. It shows the mean, standard deviation, minimum and maximum error for these

6 years. Note that the three three-parameter distributions and the two-parameter lognormal distribution lead to much larger errors than do the two-parameter gamma or Weibull distributions. The error in the worst case year for any of the former has been from 83 to 242 per cent, while the maximum error encountered for the latter is under 25 per cent. The mean and standard deviation of the rrmse from the observed maximum over all years are also substantially lower for the two-parameter gamma and Weibull distributions. Of these the gamma is clearly superior overall in predicting the observed maximum. The poor performance of the lognormal distribution is to be noted to help dispell any remaining traditional beliefs that lognormal distributions can generally be used to parameterise frequency distributions of air pollutant concentrations.

In addition to the high percentiles, the distribution chosen should be capable of fitting the higher frequency percentiles well since the hybrid modelling technique calibrates the parameters of a distribution with various properties of its pdf that are predicted by the deterministic model. In this case it is the mean, but in other cases there may be sufficient input information available for applying a deterministic model which yields reliable output over a range of mid-percentile values. In general, we use statistics based upon the likelihood value of a distribution to aid selection of an appropriate model. The likelihood is a measure of the overall fit, especially weighting higher frequency percentiles more. It also allows adoption of a formal discrimination approach based upon information criteria.

The second step in the identification procedure is therefore to calculate the likelihood of remaining candidates. Then the likelihood is inserted in a discrimination statistic or generalised information criterion which allows selection of the superior or not significantly inferior distributional candidate for the majority of data sets. The development of this generalised information criterion (GIC) was given in Chapters 6 and 8. Essentially, it allows evaluation of candidate distributional forms in terms of the likelihood plus a tolerance which varies according to the number of parameters in the distribution and the significance level required. In this chapter we have considered selection at the 98 per cent confidence level, although sensitivity to the level chosen

can easily be investigated. For non-nested model candidates, the GIC statistic, at the 98 per cent confidence level, accepts an alternative model i for a given set of samples within a period if

$$GIC = -2\log L_1 + 2.1089 > -2\log L_i \tag{11.6}$$

where $\log L_1$ is the highest value of the maximum likelihood over all parametric candidates and $\log L_i$ is the value of the maximum likelihood of alternative distributions.

Table 11.6 shows the results of applying GIC and other standard criteria to data sets of evening peak hour CO for the six winter periods from 1982-1987. Only the two-parameter gamma and Weibull distributions are considered as candidates, the other distributions being rejected after the first step. Both parametric forms perform adequately at this step in terms of GIC but the other results indicate the danger of applying some traditional statistical tests which are out of context with the specific objective of the discrimination exercise. The ML criterion merely selects the parametric form with the highest maximum likelihood value while the AIC (Akaike, 1974) and SIC (Schwarz, 1978) statistics adjust the likelihood value with a penalty for the number of parameters. The SIC also adjusts for sample size. All three statistics infer a strong superiority for gamma over Weibull. The chi-square (CHI) and the Kolmogorov-Smirnov (KS) tests (Kendall and Stuart, 1979) are both equally divided in their selection between gamma and Weibull distributions. The GIC finds that the gamma and Weibull distributions are not significantly different at the 98 per cent level in terms of likelihood. Consistent with step one, the result infers that use of either distribution is appropriate. The gamma is selected here because of its slightly better performance in fitting historical maxima.

11.5 A Hybrid Method and Prediction

One possible implementation of the hybrid approach is demonstrated for predicting the winter frequency distribution of evening peak hour CO concentrations. In selecting

a method it must be appreciated that the deterministic model POLDIF, when driven by the basic meteorological data available and when used in isolation from statistical models, predicts only mean values with reasonable accuracy. Because the distribution for CO is two parameter (i.e. with positive shape and scale and zero location value), two pieces of information are needed to describe it. Yet no variability of the mean provided by POLDIF is available as model output. In order to circumvent this difficulty, a simple strategy used in Jakeman et al. (in press) for predicting water quality extremes, where the concentrations conform to a two-parameter Weibull distribution, can be analogously invoked.

The strategy adopted is the following: use the POLDIF model to provide the mean of the gamma distribution; assume a range of values for its shape parameter; compute the scale parameter from the mean and shape; infer the percentile result of interest; and evaluate the sensitivity of the result both to values within the shape range and to errors in the deterministic prediction of the mean. Table 11.7 contains the results for the maximum concentration, the most difficult percentile of the distribution to estimate accurately. The results are reasonably insensitive to the variation in shape value as any of the shape values yield sufficiently accurate results by accepted standards (Hanna, 1982). Also, the range of shapes addressed in Table 11.7 covers those found from fitting two-parameter gamma distributions to the annual historical winter concentration data sets from 1982-1987. The shape parameter ranged from 0.84 to 1.31 in those fitted data sets. Its mean is 1.16, median is 1.21 and standard deviation is 0.15. The table gives the estimate of the winter maximum assuming the POLDIF model predicts the observed mean precisely. It also shows predicted maxima for errors of 30 per cent in the mean. The errors of 30 per cent in the mean were felt to represent the maximum errors that might be obtained by using POLDIF in conjunction with a crude model of background concentrations to predict the effective mean concentration. However, only further application of POLDIF to other years of data for which traffic and CO data are available can confirm estimates of the mean error.

11.6 Concluding Remarks

Air quality management often has as an objective the prediction of annual extremes of pollutant concentration in response to possible emission and meteorological conditions. A method has been constructed to extend the utility of a traditional and simple physically based model of pollutant dispersion from line sources in an urban area to predict seasonal extremes of concentrations. The extension requires the use of additional prior knowledge based upon analysis of the parametric form of historical concentrations. It also requires simplifying the objectives of the modelling exercise to prediction of concentrations for separate hours within seasonal periods to overcome problems of nonstationarity and autocorrelation. The three assumptions of the method were shown to apply for CO for the years available. The accuracy of predictions of the deterministic component of the hybrid model could be quantified further when other traffic and ambient concentration measurements become available. Alternatively, POLDIF or other deterministic models could be applied with more detailed model inputs if additional meteorological information were known. On the other hand, the method here offers a simple and often adequate means of inferring short-term extremes, particularly when only basic information is available.

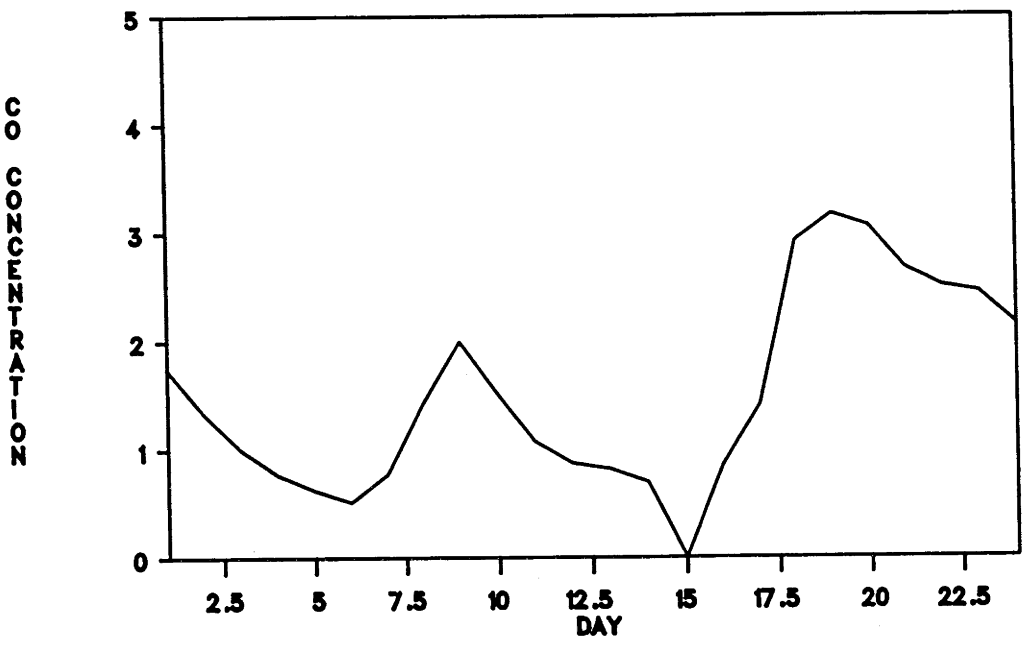

Figure 11.1: Diurnal variation of 1-hour CO (mg/m) averaged over the 1987 winter season

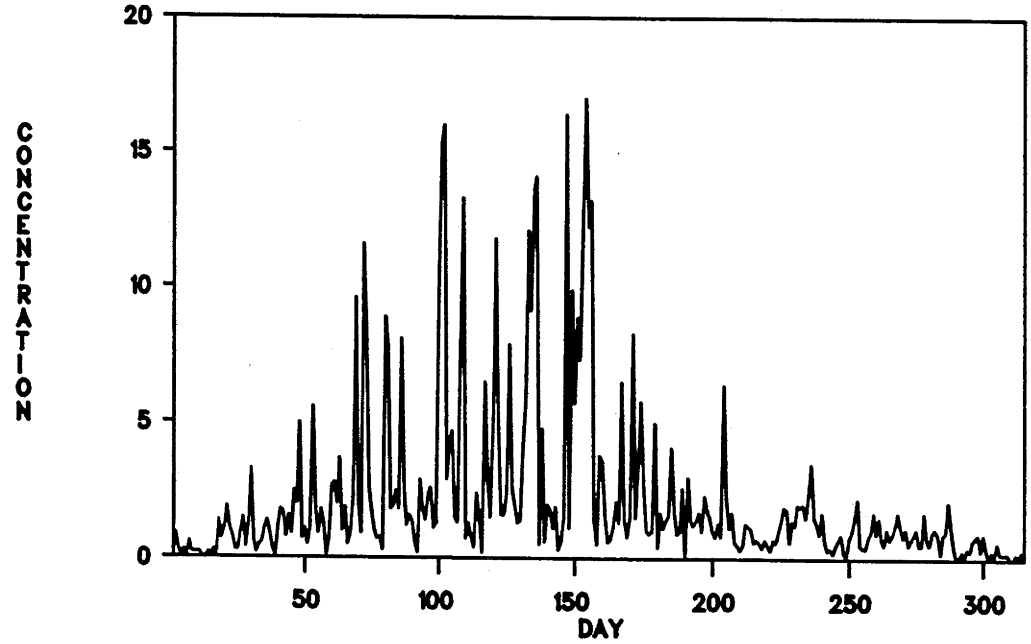

Figure 11.2: Time series of evening peak hour (1800-1900) concentrations of CO (mg/m) in 1987

TABLE 11.1
Autocorrelation at lag k in winter time series of evening peak hour concentrations of CO from 1982-87 at Civic site, Canberra.

Year	Sample Size	Autocorrelation at Lag k k=1	k=2	k=3	k=4	k=5
1982	156	0.312	0.131	0.070	-0.010	0.003
1983	97	0.006	0.072	0.083	-0.085	-0.037
1984	77	0.226	0.034	0.000	0.038	-0.090
1985	96	0.274	0.114	0.030	0.048	-0.057
1986	148	0.306	0.127	0.037	0.110	0.207
1987	178	0.452	0.323	0.190	0.136	0.074

TABLE 11.2
A measure r of stationarity for annual and winter time series of evening peak hour concentrations of CO.

Year	r value annual	winter
1982	0.777	0.947
1983	0.890	0.993
1984	0.901	0.973
1985	0.736	0.843
1986	0.772	0.892
1987	0.733	0.858

Note: The measure r of stationarity is the ratio of the variance of concentrations about a uniformly weighted moving average of 20 values to the variance about the constant arithmetic mean.

TABLE 11.3
Meteorological data used to produce the deterministic model estimates of the evening peak hour CO concentration for the winter seasons of 1985-87.

	Year		
	1985	1986	1987
wind speed (knots)	3.05	3.44	2.79
wind direction (from north)	315	315	315
cloud cover (10^{ths})	4.84	4.75	4.53
inversion height (m)	80	80	80

TABLE 11.4
Comparison of measured with predicted mean winter values of evening peak hour CO concentrations (mg/m)

Year	Mean	
	Measured	POLDIF
1985	3.39	2.31
1986	2.41	2.26
1987	3.19	2.36

TABLE 11.5

Descriptive Statistics for RRMSE fitted to the observed maximum of evening peak hour CO concentrations over the winter seasons 1982 - 1987 for two- and three-parameter gamma, Weibull and lognormal distributions

Statistics for RRMSE	Six Estimated Models					
	G3	G2	W3	W2	LN3	LN2
MEAN	0.302	0.105	0.851	0.876	0.676	0.116
SD	0.263	0.061	0.603	0.617	0.808	0.072
MIN	0.019	0.023	0.296	0.287	0.040	0.016
MAX	0.829	0.190	2.101	2.072	2.423	0.248

TABLE 11.6

Frequency of accepting the estimated two-parameter gamma and Weibull distributions for each of the evening peak hour winter CO data sets (from 1982-87) using two tests and four criteria

Estimated Model	Tests and Criteria					
	GIC	KL	AIC	SIC	CHI	KS
G2	6	5	5	5	3	3
W2	6	1	1	1	3	3

TABLE 11.7
Comparison of observed with a predicted range of maximum winter values of evening peak hour CO concentrations

Year	Relative Error in Mean Predicted	Predicted Maxima vs Shape 0.75	1.00	1.25	1.50	Observed Maxima
	0%	20.79	17.73	15.75	14.34	
1985	-30%	14.56	12.41	11.02	10.04	18.10
	+30%	27.02	23.05	20.47	18.64	
	0%	16.17	13.69	12.09	10.97	
1986	-30%	11.32	9.58	8.47	7.68	14.30
	+30%	21.02	17.79	15.72	14.26	
	0%	22.15	18.70	16.49	14.94	
1987	-30%	15.50	13.09	11.55	10.46	17.00
	+30%	28.79	24.32	21.44	19.42	

Chapter 12

On the Variability of the Wind Speed Exponent in Urban Air Pollution Models

12.1 Introduction

Benarie (1978, 1980, 1980a) has provided experimental evidence based on regression analysis that the assumption of simple inverse proportionality between pollutant concentration χ and wind speed U is not always valid for urban areas. In particular, Benarie (1980a) used a generalised box model of the form

$$\chi = CQU^b \tag{12.1}$$

to investigate seasonal variations in the wind speed exponent b for 24-hour SO_2 and black smoke (reflectivity) data. C is the atmospheric stability factor and Q the source strength (see Chapter 2). For both pollutants, he used data sets recorded over three years at three sites from Rouen and recorded over two years at five sites in Strasbourg, France. In both cities the spatial average of the exponent for SO_2 in summer was not significantly different from zero whereas during the winter heating season the spatial average of the exponent was -0.25. For black smoke the summer exponent was near -0.2 and in winter it varied from -0.3 to -0.5. The model described by equation (1) is commonly employed assuming an exponent of -1 (e.g. Hanna 1971).

It is argued by Benarie (1980a) that in real urban situations the exponent will be between 0 and -1. The value of the exponent is considered to depend on the extent of the urban area, the siting of the air quality monitors within it and the intensity of vertical mixing relative to advection. The argument is based on consideration of idealised situations. At one extreme, an infinite, relatively sourceless plane requires the wind speed exponent to be -1 as in the usual box model. At the other extreme, a constant ground concentration from uniform emissions over a sufficiently extended plane will not be strongly affected by wind speed so that the exponent tends towards zero.

In this chapter we investigate the seasonal variability in the wind speed exponent of model (1) for 24-hour average total suspended particulates TSP, β-scattering and CO data collected in Canberra, Australia. It is recommended that the ensuing generalised box model should be used in preference to the standard model and its use in predicting extremes and for forecasting is briefly explored.

12.2 The Data Set and Airshed Characteristics

TSP data have been collected in Canberra since 1980. The monitoring sites are in the major town centres of Civic and Woden, and, since 1981, in Belconnen and Kambah. TSP concentrations are recorded at each site and are based on 24-hour samples collected every 6 days. β-scattering and CO measurements are also taken and these are investigated for the year 1982 only. β-scattering and CO concentrations as 1-hour averages are measured at the Civic site only.

Meteorological data are collected at two monitoring sites; one near the Civic site and the other at the airport. Wind speed measurements at the sites differ greatly but the Civic site is considered to be more representative of the wind speeds affecting the level of air pollution concentrations.

Sources of particulate matter in ACT can broadly be divided into two categories: anthropogenic and natural. As Canberra has only light industry, the two most sig-

nificant sources of airborne emissions are motor vehicles and wood fire burning both for heating and cooking. The major natural source of particles is non-carbon matter. For carbon matter from wood fires and non-carbon matter, seasonal variation is significant. Traffic levels and hence the associated carbon matter emissions vary little seasonally. Taylor and others (1987) estimate that wood fires contribute nearly 50 % of the non-background particulate pollution in Canberra's commercial areas and about 80 % in suburban areas during the winter months. In the summer months a much higher non-carbon and organic matter component will be contributed because of the drier conditions.

The next section will attempt to quantify the broad influence of wind speed on TSP, β-scattering and CO as well as the carbon and non-carbon components of TSP. The variability of the influence of wind speed between seasons is significant. The winter season is taken to cover the months from April to September inclusive and the summer season the remaining months.

12.3 Method and Results

Time-wise regressions were applied firstly to the logarithmic form of the model (12.1), viz.

$$\log \chi = a + b \log U \tag{12.2}$$

where a denotes the natural logarithm of the atmospheric stability term C times the emission strength Q. These were performed for the annual summer and winter periods at individual sites. For TSP, the data were also amalgamated to obtain both a seasonal spatial average across the four monitoring sites in each year and a seasonal time average over the years at each site.

The results of the annual analyses are shown in Table 12.1 for TSP and in Table 12.2 for the carbon and non-carbon components of TSP, for β-scattering and CO. They largely corroborate the findings of Benarie. The wind speed exponent in Table

12.1, for example, varies from -0.16 to -0.63 in winter, depending on the site and the year. In summer, the associated t-values indicate that the exponent is not estimated as being negative with any reasonable statistical significance. Indeed, the exponent generally seems to be positive in summer. Histograms of concentration against wind speed categories of 1m/s width confirm the strength of the inverse relationship in winter and the weakness of it in summer. Figures 12.1-12.4 for the Civic site in 1984 and 1981 are typical results. Indeed, the inverse relationship holds for all sites and all years in winter. The associated t-values corroborate this as do the associated plots (not shown here) which are of similar behaviour to those of Figures 12.1 and 12.3. Conversely, for summer the t-values and the corresponding plots confirm that no inverse relationship can be assumed.

The estimate of the exponent b is reasonably constant from year to year in winter at individual sites. Except for the one odd result at Woden in 1982, at individual sites the variation in the mean value of the exponent in a given year is within one standard error of the mean in any other year. Observation of Figure 12.2 and the corresponding plots for the other sites suggests that the TSP concentration in summer does not decrease and perhaps increases with wind speed. For β-scattering and CO the seasonal variation is similar but the estimated exponent remains negative in summer as shown in Table 12.2. This is a situation we might expect since summer is a much drier season and in the case of TSP, wind may import dust as well as re-entrain it. Also, TSP is a measure of total particle mass. As wind speed increases larger particles may be suspended in the atmosphere. A few large particles would have a significant effect on TSP concentrations. It can be shown also that both the carbon and non-carbon components exhibit an inverse relationship with wind speed in winter but not in summer. Taylor and others (1987) have undertaken an elemental analysis of all TSP samples from the Civic site in 1981. Figures 12.5 and 12.6 show the contribution to TSP from carbon sources against wind speed class in winter and summer, respectively. Figures 12.7 and 12.8 show the same plots for the non-carbon component of TSP. Table 12.2 shows the estimated parameter values for model (12.2).

12.4 Hybrid Modelling Approach

The generalised box model has many good features which facilitate its use for constructing hybrid models to predict frequency distributions of air pollutant concentrations. As a deterministic component, it is simple but functional, linking the major meterorological and emission variables, such as wind speed and emission sources. Jakeman and Taylor (1985) introduced the box model structure as a component in the hybrid approach to develop a method for predicting the distribution of acid gas concentrations in an urban airshed. Their results were demonstrated for the 98-percentile predictions of 24-hour average data over annual periods at six monitoring sites.

Analysis of the seasonal variability in the exponent of wind speed for a box model confirms that the assumption of the exponent being equal to -1 can be inaccurate. To improve model performance, seasonally calibrated generalised box models should be used to predict aggregate properties of the seasonal distribution of urban air pollutant concentrations. Seasonal models for urban pollutants are also important when it comes to predicting extremes of concentrations. As shown in Chapter 11, stronger stationarity of the distribution is possible if periods of analysis are restricted to seasonal rather than annual urban pollutant concentrations.

One way to implement the generalised box model in a hybrid approach is to derive the analogue of the Simpson et al. (1983) treatment of the standard box model. This generalised analogue is

$$X_p = \frac{K}{U_{100-p}^b} \tag{12.3}$$

where X_p is the p-percentile of concentration and U_{100-p} is the opposite percentile of the distribution for wind speed. This model may work well over a mid-percentile range so that the two- or three-parameters in any parametric model of the frequency distribution can then be calibrated. In contrast to the POLDIF deterministic model of the last chapter, the model (12.3) has the potential to predict more than just the mean of the distribution of concentrations.

12.5 Use of Filtering and Smoothing Algorithms

The analysis has concentrated on a broad temporal exploration of behaviour to find average exponent values over a season. Of course it is unlikely that the exponent value is best represented over a year by two step functions which are discontinous at an arbitrarily chosen change of season. It is more likely to change smoothly over time. There are straightforward techniques available which allow the estimation of time variable parameters in linear models. Either or both of the parameters a and b in equation (12.2), for example, can be allowed to vary with time and estimates can be derived at each time point based upon a weighting of the observations of the dependent and independent variables. The simplest moving average model applied at time t would weight observations by unity in the interval $(t-k, t+k)$ and by zero outside this interval. The larger the value of k the larger the degree of smoothing of the estimate. It must be chosen as a compromise between achieving a good resolution of the parameter estimates over time k and sufficiently damping the effects of noise on the observations.

Other more sophisticated but algorithmically straightforward techniques (e.g. Jakeman and Young, 1984; Young, 1984) are based on assuming flexible but simple parameterisations of the parameter changes over time. Gauss-Markov process assumptions allow use of Kalman filtering and smoothing algorithms. The degree of resolution employed in the estimation of the value of the changing exponent should obviously depend on the quality and quantity of observations available and the accuracy required for the model predictions. In our case, TSP observations are made at most every 6 days so that little more is warranted than the estimate of the season step functions. Even in this case, however, the model (12.1) can be augmented with a parametric probability distribution component to provide, in addition to the mean, hybrid model estimates of extreme values.

To demonstrate the application of a more sophisticated smoothing procedure, the weekly average CO data collected in Canberra, Australia has been used. A random

walk process has been used to describe the parameter variation in time and a Kalman smoothing type algorithm to generate the parameter values. Results of estimation of the parameter b and K are shown in Figure 12.9. Note the consistent pattern in the parameters from year to year.

In Figure 12.9, the parameter b varies between - 1.892 and 0.755, and the parameter K between -1.287 and 0.040. This model yields a very good fit to the data as shown in Figure 12.10, the value of R^2 being 96 per cent. Note that the CO time series has been normalised by its average annual weekly value and then logorithmically transformed so that the linear model form (12.2) can be used. Figures 12.10 and 12.11 show the time series of historical CO concentrations and wind speed data.

The above smoothing algorithm employs all the data to derive a smoothed estimate. Filtering estimates, on the other hand, of a parameter at a particular point t in time use all the data up to time t only. By incorporating the present meterological information and CO levels with historical fluctuations of the parameter b and K, CO concentration could be forecast short time steps ahead. This filtering of box model parameters for forecasting applications is explored in a forthcoming paper by Jakeman, Bai and Young.

12.6 Concluding Remarks

This chapter has investigated the seasonal variability in the wind speed exponent of the generalised box model for 24-hour average total suspended particulates (TSP), β-scattering and CO data collected in Canberra, Australia. From the results here and those of Benarie, it is clear that the assumption of the exponent of the horizontal wind speed in urban air pollution models being equal to -1 is incorrect. Since the sensitivity of predicted concentration by a box model is high with respect to the exponent value, it can be argued that calibration of a generalised box model for even just a single season is likely to lead to much better predictions in future years than simply assuming an exponent value of -1, as is the present practice. It is also suggested that if extremes of the distribution of urban pollutant concentrations are required, a seasonal box model

component could be used in a hybrid modelling approach. This chapter also notes the relevance of the generalised box model in forecasting applications. A recursive methodology, which employs Kalman filtering algorithms in handling the data sets and updates of generalised box model parameter estimates, can be used.

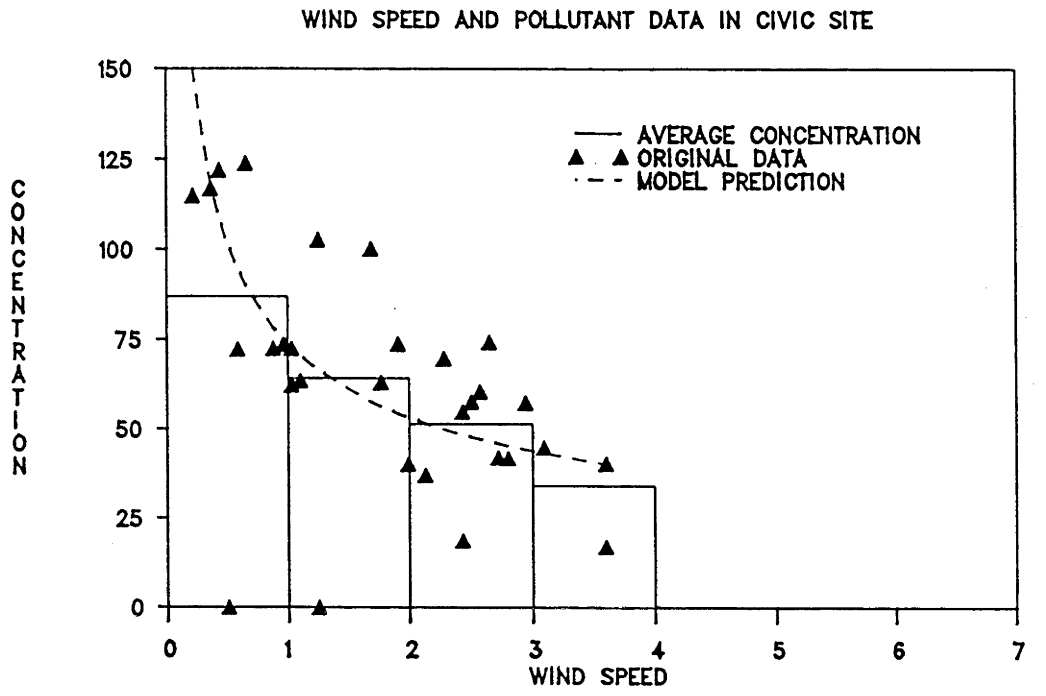

Figure 12.1: Histogram of TSP concentration against wind speed category for the Civic site in winter 1984. All figures also include raw pollutant measurements and regression model output for Equation (2). Wind speeds are in ms^{-1} and concentrations in μgm^{-3}

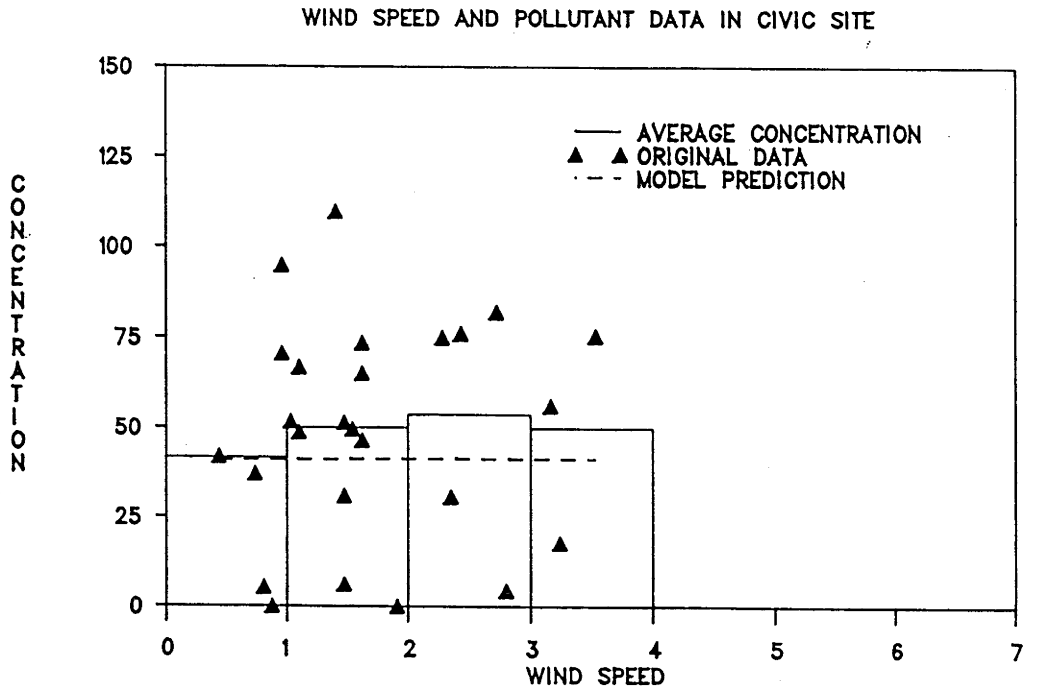

Figure 12.2: Histogram of TSP concentration against wind speed category for the Civic site in summer 1984

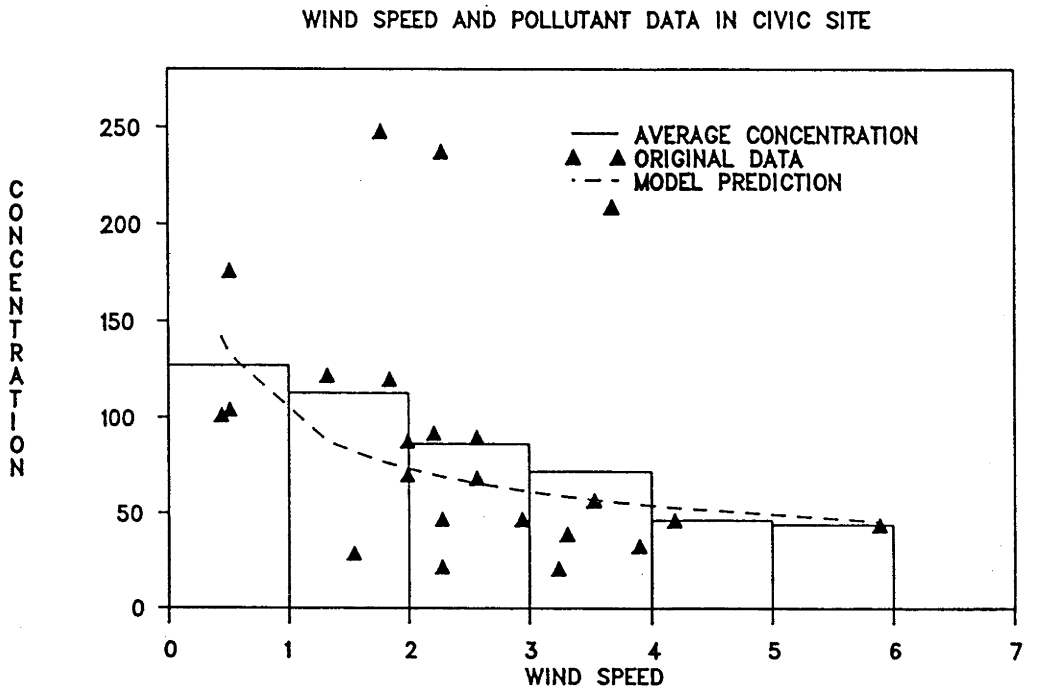

Figure 12.3: Histogram of TSP concentration against wind speed category for the Civic site in winter 1981

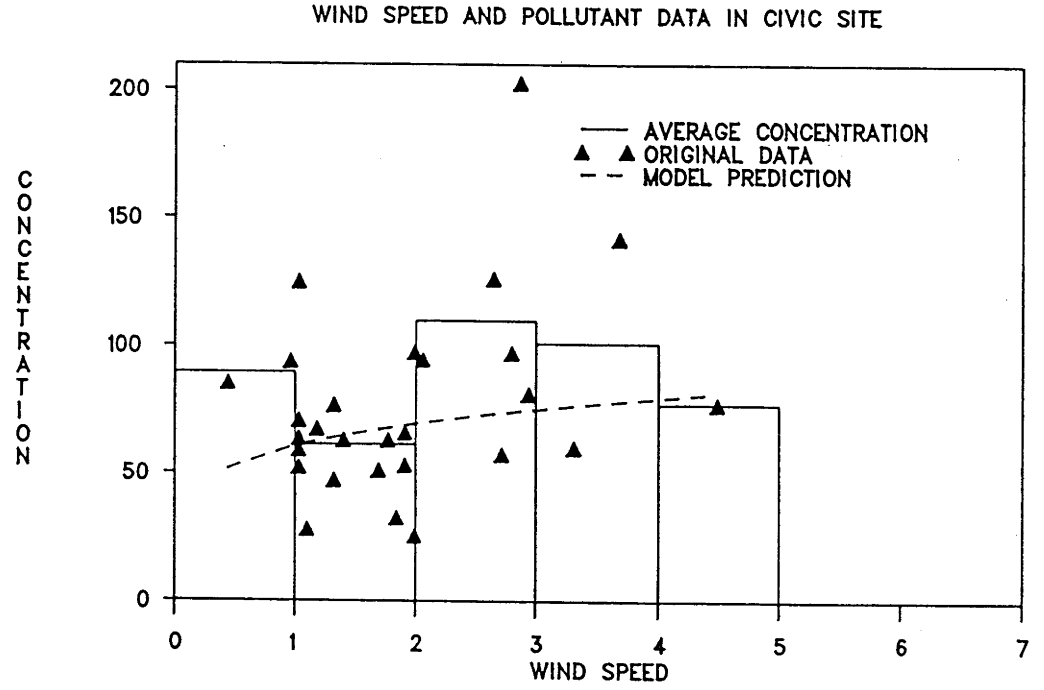

Figure 12.4: Histogram of TSP concentration against wind speed category for the Civic site in summer 1981

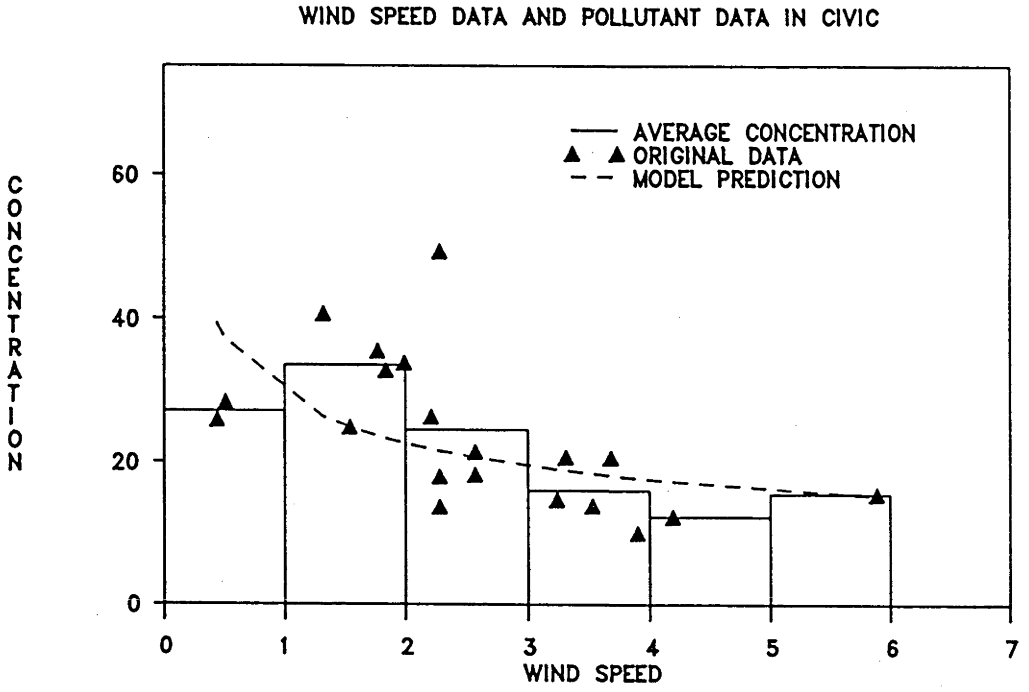

Figure 12.5: Histogram of the C component of TSP concentration against wind speed category for the Civic site in winter 1981

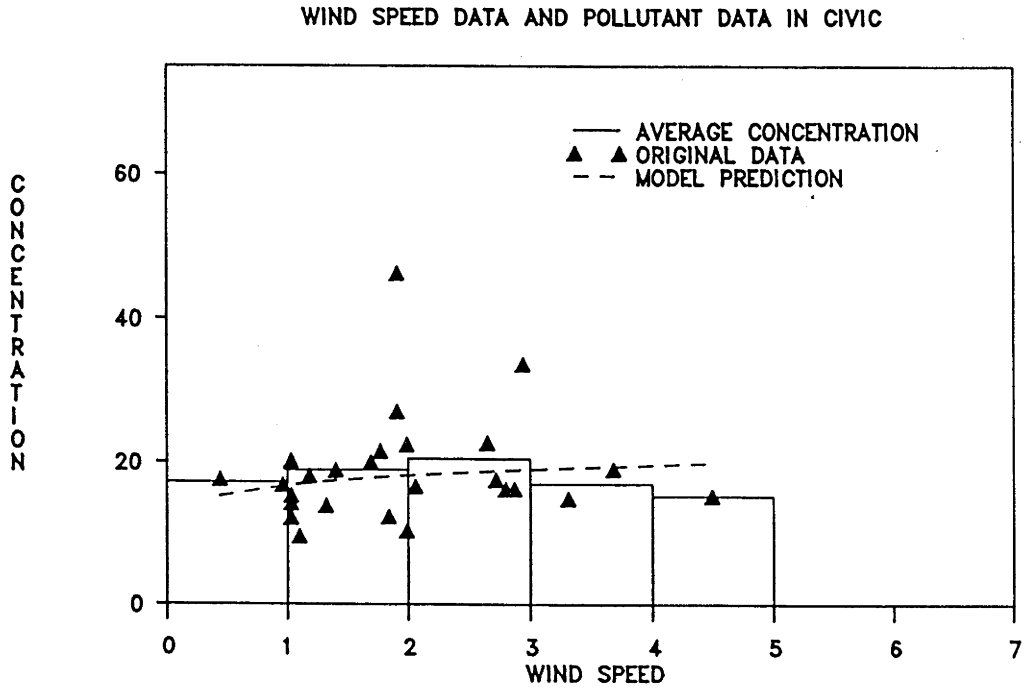

Figure 12.6: Histogram of the C component of TSP concentration against wind speed category for the Civic site in summer 1981

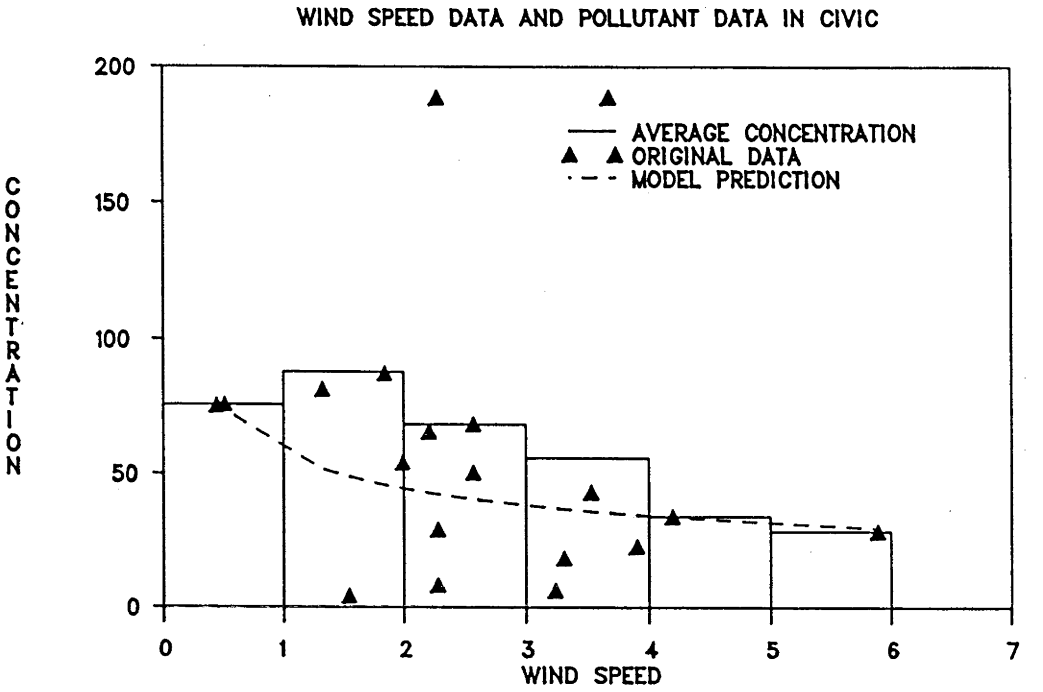

Figure 12.7: Histogram of the non-C component of TSP concentration against wind speed category for the Civic site in winter 1981

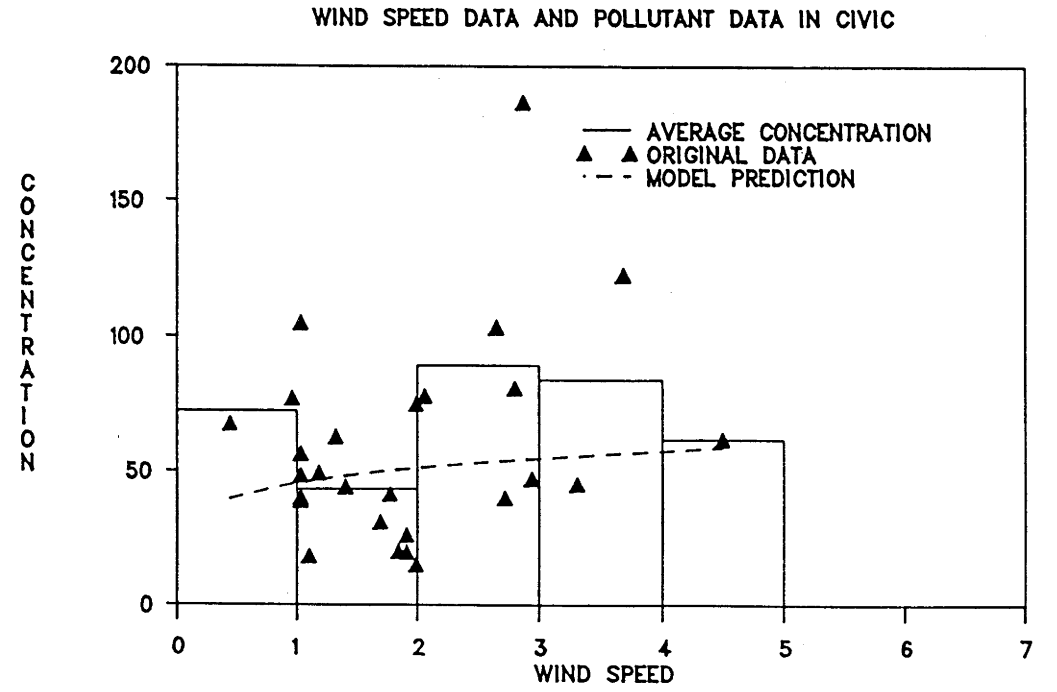

Figure 12.8: Histogram of the non-C component of TSP concentration against wind speed category for the Civic site in summer 1981

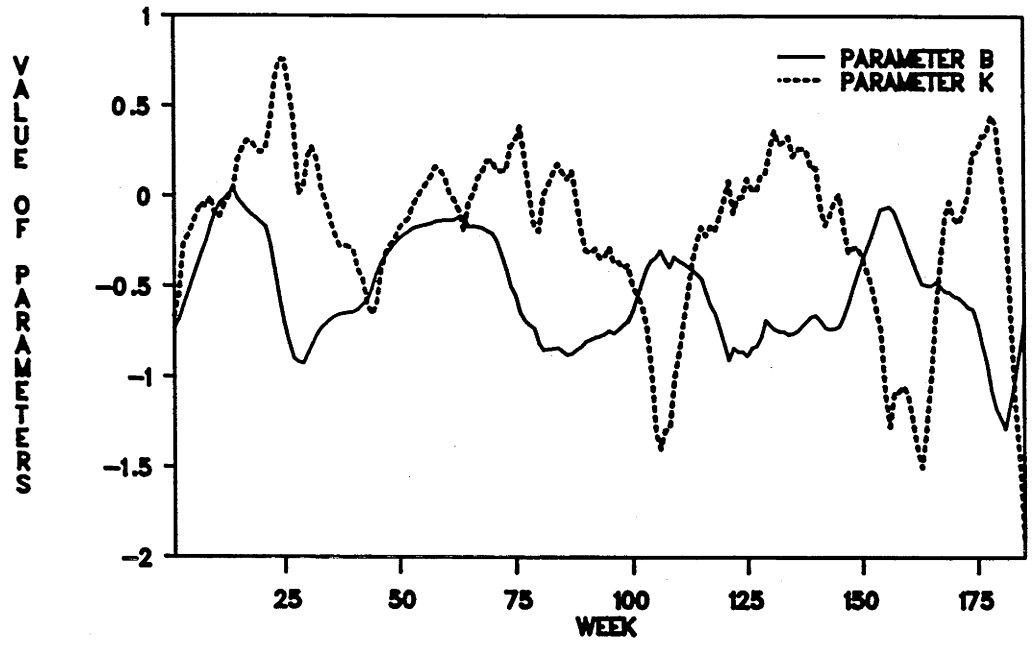

Figure 12.9: Weekly variation of box model parameters

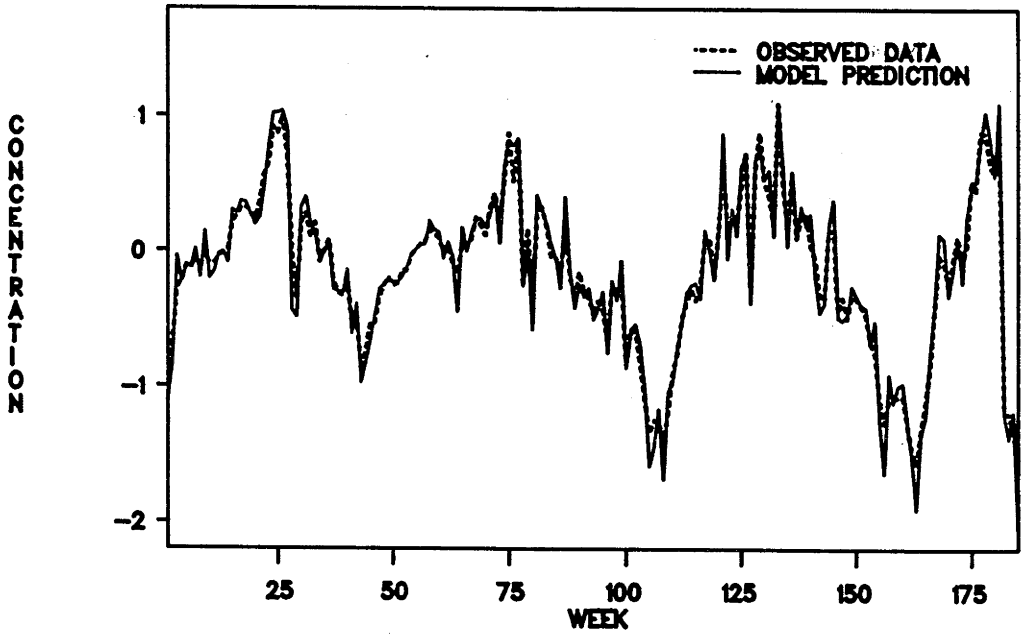

Figure 12.10: Time-varying model fit of the logarithm of weekly CO concentration as a proportion of annual weekly CO at the Civic site

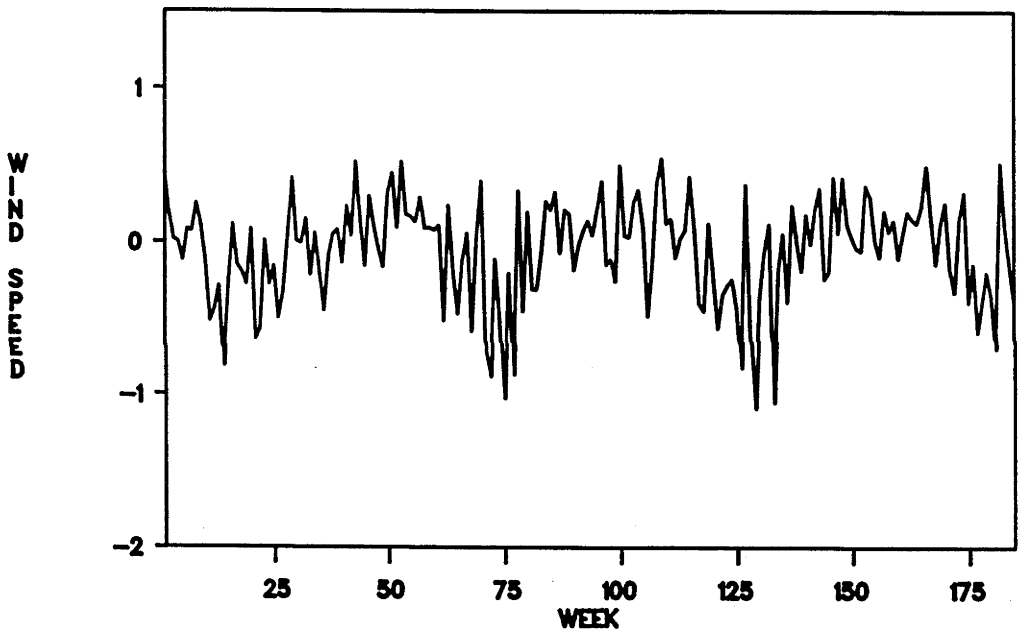

Figure 12.11: Weekly Variation in windspeed (m/sec) at Civic site

TABLE 12.1

Estimated windspeed exponent b and k (=CQ) values, and average $\bar{\chi}$ of measured values (μgm^{-3}) for Canberra TSP

Site	Season	1981			1982			1983			1984			1981-1984		
		b	k	$\bar{\chi}$	b	k	$\bar{\chi}$	b	k	$\bar{\chi}$	b	k	$\bar{\chi}$	b	k	$\bar{\chi}$
Civic	Winter	-0.44	99.48	92	-0.31	90.92	87	-0.40	66.69	70	-0.47	73.70	68	-0.37	79.04	78
	Summer	0.19	62.80	77	0.08	67.35	78	0.33	68.03	100	0.00	41.26	53	0.18	57.97	78
Woden	Winter	-0.37	55.70	51	-0.16	68.72	72	-0.24	62.80	65	-0.26	70.11	68	-0.28	64.72	64
	Summer	0.17	39.65	50	0.20	43.82	56	0.36	57.40	81	0.25	33.45	50	0.23	43.38	60
Belconnen	Winter	-	-	-	-0.41	38.47	37	-0.37	31.50	32	-0.63	33.12	32	-0.48	33.78	33
	Summer	-	-	-	0.19	36.23	49	0.51	38.47	67	0.06	21.12	29	0.25	30.27	49
Kambah	Winter	-	-	-	-0.34	41.68	42	-0.40	31.19	35	-0.58	35.87	40	-0.40	36.97	40
	Summer	-	-	-	0.27	33.78	46	0.38	31.82	54	-0.01	19.30	26	0.30	26.84	42
Spatial Average	Winter	-0.37	70.81	70	-0.28	55.15	59	-0.31	47.47	53	-0.44	50.40	53	-0.32	49.90	57
	Summer	0.19	50.91	64	0.17	45.15	58	0.42	46.53	76	0.08	27.39	40	0.25	40.04	60

TABLE 12.2

Estimated windspeed exponent b and k (=CQ) values, and average $\bar{\chi}$ of measured values for Canberra TSP components, β-scattering and CO at the Civic site

Year	Pollutant	Season	b	k	$\bar{\chi}$
1981	carbon TSP (μgm^{-3})	winter	-0.36	29.08	24
		summer	0.12	16.61	19
		full year	-0.09	20.49	21
	non-carbon (μgm^{-3})	winter	-0.37	57.40	67
		summer	0.17	45.60	59
		full year	-0.13	56.91	63
1982	β-scattering (m^{-3})	winter	-0.41	0.50	0.54
		summer	-0.07	0.31	0.41
		full year	-0.26	0.41	0.47
1983	CO (ppm)	winter	-0.31	1.57	1.73
		summer	-0.10	0.82	0.95
		full year	-0.28	1.27	1.42

APPENDIX I

Appendix 1:
History, Generation and Classification of Air Pollution

A1.1 Historical Review of Air Pollution Problems

The incidence of human-induced or anthropogenic air pollution can be traced as far back as the discovery of fire. While cooking and warming improved the function of the human health and life, the fire also represented the beginning of air pollution created by human beings. Air pollution history follows closely the evolution of social communities and the revolution of industry, science and technology. If one were to use a convex curve to express the level of air pollution over time, the peak of air pollution levels would be located somewhere in the past few decades, and the turning point would be at the time when the quality of our atmosphere was recognised as an important factor in quality of life and life support systems (Chambers, 1976). Of course, this assessment ignores the continuing build up of greenhouse gases globally.

Originally, many of the major air pollutants arose from natural causes during earth's early beginnings. Restricted meteorological conditions occurred from time to time and resulted in crucial air pollution effects. High temperatures and poor ventilation caused grass, brush or forest fires and subsequently produced contamination of the air by thick palls of smoke. Ocean evaporation and sea sprays increased in hot weather, transferring trace gases and particulate matter into the atmosphere. Sandstorms bore deep layers of dust carried from deserts by Trade Winds. Volcanoes released large amounts of sulfur dioxide and particulates. Forests released volatile hydrocarbons which significantly affected world climate and participated in photochemical reactions during sunny weather (Whelpdale and Munn, 1976). All the natural sources were major emitters yielding global background concentrations in the early stages of air

pollution history.

However, air pollution problems are generally associated with emission sources related to human activities in the urban, industrial and agricultural sectors. The term 'air pollution' is normally defined as the presence of foreign substances in the atmosphere. When air pollution occurs, it indicates that the concentrations of these substances are substantially above their normal ambient levels, effectively change the characteristics of air, significantly increase the duration of pollutants in the environment, and cause dangerous damage, such as injury to human health, ecosystems or property and materials. These foreign contaminants may be in gaseous, vaporous or solid particulate states.

Air Pollution Related to Energy

In the early stages of air pollution in human history, air pollution was mainly regarded as a local problem. Highly developed industry and rapidly increasing populations in the centralised cities became two key factors for air pollution increases. With the discovery of the energy potential of coal, its extensive usage culminated in the Industrial Revolution and major economic improvements. However, coal smoke and associated gases as traditional pollutants occupied the centre stage from the beginning of the fourteenth century to the early part of the twentieth. Even though there was very early recognition of a public health problem and public protest, coal-based industrial economies continued to develop without restriction in large cities, so that finally some major air pollution disasters occurred:

* In 1930, a strong atmospheric inversion settled over the Meuse Valley in Belgium, and black smoke which chiefly contained oxides of sulfur, inorganic acids, metallic oxides and soot caused sixty-three deaths and several hundred others became ill.

* In 1948, during a particularly calm and meteorologically stable period,

the Donora Smog in Pennsylvania resulted in twenty deaths and several thousands persons were hospitalised.

* In 1952, the great London Smog lasted 5 days last over the period December 5 through 9, and caused 4000 deaths.

The three most well-known air pollution disasters in this century threatened personal survival, and led to public reaction, legislative action and enforced emission reduction. Since then, the prevention of high smoke episodes has achieved substantial success in most cities of the world.

After more than 400 years of burning coal, the industrial world has rapidly shifted to petroleum and natural gas in recent years. The combustion of oil and gas has not only diminished the coal smoke nuisance, but has also resulted in a revolution in machinery and transportation, and provided the raw materials for petrochemical products. Unfortunately, this has led to the introduction of new forms of air pollution. Secondary air pollutants such as those in photochemical smog are now generated from a complex sequence of chemical reactions involving the exposure to sunlight of mixtures of olefins and other reacting petroleum products which contain organic nitrogen components and nitrogen oxides. It is recognised that such pollutants have intrinsic toxic or irritative potential at relatively low concentrations (Faith and Atkisson, 1972), the photochemical reaction affecting biological systems as well as restricting visibility. The first publicised photochemical smog problem was that of Los Angeles where most of the energy supply derives from petroleum rather than from burning coal. With stable atmospheric conditions in the late summer of 1943, large quantities of photochemical smog appeared as a grey-blue pall settled over the city causing burning eyes, chafing throats and damaging green leaves. Mainly it is produced in exhaust gases from internal-combustion engines. Photochemical problems have now become common pollutants in most cities of the world that are reliant on mechanised transportation.

Air Pollution by Industry and Agriculture

There is no doubt that the usages of energy are the major air pollution sources in the urban and industrial sectors. Most factories with modernized production have also contributed to air pollution concentrations. For example, the chemical industrial factories turning out products such as fertilizer, plastic materials, and medicine or daily chemical articles like skin cream and perfumery, cannot operate without discharging high concentrations of various substances. The cement industry, open cast mining industry and even building industry normally emit large quantities of particulates into the air. Paper mills not only pollute water resources but also spread their contamination into the atmospheric environment. Different industries, whether as large as iron and steel plants or manufacturing companies, or as small as flourmills, wool, cotton or food processing shops, and even electronic industries involved for example electroplating, acid, polishing treatments, are all contributors of harmful material to the air. In high density population centres and industrial areas, the collection of all individual emitters result in large volumes of emissions and high levels of air pollution concentrations, leading to serious air pollution problems.

Generally there is no argument that the urban and industrial sectors are major contributers of air pollutants concentrations. However, it is easy to ignore the existence of other important pollution sources. Anthropogenic emissions discharging to the atmospheric environment should also take into account the significant quantities deriving from agricultural and forest practices over large rural areas. It is recorded history that agricultural mismanagement has resulted in deserts such as the Rajasthan desert (Whelpdale and Munn, 1976) and increases in soil erosion from wind-blown dust during dry weather.

Over many centuries, slash burning has been used to prevent development of disease organisms, to reduce fire hazards and to produce ash content for fertilizers. Also, decaying farm wastes such as animal and vegetable wastes are regarded as quality fertilizers even now in many countries. They release a wide variety of substances into

the atmosphere, such as ammonia, hydrogen sulfide, methane, sulfides, mercaptans, carbon monoxide and carbon dioxide. With the development of new chemical fertilizers and the introduction of modern operational procedures, new forms of pollution were also introduced. Large quantities of fertilizers, especially chemical types, directly or indirectly enter the atmosphere every year. Pesticides are recommended for agricultural and forested treatment in many countries. Spraying operations, especially aircraft spraying, pollute the atmosphere on a large-scale and an extensive range of substances is used.

A1.1.2 Global pollution problem

Over the last several hundred years, traces of substances emitted from human activities from different pollutant sources have gradually accumulated into larger and larger volumes of air and have been distributed around the world. In other words, the local-scale pollution problems have extended to continental and global scale pollution problems. Certainly some atmospheric processes can dilute pollutants through "sink mechanisms", deposition of particles by gravity forces, scavenging by precipitation, or absorption at the surface of the earth. Chemical transformation in the atmosphere itself can decrease a related substance such as sulfate particles. However, the general assimilative capacity of the atmosphere is not unlimited. Basically, continental or global air pollution problems involve the two aspects covered seperately in the following sections.

Neighbourhood Influence

The effects of air pollution emitted either from urban and industrial areas or rural areas are not only strictly constrained to local regions because the air is not defined by political boundaries. Undesirable effects can significantly influence large areas or transfer to other places. For example, it has been found that sulphur dioxide and nitrogen

oxide pollutants contributed from industry areas can cause increases in the acidity of rainfall. This can spread over many countries and affect ecological systems, especially agricultural production. Photochemical smog clouds can cover areas of thousands of square kilometers, automatically extending the problem to neighbouring countries as has already found in Europe (Guicherit, 1976). Long range transport of concentrations of pollutants and photochemical smog has led to problems on a global scale. The seriousness of air pollution episodes has sometimes only been recognised well after the related occurrence of specific diseases when statistical studies and medical research brought excess mortality and morbidity to light. Such neighbourhood influence over countries need to be controlled by international cooperation.

GreenHouse and Ozone Layer Effects

In recent years greenhouse effects and depletion of the ozone layer has become of major concern as a continental and global air pollution problem. Carbon dioxide is not often considered to be an air pollutant, although it can have significant meteorological and physiological effects when reaching high levels. There is a consensus among scientists that high levels of carbon dioxide with other trace gases could lead to climate slowly warming, particularly in the cold and temperate latitudes. Change of climate could have other strong consequences such as more floods and droughts, more southern cyclones and less snow, even catastrophic sea-level rises in the future. Measurements of trace gases began at the Mauna Loa Observatory in Hawaii and at the South Pole in 1958. Since then it has been realised that the concentrations of carbon dioxide have been increasing steadily in the atmosphere world wide. In general, the biological and geochemical process of the atmosphere can provide a sufficient natural disposal system for CO_2. Unfortunately since the beginning of the twentieth century, the extensive burning of fuels has generated high concentrations of carbon dioxide so high as to break carbon balance between the rate of emission and disposal, especially after deforestation caused by industry or bush fire. The continue increase of CO_2 could be

a potential danger to our society and ecological systems.

Similarly, the depletion of the ozone layer has also recieved much more attention from both scientific researchers and the public recently. It is known that the ozone layer is a thin layer in the stratosphere, absorbing many harmful sunlight frequences, such as ultraviolet. Ultraviolet light can cause many health problems, such as sunburn, skin cancer and eye cataracts, and can also destroy vegetation in biological systems. Recently, it has been found that a special group of chemicals named Chlorofluorocarbons (CFCs) can damage the ozone layer and release harmful ultraviolet rays leading to the risk of deterioration of our living environment. CFCs have been widely and routinely used, for example, furniture, fridges and air conditioners. Their effects might become even more serious a problem in the future without restrictions on their use. Much research is required in the understanding of atmospheric behaviour, the complex chemical interactions and the possible legislation, standards and control mechanisms.

A1.2 The Atmosphere and Air Pollutants

A1.2.1 The Concept of the Atmosphere

The term 'atmosphere' is commonly defined as the layer of a gaseous mixture envoloped around the earth. Gaseous density of the atmosphere is distributed crudely in the following way: about 50% at the earth's surface which is below 3.5 miles and 99% below 18 miles. Beyond the latter distance its density rapidly decreases with increasing height.

Based upon its molecular composition, the atmosphere can be divided into the homosphere and heterosphere (Urone, 1976). The homosphere extends outward to some 55 miles from earth and is a result of the continuous turbulent movements of the air and winds in the lower regions of the atmosphere. The heterosphere has four distinguishable layers: the first layer is from approximately 55 to 125 miles, the second from 125 to 700 miles, the third from 700 to 2000 miles and the fourth 2000 to 6000

miles from the earth. It helps filter out the highly energetic portions of the sun's rays and forms a very strong oxidation reservoir for impurities.

On the other hand, according to its characteristic temperature, the atmosphere can be distinguished by four temperature regions: troposhere, stratosphere, mesophere and thermosphere. Up to 5-10 miles height is the troposhere whose lapse rate of temperature for a given locality varies with ambient weather conditions and is important with respect to the effect on the turbulent vertical movements of air. The stratosphere has a height of approximately 20 miles and forms a stratified cover to the troposphere as temperature stabilised air. The mesosphere is located with increases of temperature to about 30 miles. Above the mesosphere, is the thermosphere. Air becomes extremely thin and the molecules and atoms acquire high kinetic energy. In addition, between the troposhere and the stratosphere is the tropopause, between the mesosphere and the stratosphere is the mesosphere, and between the mesosphere and the thermosphere is the mesopause.

Finally, in terms of its chemical and physical properties, the atmosphere can be separated into two parts: the chemosphere and the zonosphere. The chemosphere exhibits the chemical properties of the atmosphere, and it includes the ozone layer, atomic oxygen, hydrogen and hydroxyl and hydrogen oxide, as well as radicals. The zonosphere contains relatively large numbers of ions and is recognised by its ability to reflect radiowaves. It can be also partitioned into several layers categorised simply as D, E, F1, F2, and G but the interested reader should seek more detail in (Urone, 1976).

A1.2.2 The Composition of Clean Air

Generally, the term 'clean air' is very much concerned with the original volume percent composition of the air in the homosphere on a dry basis. In chemical analysis, it is known that the relationships of mole percent and volume percent are numerically equal, and that the partial pressures of the individual components of a gaseous mixture are proportional to either the volume or mole percent. The substances of dry air

normally include: nitrogen, oxygen, argon, carbon dioxide and all others, and composition is approximately to be assigned an average "molecular" weight of 28.96. It has been determined that nitrogen occupies 78.09% of the volume of the atmosphere, oxygen 20.94%, argon at slightly less than 1%, carbon dioxide 0.032% and all others only 0.004%. Nitrogen is the most abundant gas and relatively inert except under special conditions, such as at high temperatures or pressures or in the pressure of lightning. Oxygen is an essential gas to life and is the necessary ingredient for metabolic processes. It is very active in chemical reaction with both living and inert matter. Argon is essentially inert and relatively heavy and is generally ignored in consideration of atmospheric processes. Carbon dioxide is the most abundant of the minor gases. All the gases found in the air including helium, neon, ozone, krypton, hydrogen, xenon and methane are essentially inert, forming a rare compounds under special conditions.

A1.2.3 Background Pollutant Concentrations

The goals of air quality management cannot be determined by analysis of the chemical composition of air since the atmospheric environment was naturally polluted in the first place. It has recognised that the total global mass of some trace gases emitted by nature exceed those emited by man by several orders of magnitude. (Urone, 1976). Obviously, even neglecting the contributions of pollutant concentrations from human activities, the atmosphere can still never have lower concentrations of any pollutants than the atmospheric background concentrations. Therefore, the definition of 'clean air' must depart from the results of pure chemical analysis and depended on the estimation of global background concentrations.

Global background concentrations have been estimated in the unit of parts per million. The results were reported by Robinson and Robbins (1972), and by Newill (1976), in micrograms per cubic meter (mg/m^3). Since atmospheric background concentrations are the lowest concentrations that can possibly be achieved, they become important references when setting air quality goals and air quality standards. However,

shoubububububuld be noticed that the background concentrations of different pollutants vary m tirtirtirtirtirime to time, and may be either decreasing or increasing. There are requirements botbtbtbtbtbtth a world-wide monitoring network and global air pollution modelling in or- to p p p p p p provide information on changes in global background concentrations. Australia renthtntntntntitlly contributes to both the monitoring and modelling work.

.2.4.4.4.4.44 Classification of Air Pollutants

Acccccccccccording to statistical analysis in 1972 by Robinso and Robbins, there are on a rld-wl-wl-wl-wl-w-wide basis, more than 6 billion tons of pollutant discharged annually into the nospspspspspplhere from pollution sources, of which H_2S comprises 24%, CO 30% and others %. Ir Ir Ir Ir IrIm recent years, air pollution control has made significant progress. Emissions of ne p(p(p(p(ppollutants such as CO and SO_2 have decreased in large quantities from developed ntritritritritriries. However, some pollutants such as NO_2 seem to have undergone increases.

At pt pt pt pt p present there are more than 100 substances regarded as air pollutants discharging o thththththhe atmosphere from emission sources. Major air pollutants can be classified into e catatatatatategories. A brief discussion follows (Leithe, 1972).

Carararararbon compounds in the air comprise an almost infinite number of compounds itedededededed only by the sensitivity and selectivity of the analytical method used. Hydro- bononononononns and oxygenated hydrocarbons are of major concern because they react to form otococococoochemical oxidants and secondary photochemical products, and occupy the layer ere re re re re e the predominant one is ozone. Carbon dioxide contributes mainly to the green- use se se se se e effect, and the temperature of the atmosphere slowly increases from the increase cononononononncentrations of CO_2. Carbon monoxide is regarded as a potent poison.

Sulululululfur compound gas is an important gaseous air pollutant. There are several of fur ir ir ir ir r oxides. Sulfur dioxide and sulfur trioxide used to attract more attention and are gardrdrdrdrdded as major atmospheric pollutants. Sulfur dioxide affects the respiratory organs d thththththhis is facilitated by the presence of water vapor and smog. Polluted respiratory air s anananananann adverse effect on persons suffering from bronchitis and can lead to a significantly

higher mortality rate. Low concentrations may lead to acute damage in the form of localised destruction of leaf tissues. In air, sulfur dioxide rapidly combines with water to form sulfuric acid (H_2SO_4), and they are apparently responsible for the noxious effects of flue gases in respiratory air and are frequently associated with poor visibility.

The oxides of nitrogen include a number of products. However, only three of them, nitrous oxide (N_2O), nitric oxide (NO), and nitrogen dioxide (NO_2) are found in any appreciable quantities. NO and NO_2 are often treated together in air and are lumped as "nitrogen oxides" and given the symbol NO_x. Normally, N_2O has no toxic chemical effects at room temperature unless the gas is compressed to high concentrations which might lead to the nasotic effect occurring.

Suspended particles are very common pollutants. For nonviable particles, the coarse dust and fine dust cause damage to materials, such as linen, clothing and buildings. They also pollute residential areas and cause eye irritation or injury. These dusts can also damage plants and vegetables. Very fine dust will stick on surfaces because of thermal diffusion. Some chemical particles may lead to poisoning. For viable particles, problems caused include asthma, high fever, sinusitus, catarrh, bronchial problems, and forms of dermatitis. Microorganisms may cause airborne infection, spread of respiratory diseases and many bacterial diseases.

It is well-known that radioactive radiation, such as α, β, and γ radiation is dangerous to life. However, the chemical nature of the carriers, known as radio-nuclides, are also a major concern. For example, the β emitter Sr_{90} is considered a very harmful substance which can deposit in the bones, having affinity for calcium. Other radio-nuclides, such as I_{131}, can accumulate in the human thyroid gland or get into animals' milk after only a short time.

A1.2.5 A Procedure for Setting Air Quality Standards

The setting of air quality criteria either as legislated standards, goals or guidelines by public authorities has been one of the major steps towards controlling widespread

air pollution. Another major step has been the establishment, often through Clean Air Acts in developed conntries, of targets for emission levels. A possible sequence of setting or developing air quality standards can be described as follows (Newill, 1976):

a. Based on dose-response analyses, determine air quality criteria.

b. In terms of air quality criteria, evaluate acceptable concentrations of pollutants and establish air quality goals.

c. Give consideration to feasibility of achievement in the light of nationally available air quality control technology, as well as economic, social and political issues, in order to select appropriate and achievable air quality levels.

d. Using the defined air quality criteria, goals, and the achievable air quality levels, set up air quality standards for the region, state or country.

e. Establish a standard for measurement and testing of the ambient air and air pollution effects must be set up so that air standards can be measurable directly.

f. Examine the consequence of the draft of air quality standards performed in practice, and modify them if necessary.

g. With economic, social and political progress and improvements in air quality control technology, progressively update the quality standards in order to achieve higher levels of air quality.

APPENDIX II

TABLE A2.1

Estimated parameter values, maximized log-likelihood values and RRMSE values for different percentiles fitted to NO samples over different years at the Museum site

Year	n	Estimated Model	Shape α	Scale β	Location γ	Max(log L)	RRMSE for different percentiles				
							MAX1	MAX2	98	MEAN-U	MEAN-F
1975	116	G3	2.70	1.50	0.24	-253.42	0.125	0.083	0.085	0.057	0.093
		G2	3.14	1.37	0.00	-253.70	0.148	0.102	0.113	0.068	0.088
		W3	1.55	4.25	0.47	-255.41	0.155	0.096	0.101	0.069	0.106
		W2	1.78	4.84	0.00	-258.38	0.192	0.124	0.123	0.086	0.135
		LN3	0.50	1.46	-0.57	-252.23	0.020	0.019	0.054	0.011	0.091
		LN2	0.59	1.29	0.00	-253.44	0.109	0.077	0.024	0.050	0.117
1976	280	G3	2.06	2.35	0.02	-687.36	0.262	0.292	0.161	0.278	0.184
		G2	2.09	2.33	0.00	-687.39	0.265	0.295	0.164	0.280	0.179
		W3	1.38	5.26	0.10	-695.02	0.266	0.285	0.146	0.267	0.188
		W2	1.42	5.41	0.00	-696.70	0.276	0.292	0.153	0.273	0.186
		LN3	0.62	1.49	-0.51	-678.64	0.112	0.209	0.094	0.228	0.232
		LN2	0.75	1.33	0.00	-686.63	0.083	0.077	0.035	0.152	0.326
1977	297	G3	3.44	1.42	-0.13	-678.94	0.190	0.219	0.071	0.216	0.157
		G2	3.20	1.49	0.00	-679.27	0.180	0.211	0.062	0.209	0.178
		W3	1.80	5.28	0.09	-687.24	0.235	0.241	0.086	0.228	0.129
		W2	1.84	5.39	0.00	-687.89	0.240	0.245	0.089	0.230	0.121
		LN3	0.45	1.66	-1.05	-676.54	0.100	0.166	0.025	0.182	0.172
		LN2	0.60	1.40	0.00	-687.37	0.112	0.019	0.119	0.105	0.302
1978	227	G3	1.32	4.24	0.25	-613.74	0.165	0.223	0.211	0.229	1.707
		G2	1.57	3.74	0.00	-616.37	0.209	0.257	0.243	0.253	1.496
		W3	1.15	5.90	0.26	-614.66	0.181	0.229	0.214	0.232	0.112
		W2	1.25	6.33	0.00	-619.83	0.223	0.260	0.240	0.252	0.143
		LN3	0.76	1.56	-0.43	-614.28	0.097	0.080	0.110	0.167	0.505
		LN2	0.89	1.42	0.00	-616.49	0.330	0.068	0.011	0.165	0.881
1979	271	G3	1.77	1.96	0.52	-586.21	0.256	0.061	0.045	0.155	0.044
		G2	2.60	1.54	0.00	-592.42	0.313	0.119	0.096	0.192	0.087
		W3	1.34	3.75	0.56	-588.42	0.289	0.084	0.060	0.168	0.061
		W2	1.61	4.50	0.00	-602.57	0.340	0.132	0.099	0.198	0.153
		LN3	0.66	1.17	0.04	-586.16	0.078	0.074	0.054	0.088	0.048
		LN2	0.65	1.18	0.00	-586.18	0.091	0.063	0.045	0.092	0.046

TABLE A2.1 continued

Year	n	Estimated Model	Shape α	Scale β	Location γ	Max(log L)	RRMSE for different percentiles				
							MAX1	MAX2	98	MEAN-U	MEAN-F
1980	175	G3	2.32	1.51	0.17	-365.55	0.198	0.128	0.100	0.102	0.082
		G2	2.65	1.39	0.00	-366.27	0.219	0.146	0.117	0.114	0.079
		W3	1.49	3.80	0.25	-369.59	0.226	0.140	0.103	0.113	0.114
		W2	1.62	4.12	0.00	-373.07	0.248	0.158	0.118	0.125	0.149
		LN3	0.57	1.22	-0.31	-363.01	0.053	0.033	0.031	0.035	0.074
		LN2	0.65	1.10	0.00	-364.14	0.047	0.043	0.033	0.024	0.103
1981	276	G3	2.06	1.33	0.14	-518.99	0.360	0.329	0.242	0.326	0.107
		G2	2.35	1.22	0.00	-521.10	0.377	0.343	0.256	0.337	0.113
		W3	1.37	2.98	0.16	-528.29	0.357	0.315	0.222	0.311	0.157
		W2	1.46	3.20	0.00	-534.84	0.372	0.326	0.231	0.318	0.199
		LN3	0.63	0.88	-0.11	-508.92	0.222	0.244	0.177	0.272	0.078
		LN2	0.67	0.83	0.00	-509.56	0.178	0.213	0.148	0.250	0.091
1982	292	G3	1.57	3.49	0.23	-773.77	0.162	0.268	0.116	0.227	0.108
		G2	1.80	3.18	0.00	-776.13	0.193	0.290	0.139	0.245	0.118
		W3	1.24	5.88	0.26	-777.48	0.173	0.268	0.110	0.223	0.137
		W2	1.32	6.27	0.00	-783.42	0.203	0.287	0.130	0.238	0.181
		LN3	0.72	1.54	-0.31	-768.14	0.084	0.141	0.008	0.161	0.080
		LN2	0.80	1.44	0.00	-770.05	0.217	0.061	0.069	0.140	0.117
1983	313	G3	2.69	1.80	0.10	-741.46	0.180	0.221	0.101	0.208	0.084
		G2	2.85	1.74	0.00	-741.69	0.189	0.227	0.108	0.213	0.079
		W3	1.62	5.30	0.24	-748.57	0.222	0.241	0.114	0.219	0.111
		W2	1.71	5.60	0.00	-751.88	0.237	0.252	0.124	0.226	0.132
		LN3	0.53	1.58	-0.62	-738.46	0.048	0.144	0.037	0.160	0.085
		LN2	0.63	1.42	0.00	-742.69	0.092	0.049	0.052	0.112	0.135
1984	82	G3	2.65	1.30	1.13	-166.19	0.120	0.148	0.073	0.072	2.453
		G2	5.23	0.87	0.00	-167.39	0.022	0.069	0.010	0.028	1.605
		W3	1.62	3.61	1.32	-165.75	0.032	0.087	0.029	0.035	0.108
		W2	2.39	5.16	0.00	-170.43	0.065	0.012	0.028	0.025	0.108
		LN3	0.45	1.43	-0.03	-167.05	0.194	0.189	0.097	0.102	0.327
		LN2	0.45	1.42	0.00	-167.05	0.198	0.192	0.099	0.104	0.331

TABLE A2.2

Summary of tests and criteria applied to NO samples by six distributions over different years at the Museum site

Selection Set	Estimated Model	Tests and Criteria GIC	KL	AIC	SIC	CHI	KS	UPE	FPE
2- and 3-Parameter Models	G3	7	3	1	0	5	7	5	2
	G2	6	1	2	3	6	8	1	4
	W3	3	1	1	0	5	7	1	2
	W2	0	0	0	0	4	6	2	1
	LN3	1	1	0	0	1	1	0	0
	LN2	8	4	6	7	10	10	1	1
3-Parameter Models	G3	9	7	7	7	5	7	5	7
	W3	4	2	2	2	5	7	5	3
	LN3	1	1	1	1	1	1	0	0
2-Parameter Models	G2	6	3	3	3	6	8	7	4
	W2	0	0	0	0	4	6	2	3
	LN2	9	7	7	7	10	10	1	3

Table A2.3

Summary of RRMSE values for different percentiles fitted to NO samples over different years at the Museum site

Estimated Model	Upper Percentiles MEAN	SD	MIN	MAX	Full Percentiles MEAN	SD	MIN	MAX	Maximum Percentile MEAN	SD	MIN	MAX
G3	0.187	0.084	0.057	0.326	0.502	0.807	0.044	2.453	0.202	0.069	0.120	0.360
G2	0.194	0.092	0.028	0.337	0.402	0.576	0.079	1.605	0.211	0.091	0.022	0.377
W3	0.186	0.084	0.035	0.311	0.122	0.032	0.061	0.188	0.214	0.083	0.032	0.357
W2	0.197	0.086	0.025	0.318	0.151	0.028	0.108	0.199	0.240	0.080	0.065	0.372
LN3	0.141	0.078	0.011	0.272	0.169	0.140	0.048	0.505	0.101	0.060	0.020	0.222
LN2	0.120	0.060	0.024	0.250	0.245	0.235	0.046	0.881	0.146	0.081	0.047	0.330

TABLE A2.4

Estimated parameter values, maximized log-likelihood values and RRMSE values for different percentiles fitted to NO_2 samples over different years at the Museum site

Year	n	Estimated Model	Shape α	Scale β	Location γ	Max(log L)	RRMSE for different percentiles				
							MAX1	MAX2	98	MEAN-U	MEAN-F
1975	92	G3	9.44	0.22	-0.63	-92.81	0.038	0.002	0.007	0.015	0.165
		G2	3.95	0.37	0.00	-95.12	0.069	0.084	0.022	0.041	0.233
		W3	2.22	1.66	0.00	-93.93	0.071	0.017	0.054	0.028	0.149
		W2	2.22	1.67	0.00	-93.93	0.072	0.018	0.055	0.028	0.148
		LN3	0.23	1.04	-1.42	-92.62	0.019	0.009	0.042	0.008	0.170
		LN2	0.57	0.26	0.00	-103.02	0.472	0.371	0.240	0.227	0.291
1976	287	G3	3.82	0.55	0.19	-401.18	0.356	0.244	0.048	0.214	0.048
		G2	4.71	0.49	0.00	-401.58	0.373	0.259	0.064	0.226	0.044
		W3	1.77	2.09	0.43	-403.72	0.384	0.260	0.058	0.224	0.047
		W2	2.19	2.59	0.00	-414.59	0.409	0.280	0.077	0.239	0.092
		LN3	0.35	1.02	-0.65	-400.92	0.340	0.237	0.046	0.209	0.048
		LN2	0.48	0.72	0.00	-404.06	0.250	0.163	0.029	0.151	0.071
1977	257	G3	21.10	0.22	-2.65	-367.66	0.310	0.095	0.107	0.170	0.372
		G2	2.77	0.75	0.00	-387.08	0.113	0.107	0.063	0.071	0.412
		W3	2.32	2.60	-0.24	-369.62	0.315	0.091	0.099	0.167	0.291
		W2	2.01	2.32	0.00	-372.59	0.285	0.061	0.075	0.149	0.377
		LN3	0.15	1.90	-4.71	-367.30	0.305	0.091	0.105	0.168	0.381
		LN2	0.75	0.54	0.00	-428.43	0.559	0.695	0.522	0.404	0.486
1978	241	G3	4.32	0.58	0.04	-368.19	0.163	0.271	0.002	0.201	0.073
		G2	4.47	0.57	0.00	-368.20	0.166	0.273	0.000	0.203	0.073
		W3	1.85	2.44	0.39	-368.96	0.206	0.292	0.016	0.216	0.068
		W2	2.21	2.88	0.00	-374.49	0.236	0.310	0.035	0.229	0.093
		LN3	0.32	1.26	-1.17	-368.56	0.154	0.271	0.002	0.203	0.076
		LN2	0.50	0.82	0.00	-373.40	0.024	0.162	0.118	0.132	0.102
1979	280	G3	13.39	0.18	-0.57	-276.51	0.216	0.196	0.068	0.172	0.053
		G2	7.41	0.25	0.00	-278.25	0.182	0.170	0.042	0.152	0.071
		W3	2.46	1.80	0.27	-282.26	0.239	0.206	0.071	0.177	0.059
		W2	2.87	2.09	0.00	-285.75	0.252	0.215	0.079	0.184	0.066
		LN3	0.20	1.19	-1.50	-275.88	0.205	0.191	0.063	0.169	0.053
		LN2	0.39	0.55	0.00	-286.79	0.059	0.085	0.036	0.089	0.098

TABLE A2.4 continued

Year	n	Estimated Model	Shape α	Scale β	Location γ	Max(log L)	RRMSE for different percentiles				
							MAX1	MAX2	98	MEAN-U	MEAN-F
1980	181	G3	6.81	0.34	-0.13	-227.19	0.119	0.175	0.049	0.124	0.055
		G2	6.01	0.37	0.00	-227.24	0.108	0.167	0.042	0.119	0.057
		W3	1.99	1.94	0.48	-227.74	0.146	0.186	0.054	0.131	0.056
		W2	2.57	2.48	0.00	-232.74	0.183	0.210	0.075	0.148	0.077
		LN3	0.26	1.19	-1.18	-227.05	0.109	0.172	0.048	0.122	0.055
		LN2	0.43	0.71	0.00	-230.71	0.041	0.074	0.042	0.066	0.080
1981	272	G3	19.87	0.19	-1.69	-334.14	0.058	0.115	0.008	0.073	0.047
		G2	5.15	0.40	0.00	-341.04	0.055	0.034	0.085	0.037	0.160
		W3	2.65	2.34	-0.02	-334.31	0.100	0.138	0.009	0.088	0.069
		W2	2.62	2.32	0.00	-334.32	0.098	0.137	0.008	0.087	0.074
		LN3	0.15	1.72	-3.58	-334.27	0.055	0.114	0.008	0.073	0.041
		LN2	0.49	0.62	0.00	-360.46	0.342	0.151	0.253	0.165	0.221
1982	275	G3	8.69	0.42	-0.59	-436.50	0.263	0.177	0.105	0.188	0.067
		G2	5.78	0.53	0.00	-437.85	0.234	0.153	0.082	0.170	0.085
		W3	2.24	3.07	0.31	-443.50	0.284	0.185	0.105	0.191	0.076
		W2	2.50	3.42	0.00	-446.50	0.295	0.193	0.112	0.197	0.084
		LN3	0.25	1.55	-1.84	-435.24	0.247	0.168	0.098	0.182	0.067
		LN2	0.44	1.02	0.00	-446.88	0.087	0.042	0.013	0.096	0.120
1983	281	G3	5.11	0.39	0.93	-347.00	0.008	0.065	0.054	0.111	0.019
		G2	11.50	0.26	0.00	-350.25	0.059	0.100	0.083	0.131	0.029
		W3	2.05	1.98	1.19	-350.22	0.060	0.093	0.072	0.122	0.029
		W2	3.40	3.27	0.00	-370.41	0.110	0.124	0.094	0.139	0.094
		LN3	0.33	0.92	0.29	-346.92	0.028	0.047	0.042	0.106	0.020
		LN2	0.30	1.03	0.00	-347.19	0.003	0.063	0.055	0.113	0.019

TABLE A2.5

Summary of tests and criteria applied to NO_2 samples by six distributions over different years at the Museum site

Selection Set	Estimated Model	Tests and Criteria							
		GIC	KL	AIC	SIC	CHI	KS	UPE	FPE
2- and 3-Parameter Models	G3	3	2	0	0	2	3	2	1
	G2	6	3	5	5	6	8	3	1
	W3	3	1	0	0	4	6	2	4
	W2	3	2	3	3	4	6	2	3
	LN3	1	1	0	0	1	1	0	0
	LN2	1	0	1	1	1	3	0	0
3-Parameter Models	G3	3	2	2	2	2	3	2	1
	W3	5	4	4	4	4	6	5	6
	LN3	1	1	1	1	1	1	0	0
2-Parameter Models	G2	6	5	5	5	6	8	4	3
	W2	3	3	3	3	4	6	3	5
	LN2	1	1	1	1	1	3	2	1

TABLE A2.6

Summary of RRMSE values for different percentiles fitted to NO_2 samples over different years at the Museum site

Estimated Model	Upper Percentiles				Full Percentiles				Maximum Percentile			
	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX
G3	0.141	0.062	0.015	0.214	0.100	0.104	0.019	0.372	0.170	0.117	0.008	0.356
G2	0.128	0.064	0.037	0.226	0.129	0.117	0.029	0.412	0.151	0.097	0.055	0.373
W3	0.149	0.060	0.028	0.224	0.094	0.077	0.029	0.291	0.200	0.108	0.060	0.384
W2	0.156	0.063	0.028	0.239	0.123	0.093	0.066	0.377	0.216	0.104	0.072	0.409
LN3	0.138	0.063	0.008	0.209	0.101	0.107	0.020	0.381	0.162	0.113	0.019	0.340
LN2	0.160	0.097	0.066	0.404	0.165	0.137	0.019	0.486	0.204	0.198	0.003	0.559

TABLE A2.7

Estimated parameter values, maximized log-likelihood values and RRMSE values for different percentiles fitted to NO_x samples over different years at the Museum site

Year	n	Estimated Model	Shape α	Scale β	Location γ	Max(log L)	RRMSE for different percentiles				
							MAX1	MAX2	98	MEAN-U	MEAN-F
1975	91	G3	4.42	1.33	-0.51	-215.28	0.154	0.167	0.184	0.086	0.436
		G2	3.43	1.56	0.00	-216.35	0.116	0.137	0.201	0.068	0.491
		W3	1.88	6.02	0.03	-219.09	0.171	0.166	0.216	0.090	0.285
		W2	1.89	6.05	0.00	-219.10	0.173	0.167	0.217	0.091	0.276
		LN3	0.39	1.87	-1.63	-213.77	0.092	0.133	0.205	0.061	0.515
		LN2	0.61	1.53	0.00	-222.37	0.226	0.092	0.032	0.092	0.612
1976	269	G3	2.49	2.63	0.50	-725.25	0.266	0.281	0.202	0.274	0.091
		G2	3.02	2.34	0.00	-726.88	0.289	0.299	0.219	0.287	0.092
		W3	1.50	7.09	0.70	-733.26	0.275	0.277	0.190	0.268	0.127
		W2	1.68	7.98	0.00	-741.80	0.298	0.291	0.202	0.277	0.177
		LN3	0.53	1.88	-0.53	-719.08	0.174	0.233	0.168	0.246	0.074
		LN2	0.59	1.78	0.00	-720.29	0.113	0.192	0.133	0.221	0.088
1977	274	G3	4.84	1.43	-0.40	-683.47	0.042	0.066	0.052	0.102	0.116
		G2	4.17	1.57	0.00	-684.02	0.023	0.052	0.039	0.094	0.137
		W3	2.12	7.16	0.20	-687.47	0.114	0.109	0.081	0.125	0.099
		W2	2.19	7.39	0.00	-688.10	0.121	0.114	0.085	0.128	0.092
		LN3	0.36	2.11	-2.25	-683.59	0.016	0.033	0.028	0.089	0.115
		LN2	0.53	1.75	0.00	-692.96	0.233	0.115	0.094	0.105	0.209
1978	228	G3	1.82	4.26	0.54	-674.75	0.200	0.298	0.221	0.269	0.071
		G2	2.22	3.73	0.00	-676.91	0.234	0.322	0.244	0.287	0.084
		W3	1.35	8.42	0.61	-677.55	0.228	0.309	0.226	0.275	0.087
		W2	1.49	9.26	0.00	-683.67	0.258	0.328	0.242	0.287	0.141
		LN3	0.63	2.00	-0.66	-674.44	0.043	0.220	0.164	0.227	0.075
		LN2	0.72	1.88	0.00	-675.50	0.071	0.153	0.104	0.194	0.095
1979	263	G3	3.46	1.57	0.35	-627.46	0.291	0.151	0.072	0.200	0.049
		G2	4.03	1.43	0.00	-627.96	0.306	0.165	0.084	0.209	0.046
		W3	1.78	5.77	0.65	-633.76	0.324	0.170	0.081	0.210	0.079
		W2	2.02	6.54	0.00	-640.09	0.342	0.184	0.093	0.219	0.122
		LN3	0.43	1.80	-0.89	-625.74	0.233	0.113	0.044	0.177	0.043
		LN2	0.52	1.62	0.00	-627.79	0.156	0.049	0.012	0.137	0.069

TABLE A2.7 continued

Year	n	Estimated Model	Shape α	Scale β	Location γ	Max(log L)	RRMSE for different percentiles MAX1	MAX2	98	MEAN-U	MEAN-F
1980	169	G3	3.07	1.63	0.85	-397.53	0.172	0.161	0.038	0.101	0.045
		G2	4.48	1.31	0.00	-398.61	0.212	0.193	0.068	0.123	0.041
		W3	1.63	5.10	1.29	-398.82	0.204	0.177	0.047	0.114	0.056
		W2	2.13	6.64	0.00	-407.14	0.254	0.212	0.076	0.139	0.112
		LN3	0.43	1.79	-0.67	-397.19	0.124	0.138	0.022	0.082	0.040
		LN2	0.49	1.65	0.00	-397.69	0.066	0.097	0.014	0.054	0.052
1981	268	G3	4.19	1.16	0.06	-588.78	0.354	0.260	0.181	0.251	0.088
		G2	4.32	1.14	0.00	-588.81	0.356	0.262	0.183	0.253	0.086
		W3	1.85	5.06	0.43	-600.87	0.364	0.256	0.168	0.246	0.121
		W2	2.02	5.56	0.00	-605.92	0.374	0.263	0.173	0.250	0.147
		LN3	0.40	1.69	-0.94	-584.08	0.311	0.234	0.163	0.234	0.088
		LN2	0.50	1.47	0.00	-588.48	0.228	0.167	0.104	0.187	0.119
1982	271	G3	2.43	3.40	0.55	-794.78	0.213	0.257	0.143	0.201	0.076
		G2	2.88	3.05	0.00	-795.99	0.237	0.276	0.161	0.215	0.073
		W3	1.51	8.86	0.83	-800.03	0.238	0.266	0.144	0.204	0.103
		W2	1.69	9.92	0.00	-807.02	0.266	0.284	0.160	0.218	0.146
		LN3	0.54	2.12	-0.87	-791.57	0.108	0.203	0.102	0.165	0.065
		LN2	0.62	1.99	0.00	-793.43	0.016	0.143	0.047	0.126	0.092
1983	261	G3	4.31	1.63	0.77	-667.20	0.077	0.189	0.139	0.179	0.051
		G2	5.46	1.43	0.00	-668.07	0.100	0.204	0.152	0.189	0.052
		W3	1.95	7.30	1.34	-674.11	0.121	0.207	0.148	0.187	0.072
		W2	2.35	8.82	0.00	-682.73	0.148	0.223	0.160	0.196	0.117
		LN3	0.39	2.07	-0.74	-665.52	0.011	0.155	0.114	0.163	0.049
		LN2	0.44	1.96	0.00	-666.11	0.033	0.129	0.092	0.150	0.056
1984	82	G3	3.03	1.62	2.66	-191.68	0.097	0.006	0.008	0.037	2.140
		G2	8.04	0.94	0.00	-193.49	0.002	0.071	0.061	0.027	1.189
		W3	1.71	5.13	3.01	-191.61	0.027	0.045	0.036	0.020	0.096
		W2	2.89	8.51	0.00	-198.34	0.061	0.101	0.078	0.044	0.104
		LN3	0.43	1.78	1.12	-192.23	0.161	0.025	0.011	0.062	0.289
		LN2	0.36	1.96	0.00	-192.46	0.096	0.016	0.021	0.037	0.213

TABLE A2.8

Summary of tests and criteria applied to NO_x samples by six distributions over different years at the Museum site

Selection Set	Estimated Model	Tests and Criteria							
		GIC	KL	AIC	SIC	CHI	KS	UPE	FPE
2- and 3-Parameter Models	G3	7	3	0	0	5	7	5	2
	G2	8	2	2	2	8	9	1	4
	W3	2	1	0	0	4	7	1	1
	W2	0	0	0	0	4	6	1	2
	LN3	1	0	0	0	0	0	0	0
	LN2	8	4	8	8	8	9	2	1
3-Parameter Models	G3	7	7	7	7	5	7	7	7
	W3	3	3	3	3	4	7	3	3
	LN3	0	0	0	0	0	0	0	0
2-Parameter Models	G2	8	2	2	2	8	9	4	6
	W2	0	0	0	0	4	6	2	3
	LN2	8	8	8	8	8	9	4	1

Table A2.9

Summary of RRMSE values for different percentiles fitted to NO_x samples over different years at the Museum site

Estimated Model	Upper Percentiles				Full Percentiles				Maximum Percentile			
	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX
G3	0.170	0.080	0.037	0.274	0.316	0.618	0.045	2.140	0.186	0.094	0.042	0.354
G2	0.175	0.088	0.027	0.287	0.229	0.344	0.041	1.189	0.187	0.115	0.002	0.356
W3	0.174	0.080	0.020	0.275	0.113	0.061	0.056	0.285	0.207	0.097	0.027	0.364
W2	0.185	0.077	0.044	0.287	0.143	0.050	0.092	0.276	0.229	0.095	0.061	0.374
LN3	0.150	0.069	0.061	0.246	0.135	0.144	0.040	0.515	0.127	0.091	0.011	0.311
LN2	0.130	0.057	0.037	0.221	0.160	0.160	0.052	0.612	0.124	0.078	0.016	0.233

TABLE A2.10

Estimated parameter values, maximized log-likelihood values and RRMSE values for different percentiles fitted to SO_2 samples over different years at the Museum site

Year	n	Estimated Model	Shape α	Scale β	Location γ	Max(log L)	RRMSE for different percentiles				
							MAX1	MAX2	98	MEAN-U	MEAN-F
1975	26	G3	21.26	0.10	-1.13	-16.63	0.209	0.158	0.200	0.079	0.305
		G2	3.63	0.28	0.00	-17.99	0.091	0.267	0.327	0.035	0.331
		W3	2.26	1.17	-0.03	-17.05	0.196	0.184	0.353	0.074	0.270
		W2	2.20	1.14	0.00	-17.06	0.192	0.187	0.353	0.073	0.280
		LN3	0.15	1.08	-1.98	-16.55	0.206	0.158	0.368	0.078	0.312
		LN2	0.61	-0.13	0.00	-20.88	0.188	0.492	0.245	0.071	0.368
1976	149	G3	0.37	1.35	0.04	36.34	0.907	0.268	0.007	0.358	0.512
		G2	1.25	0.43	0.00	-53.03	0.142	0.106	0.203	0.105	0.349
		W3	0.51	0.38	0.04	9.91	2.365	0.889	0.378	0.965	0.908
		W2	1.18	0.56	0.00	-52.41	0.084	0.132	0.216	0.104	0.369
		LN3	0.93	-0.94	-0.03	-61.26	0.598	0.057	0.129	0.233	0.460
		LN2	1.08	-1.08	0.00	-61.85	0.901	0.176	0.063	0.348	0.510
1977	174	G3	267.55	0.03	-6.09	-98.11	0.362	0.270	0.212	0.202	0.633
		G2	2.33	0.37	0.00	-121.62	0.135	0.107	0.090	0.092	0.723
		W3	2.82	1.27	-0.26	-99.07	0.355	0.261	0.202	0.196	0.634
		W2	1.97	0.97	0.00	-106.67	0.304	0.227	0.181	0.173	0.855
		LN3	0.04	2.27	-8.79	-98.07	0.361	0.270	0.212	0.201	0.641
		LN2	0.86	-0.37	0.00	-156.89	0.474	0.261	0.170	0.206	0.768
1978	153	G3	3.90	0.21	-0.18	-65.90	0.014	0.165	0.200	0.102	0.314
		G2	1.82	0.34	0.00	-66.65	0.151	0.091	0.152	0.091	0.305
		W3	1.66	0.72	-0.02	-61.58	0.022	0.178	0.206	0.107	0.329
		W2	1.57	0.69	0.00	-61.68	0.000	0.167	0.199	0.102	0.337
		LN3	0.22	0.55	-1.16	-67.02	0.054	0.202	0.224	0.119	0.398
		LN2	0.91	-0.77	0.00	-84.70	0.846	0.204	0.025	0.329	0.459
1979	77	G3	0.26	0.88	0.04	128.12	0.364	0.086	0.333	0.141	0.279
		G2	1.06	0.25	0.00	24.71	0.201	0.343	0.446	0.150	0.410
		W3	0.36	0.11	0.04	114.13	2.845	0.793	0.081	1.116	1.061
		W2	0.98	0.26	0.00	24.68	0.164	0.322	0.433	0.137	0.341
		LN3	5.38	-4.24	0.04	87.35	774.212	59.928	13.560	293.499	857.192
		LN2	1.04	-1.86	0.00	31.55	0.026	0.305	0.462	0.116	0.312

TABLE A2.10 continued

Year	n	Estimated Model	Shape α	Scale β	Location γ	Max(log L)	RRMSE for different percentiles				
							MAX1	MAX2	98	MEAN-U	MEAN-F
1980	124	G3	0.60	0.44	0.04	54.21	0.189	0.017	0.136	0.072	0.241
		G2	1.74	0.17	0.00	33.33	0.437	0.237	0.278	0.188	0.151
		W3	0.78	0.24	0.04	47.18	0.140	0.015	0.116	0.053	0.223
		W2	1.32	0.33	0.00	31.20	0.442	0.228	0.263	0.188	0.128
		LN3	0.79	-1.46	-0.01	34.23	0.295	0.160	0.250	0.127	0.187
		LN2	0.83	-1.50	0.00	34.16	0.267	0.140	0.239	0.114	0.193
1981	174	G3	0.60	0.98	0.04	-62.00	0.175	0.396	0.196	0.209	0.334
		G2	1.45	0.43	0.00	-85.65	0.419	0.529	0.329	0.307	0.190
		W3	0.78	0.54	0.04	-72.86	0.128	0.376	0.177	0.195	0.295
		W2	1.23	0.67	0.00	-86.39	0.434	0.531	0.326	0.310	0.141
		LN3	0.71	-0.56	-0.10	-89.20	0.343	0.510	0.327	0.288	0.254
		LN2	0.97	-0.85	0.00	-93.79	0.116	0.411	0.236	0.215	0.331
1982	183	G3	0.63	0.79	0.04	-39.17	0.056	0.248	0.232	0.225	0.301
		G2	1.56	0.34	0.00	-59.58	0.257	0.414	0.359	0.310	0.154
		W3	0.81	0.46	0.04	-49.28	0.086	0.238	0.223	0.221	0.254
		W2	1.30	0.58	0.00	-60.15	0.285	0.421	0.359	0.314	0.113
		LN3	0.70	-0.71	-0.08	-63.56	0.149	0.384	0.350	0.294	0.209
		LN2	0.93	-0.97	0.00	-67.36	0.093	0.282	0.279	0.250	0.281
1983	78	G3	0.42	0.46	0.04	80.79	0.425	0.099	0.117	0.165	0.351
		G2	2.08	0.11	0.00	46.26	0.194	0.255	0.324	0.121	0.192
		W3	0.58	0.16	0.04	68.62	1.070	0.427	0.085	0.435	0.495
		W2	1.48	0.26	0.00	44.86	0.215	0.254	0.314	0.126	0.159
		LN3	0.70	-1.65	-0.01	46.45	0.036	0.195	0.303	0.075	0.211
		LN2	0.75	-1.73	0.00	46.35	0.014	0.171	0.289	0.065	0.213
1984	65	G3	0.38	1.11	0.04	23.43	0.983	0.304	0.016	0.372	0.506
		G2	1.38	0.33	0.00	-12.59	0.166	0.093	0.206	0.063	0.357
		W3	0.53	0.33	0.04	12.02	2.354	0.890	0.363	0.890	0.827
		W2	1.26	0.49	0.00	-12.09	0.086	0.129	0.224	0.032	0.383
		LN3	0.80	-0.95	-0.05	-16.20	0.468	0.012	0.160	0.177	0.425
		LN2	1.02	-1.18	0.00	-16.74	0.875	0.176	0.071	0.331	0.465

TABLE A2.11

Summary of tests and criteria applied to SO_2 samples by six distributions over different years at the Museum site

Selection Set	Estimated Model	Tests and Criteria							
		GIC	KL	AIC	SIC	CHI	KS	UPE	FPE
2- and 3-Parameter Models	G3	7	7	7	7	0	0	0	1
	G2	1	0	0	0	5	7	6	4
	W3	0	0	0	0	0	0	0	0
	W2	3	3	3	3	6	8	4	5
	LN3	0	0	0	0	0	0	0	0
	LN2	0	0	0	0	2	6	0	0
3-Parameter Models	G3	7	7	7	7	0	0	7	4
	W3	0	0	0	0	0	0	0	3
	LN3	0	0	0	0	0	0	0	0
2-Parameter Models	G2	7	2	2	2	5	7	6	4
	W2	8	5	5	5	6	8	4	5
	LN2	3	3	3	3	2	6	0	1

Table A2.12

Summary of RRMSE values for different percentiles fitted to SO_2 samples over different years at the Museum site

Estimated Model	Upper Percentiles				Full Percentiles				Maximum Percentile			
	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX
G3	0.192	0.100	0.072	0.372	0.378	0.121	0.241	0.633	0.368	0.315	0.014	0.983
G2	0.146	0.091	0.035	0.310	0.316	0.162	0.151	0.723	0.219	0.113	0.091	0.437
W3	0.425	0.387	0.053	1.116	0.530	0.293	0.223	1.061	0.956	1.070	0.022	2.845
W2	0.156	0.089	0.032	0.314	0.311	0.208	0.113	0.855	0.221	0.140	0.000	0.442
LN3	29.509	87.997	0.075	293.499	86.029	257.054	0.187	857.192	77.672	232.180	0.036	774.212
LN2	0.204	0.104	0.065	0.348	0.390	0.161	0.193	0.768	0.380	0.347	0.014	0.901

TABLE A2.13

Summary of tests and criteria applied to NO samples by six distributions over different years and sites with 26 annual data sets

Selection Set	Estimated Model	Tests and Criteria							
		GIC	KL	AIC	SIC	CHI	KS	UPE	FPE
2- and 3-Parameter Models	G3	19	13	9	5	13	17	8	3
	G2	9	1	4	6	16	20	3	5
	W3	11	1	1	0	12	18	4	7
	W2	3	1	1	1	12	19	10	8
	LN3	3	2	0	0	3	5	0	0
	LN2	16	8	11	14	21	24	1	3
3-Parameter Models	G3	24	21	21	21	13	17	11	9
	W3	16	2	2	2	12	18	14	15
	LN3	3	2	2	2	3	5	0	1
2-Parameter Models	G2	11	7	7	7	16	20	9	6
	W2	7	2	2	2	12	19	15	13
	LN2	20	17	17	17	21	24	2	7

TABLE A2.14

Summary of RRMSE values for different percentiles fitted to NO samples over different years and sites with 26 annual data sets

Estimated Model	Upper Percentile				Full Percentile				Maximum Percentile			
	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX
G3	0.165	0.105	0.002	0.382	0.979	1.725	0.044	8.009	0.189	0.137	0.006	0.714
G2	0.176	0.105	0.002	0.396	0.768	1.118	0.079	4.816	0.176	0.116	0.002	0.377
W3	0.169	0.094	0.000	0.341	0.202	0.155	0.061	0.688	0.212	0.155	0.001	0.759
W2	0.170	0.100	0.019	0.368	0.214	0.151	0.108	0.888	0.181	0.098	0.013	0.372
LN3	0.207	0.145	0.011	0.588	0.511	1.004	0.048	5.109	0.342	0.390	0.020	1.397
LN2	0.268	0.193	0.024	0.938	0.723	1.404	0.046	7.146	0.541	0.514	0.047	2.165

TABLE A2.15

Summary of tests and criteria applied to NO_2 samples by six distributions over different years and sites with 18 annual data sets

Selection Set	Estimated Model	Tests and Criteria GIC	KL	AIC	SIC	CHI	KS	UPE	FPE
2- and 3-Parameter Models	G3	6	3	0	0	3	6	2	2
	G2	11	5	9	9	12	15	7	2
	W3	5	3	0	0	6	11	2	5
	W2	9	5	7	7	8	15	6	9
	LN3	1	1	0	0	1	1	0	0
	LN2	3	1	2	2	3	8	1	0
3-Parameter Models	G3	6	4	4	4	3	6	4	3
	W3	8	7	7	7	6	11	8	9
	LN3	1	1	1	1	1	1	0	0
2-Parameter Models	G2	11	9	9	9	12	15	8	4
	W2	9	7	7	7	8	15	7	12
	LN2	3	2	2	2	3	8	3	2

TABLE A2.16

Summary of RRMSE values for different percentiles fitted to NO_2 samples over different years and sites with 18 annual data sets

Estimated Model	Upper Percentile MEAN	SD	MIN	MAX	Full Percentile MEAN	SD	MIN	MAX	Maximum Percentile MEAN	SD	MIN	MAX
G3	0.126	0.058	0.015	0.214	0.207	0.353	0.019	1.596	0.157	0.102	0.008	0.356
G2	0.114	0.062	0.010	0.226	0.265	0.413	0.029	1.878	0.152	0.098	0.026	0.373
W3	0.133	0.059	0.002	0.224	0.112	0.068	0.029	0.291	0.176	0.108	0.004	0.384
W2	0.138	0.062	0.025	0.239	0.145	0.093	0.066	0.392	0.187	0.113	0.003	0.409
LN3	0.121	0.059	0.008	0.209	0.135	0.115	0.020	0.381	0.145	0.096	0.019	0.340
LN2	0.193	0.142	0.028	0.593	0.265	0.256	0.019	1.081	0.327	0.319	0.003	1.128

TABLE A2.17

Summary of tests and criteria applied to NO_x samples by six distributions over different years and sites with 26 annual data sets

Selection Set	Estimated Model	Tests and Criteria							
		GIC	KL	AIC	SIC	CHI	KS	UPE	FPE
2- and 3-Parameter Models	G3	16	7	2	1	14	18	8	3
	G2	13	4	7	7	20	21	3	6
	W3	7	2	0	0	10	21	2	5
	W2	3	1	1	1	11	18	7	5
	LN3	7	5	4	2	6	8	1	4
	LN2	15	7	12	15	22	25	5	3
3-Parameter Models	G3	17	14	14	14	14	18	13	11
	W3	10	6	6	6	10	21	11	9
	LN3	8	6	6	6	6	8	2	6
2-Parameter Models	G2	13	7	7	7	20	21	9	9
	W2	3	1	1	1	11	18	8	8
	LN2	20	18	18	18	22	25	9	9

TABLE A2.18

Summary of RRMSE values for different percentiles fitted to NO_x samples over different years and sites with 26 annual data sets

Estimated Model	Upper Percentile				Full Percentile				Maximum Percentile			
	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX	MEAN	SD	MIN	MAX
G3	0.153	0.087	0.022	0.352	0.380	0.672	0.045	2.649	0.161	0.092	0.002	0.354
G2	0.165	0.091	0.024	0.377	0.391	0.632	0.041	2.773	0.182	0.089	0.002	0.356
W3	0.155	0.086	0.020	0.336	0.137	0.095	0.056	0.542	0.178	0.094	0.026	0.364
W2	0.163	0.090	0.017	0.348	0.185	0.111	0.078	0.675	0.194	0.093	0.044	0.374
LN3	0.158	0.086	0.032	0.338	0.203	0.282	0.040	1.396	0.196	0.222	0.008	0.895
LN2	0.142	0.073	0.037	0.342	0.232	0.274	0.050	1.275	0.179	0.177	0.001	0.802

Bibliography

[1] Akaike, H. (1974). A new look at the statistical model identification, IEEE Trans. Auto. Control AC-19, 716-723.

[2] Anderson, M. (1985). POLDIF: Dispersion of Air and Noise Pollution from Urban Road Networks, CSIRO Division of Building Research, Sydney.

[3] Anderson, M. and Taylor, M.A.P. (1988). Estimating the environmental impacts of road traffic at the local level: A PC-based modelling system, In Desktop Planning: Microcomputer Applications for Infrastructure and Services Planning and Management, P.W. Newton, M.A.P. Taylor and R. Sharpe (eds), Hargreen Pub. Co., Melbourne, 251-261.

[4] Aitchison, J. and Brown, J.A.C. (1957). The Lognormal Distribution, Cambridge, Cambridge University Press.

[5] Atkinson A.C. (1970). A method for discriminating between models (with discussion), J. R. Statist. Soc. B 32, 323-353.

[6] Bai, J., Jakeman, A.J. and McAleer, M. (1989). A new approach to maximum likelihood estimation of the three-parameter gamma and Weibull distributions, Working Paper in Economics and Econometrics No. 191, Australian National University, November 1989, pp. 26.

[7] Bai, J., Jakeman, A.J. and McAleer, M. (1989). The estimating the percentiles of some misspecified non-nested distributions, Working Papers in Economics and Econometrics No. 193, Australia National University, December 1989.

[8] Bai, J., Jakeman, A.J. and McAleer, M. (1990). The effects of misspecification in estimating the percentiles of some two- and three-parameter distributions, Math. Comp. Simulation 32, 194-199.

[9] Bai, J., Jakeman, A.J. and McAleer, M. (1990). Discrimination between nested two- and three-parameter distributions: An application to models of air pollution, Working Paper in Economics and Econometrics No. 197, Australian National University, March 1990, pp. 23.

[10] Bai, J., Jakeman, A.J. and McAleer M. (1990). Discrimination procedure for fitting nested and non-nested distributions to environmental quality data, Working Papers in Economics and Econometrics No. 200, Australian National University, March 1990, pp. 58.

[11] Bai, J., Jakeman A.J. and Taylor, J.A. (1988). Statistical distribution modelling: Function, methods and application to air quality management, Math. Comp. Simulation 30, 3-9.

[12] Bai, J., Jakeman A.J. and Taylor, J.A. (1990). Percentile estimation of the three-parameter gamma and lognormal distribution: methods of moments versus maximum likelihood, Math. Comp. Simulation 32, 164-169.

[13] Bai, J. and Taylor, J.A. (1986). Estimation of the parameters and upper percentiles of statistical distributions applicable to air quality data, part I: The gamma distribution, CRES Working Paper, No 1986/39, Australian National University, pp. 131.

[14] Bain, L.J. (1978). Statistical Analysis of Reliability and Life-testing Models, Marcel Dekker, New York.

[15] Bankoff, S.G. and Hanzevack, E.L. (1973). Parameter updating for air pollution dispersion model, Proc. Symp. Am. Inst. Chem. Eng. Natl Meet. Air Pollut. Modelling, New Orleans, La.

[16] Bankoff, S.G. and Hanzevack, E.L. (1975). The adaptive filtering transport model for prediction and control of pollutant concentration in an urban airshed, Atmos. Environ. 9, 793-808.

[17] Barlow, R.E. (1971). Average time and maxima for air pollution concentration, Operations Res. Cent., Univ. Calif., Berkeley, Calif. Rep., No. ORC-71-17 (Natl Tech. Inf. Serv., No. NTIS AD-729-413).

[18] Barry, P.J. (1971). Use of argon-41 to study the dispersion of stack effluents, Proc. Symp. Nucl. Tech. Environ. Pollut., Int. At. Energy Agency, Vienna, 241-253.

[19] Barth, S.D. (1970). Federal motor vehicle emission goal for carbon monoxide, hydrocarbon and NO_x based on desired air quality levels, J. Air Pollut. Control Assoc., 20, 519-23.

[20] Benarie, M.M. (1976). Urban air pollution modelling without computers, U.S. Environ. Prot. Agency, Publ. No. EPA-600/4-76-055.

[21] Benarie, M.M. (1978). The simple box model revisited, Atmos. Environ. 12, 1929-1930.

[22] Benarie, M.M. (1980). Urban Air Pollution Modelling, Macmillan, London and Basingstoke, 143-146.

[23] Benarie, M.M. (1980a). The simple box model simplified. In Air Pollution 1980, Proc. 14th Int. Colloq., Paris. Studies in Environmental Science, Vol. 8, M.M. Benarie (ed.), Elsevier, Amsterdam.

[24] Bencala, K.E. and Seinfeld, J.H. (1976). On frequency distributions of air pollutant concentrations, Atmos. Environ. 10, 941-950.

[25] Box, G.E.P. and Jenkins, G.M. (1976). Time Series Analysis: Forecasting and Control, Revised Edition, Holden Day, San Francisco.

[26] Box, G.E.P. and Draper, N.R. (1987). Empirical Model-Building and Response Surfaces, John Wiley, New York.

[27] Breiman, L., and Meisel, W.S. (1976). Empirical techniques for analyzing air quality and meteorological data, part III, short-term changes in ground-level ozone concentrations: an empirical analysis, U.S. Environ. Prot. Agency, Publ., No. EPA-600/4-76-029c, pp. 73.

[28] Briggs, G.A. (1974). Diffusion Estimates for Small Emissions, Draft report, U.S. Atomic Energy Commission ATDL Contribution File No. 79, Oak Ridge, Tennessee, USA.

[29] Burt, E.W. and Slater, H.H. (1977). Evaluation of the Valley Model, Joint AMS/APCA Conf. on Appl. Air Pollut. Met., Salt Lake City, Utah.

[30] Bury, K.V. (1975). Statistical Models in Applied Science, John Wiley, New York.

[31] Carras, J., Ross, G., Sawford, B. and Simpson, R.W. (1989). Methods of source-receptor modelling for air quality applications, In Air Pollutant from Energy Industries:Scientific Basis of Standards and Reserach Needs, A.J. Jakeman (ed.), Centre for Resource and Environmental Studies, Australian National University, Canberra.

[32] Chambers, L.A. (1976). The pollutant, In Air Pollution, Vol. I, A.C. Stern (ed.), Academic Press, London, 3-22.

[33] Chang, T.Y. and Weinstock, B. (1973). Urban carbon monoxide concentration and vehicle emissions, J. Air Pollut. Control Assoc. 23, 691-696.

[34] Chang, T.Y. and Weinstock, B. (1974). Rollback modelling for urban air pollution control, Proc. Symp. Atmos. Diffusion Air Pollut., Santa Barbara, September 1974, Am. Met. Soc., Boston, Mass., 184-189.

[35] Cheng, R.C.H. and Amin, N.A.K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin, J. R. Statist. Soc. B 45, 394-403.

[36] Chock, D.P. (1978). A simple line-source model for dispersion near roadways, Atmos. Environ. 12, 823-829.

[37] Cohen, A.C. and Norgaard, N.J. (1977). Progressively censored sampling in the three-parameter gamma distribution, Technometrics 19, 333-340.

[38] Cohen, A.C. and Whitten, B.J. (1982). Modified moment and maximum likelihood estimators for parameters of the three-parameter gamma distribution, Commun. Statist. B-11, 197-216.

[39] Cox, D.R. (1961). Tests of separate families of hypotheses, Proc. 4th Berk. Sym. Math. Statist. and Prob. 1, Univ. of Calif. Press, Berkeley, 105-123.

[40] Cox, D.R. (1962). Further results on tests of separate families of hypotheses, J. R. Statist. Soc. B 24, 406-424.

[41] Cramer, H. (1946). Mathematical methods of statistics, Princeton University Press, Princeton.

[42] Curran, C.T. and Frank, H.N. (1975). Assessing the validity of log-normal model when predicting maximum air pollution concentrations, Proc. 68th Ann. Meet. Air Pollut. Control Assoc., Boston, Mass., June 1975, Paper, No. 75-51.3, pp. 17.

[43] Dabbert, W.F., Ludwig, F.L. and Johnson, W.B. (1973). Validation and Applications of An Urban Diffusion Model for Vehicular Pollutants, Atmos. Environ. 7, 603-618.

[44] Daly, N.J. and Steele, L.P. (1976). A predictive model for CO in Canberra, In Symposium on Air Pollution Diffusion Modelling, Australian Environment Council, Canberra, 264-275.

[45] Daw, F.A. and O'Laughlin, K.J. (1972). ACT: Dispersion climatology and meteorological effects on the distribution of air pollution, International Clean Air Conference, Clean Air Society of Australia and New Zealand, Melbourne.

[46] de Nevers, N.H. and Morris, J.R. (1973). Modified rollback models, Proc. 66th Ann. Meet. Air Pollut. Control Assoc., Chicago, Ill., June 1973, Paper, No. 73-139.

[47] de Nevers, N.H. and Morris, J.R. (1975). Rollback modelling: basic and modified, J. Air Pollut. Control Assoc. 25, 943-947.

[48] de Nevers N.H., Neligan R.E. and Slater H.H. (1977). Air quality management, pollution control strategies, modelling and evaluation, In Air Pollution, Vol. V, A.C. Stern (ed.), Academic Press, London, 3-40.

[49] Drake, R.L., McNaughton, D.J. and Huang C. (1979). Mathematical Models for Atmospheric Pollutants, Appendix D: Available Air Quality Models. EPRI report No. EA-1131.

[50] Dubey, S.D. (1966). Asymptotic efficiencies of the moment estimators for the parameters of the Weibull Laws, Naval Research Logist. Quart. 13, 265-288.

[51] Dumonceaux, R. and Antle, C.E. (1973). Discrimination between the lognormal and the Weibull distributions, Technometrics 15, 923-926.

[52] Durbin, P.A. and Hunt, J.C.R. (1980). Dispersion from elevated sources in turbulent boundary layers, J. Mecanique 19, 679.

[53] Egan, B.A. and Mahoney J.R. (1972). Numerical Modelling of Advection and Diffusion of Urban Area Source Pollutants, J.A.M. 11, 312-321.

[54] Einstein, A. (1905). Uber die von molekularkinetischen Theorie der Warme geforderte Bewegung von in ruhenden Flussigkeiten suspendierten, Teilchen. Ann. Phys. 17, 549.

[55] Epps, T.W., Singleton, K.J. and Pulley, L.B. (1982). A test of separate families of distributions based on the empirical moment generating function, Biometrika 69, 391-399.

[56] Eschenroeder, A.Q. (1975). An assessment of models for predicting air quality, ERT Document ERTW-75-03, Environment Research Techniques, Concord.

[57] Faith, W.L. and Atkisson, A.A. (1972). Air Pollution, Wiley (Interscience), New York.

[58] Ferrari, L., Hawke, G., Heggie, A. and Johnson, D. (1986). Review of air quality monitoring in Canberra: a report to the National Capital Development Commission, State Pollution Control Commission.

[59] Finlayson-Pitts, B.J. and Pitts, J.N. (1986). Atmospheric Chemistry: Fundamentals and Experimental Techniques, John Wiley, New York.

[60] Fox, D.G. (1981). Judging air quality model performance, Bull. Am. Met. Soc. 62, 599-609.

[61] Funabashi, N. (1973). A real-time air pollution prediction by a multi-box model, (Text in Japanese). Proc. 14th Ann. Meet. Japan Soc. Air Pollut., Fukushima, November 1973, Paper, No. 50.

[62] Georgopoulos P.G. and Seinfeld J.H. (1982). Statistical distributions of air pollutant concentrations, Environ. Sci. Tech. 16, 401-416A.

[63] Geraghty D. and Ricci P.F. (1984). Mathematical models for atmospheric pollutants, In Principles of Health Risk Assessment, P.F. Ricci (ed.), Prentice-Hall, Englewood Cliffs, New Jersey, 67-115.

[64] Gifford, F.A. (1970). Atmospheric Diffusion in an Urban Area, U.S. Atomic Energy Commission ATDL Contribution No. 33, Oak Ridge, Tennessee, USA.

[65] Gifford, F.A. (1973). The Simple ATDL Urban Diffusion Model, In Proc. 4th Meet. Expert Panel on Air Pollut. Modelling, Oberursel, Federal Republic of Germany, NATO Committee on the Challenges of Modern Society, Pub. No. 30, Brussels, Belgium.

[66] Gifford F.A. and Hanna R.S. (1971). Urban Air Pollution Modelling, Proc. 2th Int. Clean Air Congr., H.M. England and W.T. Beery (eds). Academic Press, New York, 1146-1151.

[67] Gifford F.A. and Hanna R.S. (1973). Modelling urban air pollution, Atmos. Environ. 7, 131-136.

[68] Graedel, T.E., Kleiner, B. and Patterson, C.C. (1974). Measurements of extreme concentration of tropospheric hydrogen sulphide, J. Geophys. Res. 79, 4467.

[69] Gualdi R. and Tebaldi S. (1982). Real-time control of air pollution: the case of Milan, In Mathematical Models for Planning and Controlling Air Quality, G. Fronza and P. Melli (eds), Pergamon, Oxford, 233-246.

[70] Hall, C.D. (1975). The simulation of particle motion in the atmosphere by a numerical random walk model, Quart. J. R. Met. Soc. 101, 235-244.

[71] Hameed, S. (1974). A modified multi-cell method for simulation of atmospheric transport, Atmos. Environ. 8, 1003-1008.

[72] Hamming, W.J., Chass, R.L., Dickenson, J.E. and Macbeth, W.G. (1973). Motor vehicle control and air quality. The path to clean air for Los Angeles, Proc. 66th Ann. Meet. Air Pollut. Control Assoc., Chicago, Ill., June 1973, Paper, No. 73-73.

[73] Hanna, R.S. (1968). A method of estimating vertical eddy transport in the planetary boundary layer using characteristics of the vertical velocity spectrum, J. Atmos. Sci. 25, 1026.

[74] Hanna, R.S. (1971). A Simple Method of Calculating Dispersion from Urban Sources, J. Air Pollut. Control Assoc. 21, 774-777.

[75] Hanna, R.S. (1973). Urban Air Pollution Models - Why? U.S. Atomic Energy Commission ATDL Contribution No. 83, Oak Ridge, Tennessee, USA.

[76] Hanna, R.S. (1978). Urban Modelling of Inert Substances. Air Quality Meteorology and Atmospheric Ozone, ASTM STP 653, Morris, A.L. and R.C. Burras (eds), American Society for Testing and Materials, 262-275.

[77] Hanna, R.S. (1978). A statistical diffusion model for use with variable wind fields, 4th Symposium on Turbulent Diffusion and Air Pollution, American Meteorological Society, Boston.

[78] Hanna R.S. (1982). Review of Atmospheric Diffusion Models for Regulatory Applications, World Meteorological Organisation Technical Note No. 177, WMO-No. 581, Secretariat of the World Meteorological Organisation, Geneva.

[79] Hanna, R.S., (1982). Natural variability of observed SO and CO concentrations in Saint Louis, Atmos. Environ. 16, 1435-1440.

[80] Hanna, R.S., Briggs, G.A. and Hosker R.P. (1982). Handbook on atmospheric dispersion, Technical Information Centre, U.S. Department of Energy.

[81] Hanna, R.S. and Gifford, F.A. (1977). Application of the ATDL Simple Urban Dispersion Model to Frankfurt, West Germany. ATDL Contribution No. 77/17, Oak Ridge, Tennessee, USA.

[82] Harter, H.L. and More, A.H. (1965). Maximum likelihood estimation of the parameters of gamma and Weibull populations from complete and from censored samples, Technometics., 7, 639-643.

[83] Horie, Y. and Overton, J.H. (1974). The effect on rollback models due to distribution of concentration, Proc. Symp. Stat. Aspects Air Quality Data, U.S. Environ. Prot. Agency, Publ. No. EPA-650/4-74-038, pp. 15-1-15-18.

[84] Horowitz, J.L. and McAleer, M. (1988). A simple method for testing a general parametric model against a non-nested alternative, Working Paper in Economics and Econometrics No. 166, Australian National University, July 1988, pp. 49.

[85] Huang, C.H. (1979). A theory of dispersion in turbulent shear flow, Atmos. Environ., 13, 453-463.

[86] Hunt, J.C.R. and Weber, A.H. (1979). A Lagrangian statistical analysis of diffusion from a ground-level source in a turbulent boundary layer, Quart. J. R. Met. Soc. 105, 423-443.

[87] Ishikawa, Y., Yagi, Y., Oshio, T., Kure, F. and Yoneda, A. (1973). Approaches to the design of calculating analysis on air pollution in Osaka, part 3. J. Japan Soc. Air Pollut. 8, 514 (in Japanese).

[88] Jakeman A.J., Bai, J. and Taylor, J.A. (1988). On the variability of the wind speed exponent in urban air pollution models, Atmos. Environ. 22, 2013-2019.

[89] Jakeman, A.J. and Simpson, R.W. (1987). Air quality and resource development: a risk assessment in the Hunter Region in Australia, CRES Monograph 16, Centre for Resource and Environmental Studies, Australian National University, Canberra.

[90] Jakeman, A.J., Simpson, R.W. and Taylor, J.A. (1988). Modelling distributions of air pollutant concentrations: III the hybrid deterministic-statistical distribution approach, Atmos. Environ. 22, 163-174.

[91] Jakeman, A.J. and Taylor, J.A. (1989). Identification, estimation and simulation of frequency distributions of pollutant concentrations for air quality management, In P.N. Cheremisinoff (ed.), Encyclopedia of Environmental Control Technology, Gulf Publishing, Houston, 135-158.

[92] Jakeman, A.J., Taylor, J.A. and Simpson, R.W. (1986). Modelling distributions of air pollutant concentrations - II, Estimation of one and two parameter statistical distributions. Atmos. Environ. 20, 2435-2447.

[93] Jakeman, A.J., Whitehead, P.G., Robson, A., Jenkins, A. and Taylor, J.A. (1989). A method for predicting the extremes of stream acidity and other water quality variables, J. Hydro. (in press).

[94] Jakeman, A.J. and Young, P.C. (1984). Recursive filtering and smoothing procedures for the inversion of ill-posed causal problems, Util. Math. 25, 351-376.

[95] Johnson, N.L and Kotz, S. (1970). Distributions in Statistics: Continuous Univariate Distributions-1, Houghton Mifflin, Boston.

[96] Johnson W.B., Sklarew R.C. and Turner D.B. (1976). Urban air quality simulation modelling, In Air Pollution, Vol. I, A.C. Stern (ed.), Academic Press, London, 503-562.

[97] Kalman, R.E. (1960). A new approach to linear filtering and prediction problems, J. Basic Engineering (Transactions of the ASME, Series B) 82, 35-45.

[98] Kendall, M.G. and Stuart, A. (1979). The Advanced Theory of Statistics, Vol.2, 4th Edition. Griffin, London.

[99] Kent, J.T. (1986). The underlying structure of nonnested hypothesis tests, Biometrika 73, 333-343.

[100] Khuri, A.I. and Cornell, J.A. (1987). Response Surfaces Designs and Analysis, Marcel Dekker, Inc., New York.

[101] Knox J.B. and Lange R. (1974). Surface air pollutant concentration frequency distributions: Implications for urban modelling, J. Air Pollut. Control Assoc. 24, 48-53.

[102] Kontnik, L.T. (1974). A multi-layer mixing cell model for the estimation of urban air pollution concentrations, Univ. North Carolina, Chapel Hill, M.S. Thesis.

[103] Lamb, R.G. (1978). A numerical simulation of dispersion from an elevated point source within a modelled convective planetary boundary layer, Atmos. Environ. 12, 1297-1304.

[104] Lamb, R.G. (1979). The effects of release height on material dispersion in the convective planetary boundary layer. Proc. 4th Sym. on Turbulence, Diffusion and Air Pollut., Am. Met. Soc., 27-33.

[105] Lamb, R.G. (1982). Diffusion in the convective layer, In Atmospheric Turbulence and Air Pollution Modelling, F.T.M. Nieawstadt and H.V. Dop (eds), Reidel D. Publishing Co., 159 230.

[106] Lamb, R.G. and Neiburger, M. (1971). An Interim Version of a Generalised Urban Air Pollution Model, Atmos. Environ. 5, 239-264.

[107] Lamb R.G. and Seinfeld J.H. (1973). Mathematical modelling of urban air pollution: General theory. Environ. Sci. Tech. 7, 253,261.

[108] Larsen R.I. (1969). A new mathematical model of air pollutant concentration averaging time and frequency, J. Air Pollut. Control Assoc. 19, 24-30.

[109] Larsen, R.I. (1971). A Mathematical Model for Relating Air Quality Measurements to Air Quality Standards, U.S. Environ. Prot. Agency, Publ. No. AP-89, pp. 56.

[110] Larsen R.I. (1973). An air quality data analysis system for interrelating effects, standards and needed source reductions, J. Air Pollut. Control Assoc. 23, 933-940.

[111] Larsen, R.I. (1974). An air quality data analysis system for interrelating effects, standards, and needed source reductions – part 2, J. Air Pollut. Control Assoc. 24, 551-558.

[112] Leahy, D.M. (1975). An application of simple advective pollution model to the city of Edmonton, Atmos. Environ. 9, 817-823.

[113] Leithe, W. (1972). The Analysis of Air Pollutants, Ann Arbor Science Publishers, Inc. (translated by Kondor, R.)

[114] Ley, A.J. (1982). A random walk simulation of two-dimensional turbulent diffusion in the neutral surface layer, Atmos. Eviron. 16, 2799-2808.

[115] Lynn, D.A. (1972). Fitting curves to urban suspended particulate data, Proc. Symp. Stat. Aspects Air Quality Data, Chapel Hill, N.C., November 1977, U.S. Environ. Protect. Agency, Publ. No. EPA-650/4-74-038, pp. 13-1-13-28.

[116] MacNicol, B. (1982). An estimation of the variation in mixing depth over the Canberra region during the period November 1974 to October 1975, Canberra College of Advanced Education, Canberra.

[117] Mage, T.D. (1975). An improved statistical model for analysing air pollution concentration data, Proc. 68th Ann. Meet. Air Pollut. Control Assoc. Boston, Mass., June 1975, Paper No. 57-51.4, pp. 28.

[118] Mann, N.R., Schafer, R.E. and Singpurwalla, N.D. (1974). Methods for Statistical Analysis of Reliability and Life Data, John Wiley, New York.

[119] McAleer, M. and Pesaran, M.H. (1986). Statistical inference in non-nested econometric models, App. Math. Comp. 20, 271-311.

[120] McElroy, J.L. and F. Pooler (1968). St Louis Dispersion Study, Vol. 2: Analysis, Report AP-53, Dept. of Health, Education and Welfare, Washington DC.

[121] Meisel, W.S. (1976). Empirical techniques for analysing air quality and meteorological data, part I, the role of empirical methods in air quality and meteorological analysis, U.S. Environ. Prot. Agency, Publ. No. EPA-600/4-76-029a, pp. 66.

[122] Meisel, W.S. and Teener, M.D. (1976). Empirical techniques, part II, feasibility study of a source-oriented empirical air quality simulation model, U.S. Environ. Prot. Agency, Publ. No. EPA-600/4-029b, pp. 61.

[123] Miles, G.H., Jakeman, A.J. and Bai, J., (1989). A method for predicting the future extremes of urban air pollution from vehicle emissions, meteorology and historical concentrations, Proc. 8th Biennial Conference, Simulation Society of Australia, Canberra, September 1989, 418-423.

[124] Miles, G. and Simpson, R.W. (1988). The general applicability of a model to control violations of short term exposure standards to TSP, Atmos. Environ. 22, 775-781.

[125] Myers, R.H. (1971). Response Surface Methodology, Allyn and Bacon, Inc., Boston.

[126] Newill, V.A. (1976). Air quality Standards, In Air Pollution, Vol. V, A.C. Stern (ed.), Academic Press, London, 445-504.

[127] Nicholson, S. (1975). A pollution model for street level, Atmos. Environ. 9, 19-31.

[128] Ott W.R., Mage D.T. and Randecker V.W. (1979). Testing the Validity of the Lognormal Probability Model: Computer Analysis of Carbon Monoxide Data from U.S. Cities, U.S. Environ. Prot. Agency, Publ. No. EPA-600/4-79-040.

[129] Pasquill, F.C. (1974). Atmospheric Diffusion, 2nd edition. John Wiley, New York, pp. 429.

[130] Pasquill, F.C. and Smith, F.B., (1983). Atmospheric Diffusion, 3 edition. Ellis Horwood, Chichester.

[131] Pearson, E.S. (1959). Note on an approximation to the distribution of non-central χ^2, Biometrika 46, 364.

[132] Pearson, K. (1900). On a criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling, Phil. Magazine, 5th Series 50, 157-175.

[133] Peterson T.W. and Moyers J.L. (1980). Emmission limits for variable sources by use of multipoint rollback, Atmos. Environ. 14, 1439-1444.

[134] Prinz, B., and Stratman, H. (1966). The statistics of propagation conditions in the light of continuous concentration measurements of gaseous pollutants, Staub 26, 4-12.

[135] Ragland, K.W. (1973). Multiple Box Model for Dispersion of Air Pollution From Area Sources, Atmos. Environ. 7, 1017-1032.

[136] Randerson, D. (1968). A Numerical Model for Predicting the Transport of Sulfur Dioxide Through the Atmosphere, Ph.D. Dissertation, Texas A & M University, pp. 153.

[137] Reid, J.D., (1979). Markov chain simulations of vertical dispersion in the neutral surface layer for surface and elevated releases, Boundary-layer Meteorology 16, 3-22.

[138] Reiquam, H.E. (1970). An atmospheric transport and accumulation model for airsheds. Atmos. Environ. 4, 233-237.

[139] Reiquam, H.E. (1971). Preliminary trial of a box model in the Oslo airshed, Proc. 2nd Int. Clean Air Congr., Washington, D.C., December 1970, H.M. England and W.T. (eds), Academic Press, New York, 1131-1135.

[140] Robinson, E. and Robbins, R.C. (1972). Air Pollution Control, Part II, W. Strauss (ed.), Wiley (Interscience), New York, 1-93.

[141] Schuck, E.A. and Papetti, R. (1973). Examination of the photochemical air pollution problem in sourthern California, U.S. Environ. Prot. Agency, Tech. Support Doc. Metropolitan Los Angeles Intrastate Air Quality Control Region Transportation Control Plan Final Promulgation.

[142] Schwarz, G. (1978). Estimating the dimension of a model. Ann. Statist. 6, 461-464.

[143] Scrase, F.J. (1930). Some characteristics of eddy motion in the atmosphere, Meteorological Office Geophysical Memoirs No. 52.

[144] Seinfeld J.H. (1975). Air pollution: physical and chemical fundamentals. McGraw Hill, New York.

[145] Shiozawa, K., Ootaki, A. and Okatomo, S. (1973). A multiple-box model for estimating air pollution in urban areas, Bull. Sci. Eng. Res. Lab., Waseda Univ., 61, 40-47 (in Japanese).

[146] Simpson, R.W. (1989). The assimilative capacity of areas affected by energy-related industrial activity: the Hunter region, In Air Pollutant from Energy Industries:Scientific Basis of Standards and Reserach Needs, A.J. Jakeman (ed.), Centre for Resource and Environmental Studies, Australian National University, Canberra.

[147] Simpson R.W., Daly N.J. and Jakeman A.J. (1983). The prediction of maximum air pollution concentrations for inert gases using Larsen's model and the ATDL model, Atmos. Environ. 17, 2497-2503.

[148] Simpson R.W. and Hanna R.S. (1981). A review of deterministic urban air quality models for inert gases, U.S. National Office of Air Assessment, Technical Memo ERL ARL-106, Maryland.

[149] Simpson R.W. and Jakeman A.J. (1985). Forecasting worst case pollution scenarios for acid gas and suspended particulates due to urban industrial development, Environmental Pollution (Series B) 9, 137-149.

[150] Simpson R.W., Jakeman A.J. and Daly N.J. (1985). The relationship between the ATDL model and the statistical distributions of wind speed and pollution data, Atmos. Environ. 19, 75-82.

[151] Singer, I.A. and M.E. Smith (1966). Atmospheric Dispersion at Brookhaven National Laboratory, Int. J. Air Water Pollut. 10, 125-135.

[152] Smith, F.B. (1976). Three Simple Models for Predicting Pollutant Concentrations in Cities, Meteorological Office Note TDN 74, Bracknell, U.K.

[153] Smith, M.E. (ed.) (1968). Recommended Guide for the Prediction of the Dispersion of Airborne Effluents, Am. Soc. Mech. Eng., New York.

[154] Stedinger, J.R. (1980). Fitting lognormal distributions to hydrologic data, Water Resour. Res. 16, 481-490.

[155] Steen, P.J. and Stickler, D.J. (1976). A Sewage Pollution Study of Beaches from Cardiff to Ogmore, Report January 1976, Cardiff: Department of Applied Biology, UWIST.

[156] Stoica, P., Eykhoff, P., Janssen, P. and Soderstrom, T. (1986). Model-structure selection by cross-validation, Int. J. Control 43, 1841-1878.

[157] Sutton, O.G. (1932). A theory of eddy diffusion in the atmosphere, Proc. R. Soc., A, 135, 143.

[158] Szepesi, D.J. (1977). Modified rollback model for air quality planning, Proc. 4th Int. Clean Air Congr., Tokyo, May 1977, 284-287.

[159] Takamatsu, T., Naito, M., Hiraoka, M., Kawata, K. and Mizoguchi (1971). Computer control system for preventing air pollution, Proc. 2nd Int. Clean Air Congr., Washington, D.C., 11 December 1970, H.M. Englund and W.T. Beery (eds), Academic Press, New York, 1135-1143.

[160] Taylor, G.I. (1915). Eddy motion in the atmosphere, Phil. Trans. R. Soc., A, 215.

[161] Taylor, J.A. (1985). Air quality models for environmental management. PhD Thesis, Australian National University.

[162] Taylor, J.A., Glikson, M., Polach, H. and Simpson, R. (1987). Total suspended particulate pollution in the Australian Capital Territory, Report to the Office of the Australian Capital Territory.

[163] Taylor, J.A., Jakeman, A.J. and Bai, J. (1986). A Monte Carlo study of estimation of the upper percentiles of the three-parameter gamma distribution using methods of moments and maximum likelihood. CRES Working Paper, No 1986/34, Australian National University, pp. 16.

[164] Taylor, J.A., Jakeman, A.J. and Bai, J. (1988). Modelling for air quality management, 12th World IMACS Congr. Sci. Comp., 18-22.

[165] Taylor, J.A., Jakeman, A.J. and Simpson, R.W. (1986). Modelling distributions of air pollutant concentrations-I, identificantion of statistical models, Atmos. Environ. 20, 1781-1789.

[166] Taylor, J.A., Simpson, R.W. and Jakeman, A.J. (1985). A hybrid model for predicting the distribution of pollutants dispersed from line sources. Science of the Total Environment 46, 191-213.

[167] Taylor, M.A.P. and Anderson, M. (1982). Modelling pollution and energy use in urban road networks, In. Proceedings 11th Australian Road Research Board Conference. 11(6): 1- 17.

[168] Taylor, M.A.P. and Anderson, M. (1984). Energy and emissions in a metropolitan road transport system, In Proceedings 12th Australian Road Research Board Conference. 12(6): 63- 78.

[169] Thomson, D.J. and Ley, A.J. (1982). A random walk dispersion model, applicable to diabatic conditions, Meteorological Office Note TDN 138, Bracknell, U.K.

[170] Tsukatani, T. and Shoyi, H. (1977). Statistical model of air pollutant concentration, Proc. 4th Clean Air Congr., Tokyo, May 1977, 315-317.

[171] Turner, D.B. (1964). A diffusion model for urban area, J. Appl. Met. 3, 83-91.

[172] Turner, D.B. (1969). Workbook of atmospheric dispersion estimates, US Public Health Service Publ. No. 999-AP-26, Revised.

[173] Turner, D.B. (1970). Workbook of Atmospheric Dispersion Estimates, Office of Air Programs, U.S. Environ. Prot. Agency, Publ. No. AP-26.

[174] Turner, D.B. (1979). Atmospheric Dispersion Modelling: A Critical Review, J. Air Pollut. Control Assoc. 29, 502-519.

[175] Urone, P. (1976). The primary air pollutants - a gaseous their occurrence, source, and effects, In Air Pollution, Vol. I, A.C. Stern(ed.), Academic Press, London, 23-75.

[176] Venkatram A. (1983). Uncertainty in predictions from air quality models, Boundary-Layer Meteorology 27, 185-196.

[177] Venkatram A. (1984). The uncertainty in estimating dispersion in the convective boundary layer, Atmos. Environ. 18, 307-310.

[178] Venkatram A. and Pleim J. (1985). Analysis of observations relevant to long-range transport and deposition of pollutants, Atmos. Environ. 19, 659-667.

[179] Weber, A.H. (1976). Atmospheric Dispersion Parameters in Gaussian Plume Modelling, Part 1. Review of Current Systems and Possible Future Developments, U.S. Environ. Prot. Agency, Publ. No. EPA-600/4-76-030a.

[180] Wells, C.H., and Lau, R.W.J. (1971). Stochastic modellijng and control of ambient air quality: a new approach, Proc. Inst. Electr. Eng. Western Electron. Show Conv., San Francisco, Calif., August 1971, Paper, No. 32/5, pp. 8.

[181] Whelpdale, D.M. and Munn, R.E. (1976). Global sources, sinks, and transport of air Pollution, In Air Pollution, Vol. I, A.C. Stern (ed.), Academic Press, London, 289-324.

[182] William S.M. and Harville D.A. (1981). Best linear recursive estimation for mixed linear models, J. Am. Statist. Assoc. 76, 860-869.

[183] Wilson, J.D., Turtell, G.D. and Kidd, G.E., (1981). Numerical simulation of particle trajectories in inhomogeneous turbulence, II: Systems with variable turbulent velocity scale, Boundary-layer meteorology 21, 423.

[184] Yeh, G.T. and Tsai, Y.J. (1976). Analytical solution of the three-dimensional diffusion equation with variable coefficients, Proc. 3rd Symp. Atmos. Turbulence, Diffusion Air Quality, Raleigh, N.C., 19 to 22 October 1976, Am. Met. Soc., Boston, Mass., 194-199.

[185] Young P.C. (1978). A general theory of modelling for badly defined systems, In Modelling, identification and control in environmental systems, G.C. Vansteenkiste (ed.), North-Holland, Amsterdam.

[186] Young, P.C. (1984). Recursive Estimation and Time Series Analysis: An Introduction, Springer-Verlag, Berlin.

[187] Zimmerman, J.R. and Thompson, R.S., (1975). User's guide for HIWAY, a highway air pollution model, U.S. Environ. Prot. Agency, Publ. No. EPA-650/4-74-008.