UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

A CORROSION SEVERITY RANKING METHODOLOGY AND A PREDICTIVE MODEL FOR CORROSION GROWTH BASED ON ENVIRONMENTAL AND CORROSION GROWTH DATA

A Dissertation

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

degree of

Doctor of Philosophy

By

CHARNNARONG SAIKAEW
Norman, Oklahoma
2003

Downloaded from SHAREOK. Persistent Handle URL: https://hdl.handle.net/11244/624

UMI Number: 3094295

UMI®

UMI Microform 3094295

Copyright 2003 by ProQuest Information and Learning Company.

All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

© Copyright by CHARNNARONG SAIKAEW 2003
All Rights Reserved.

A CORROSION SEVERITY RANKING METHODOLOGY AND A PREDICTIVE MODEL FOR CORROSION GROWTH BASED ON ENVIRONMENTAL AND CORROSION GROWTH DATA

A Dissertation APPROVED FOR THE
SCHOOL OF INDUSTRIAL ENGINEERING

BY

Dr. Mary C. Court (chair)

Dr. Fakhrildeen Albahadily

Dr. Teri Reed Rhoads

Dr. Randa L. Shehab

Dr. Krishnaiya Thulasiraman

To my parents

นาย สุวัฒน์ สายแก้ว
นาง เพียง สายแก้ว

Acknowledgements

I would like to thank my parents for supporting me throughout the completion of my study. I also want to thank my brother and sisters who have taken care of our parents during my time away from home. I am indebted to S.N. Goenka who enlightened me on meditation. Special thank should go to the Royal Thai Embassy Office of Education Affairs, which takes care of Thai scholars studying abroad.

I would like to express my deep appreciation to my advisor, Dr. Mary C. Court, for her supervision and guidance, and for providing me with supporting financial assistantship and a powerful computer. I would also like to thank Drs. Randa L. Shehab, Teri Reed Rhoads, K. Thulasiraman, and F.A. Albahadily for reading the final manuscript and serving on the committee. I am grateful to Dr. Lance L. Lobban who gave me an explanation of the atmospheric condition phenomenon. I am also grateful to Arinc, Inc. of Oklahoma City for providing me the atmospheric condition data sets and the corrosion growth for my research.

During the preparation of this dissertation, I have benefited from the help offered by many friends. Among these friends, I am thankful to Khong Farn Yu, Budi Santosa, Anurat Wisitsora-at, Amitava Majumdar, and Tinnakorn Komsan. Special friends including Kulawadee L. Pigott, Nopawan Rattasuk, Pherapha Jaidee, Thirawudh Pongprayoon, Nisarat and Eakasit Vorakitolarn, Ayodeji Farjebe, and Joseph Zume have given support when frustrations and discouragements appeared.

Finally, I would like to express appreciation to the office staff of the School of Industrial Engineering for providing equipment and the library staff for helping me locate research materials.

TABLE OF CONTENTS

LIST OF TABLES viii

LIST OF FIGURES x

ABSTRACT xii

Chapter

1. INTRODUCTION 1

1.1 Problem Statement and Research Objectives 8
1.1.1 Research objectives 9
1.2 Research Methodology Outline 10
1.3 Research Organization 12

2. LITERATURE REVIEW 13

2.1 Atmospheric Corrosion 13
2.2 Data Screening in Time-Series Data 19
2.2.1 Outliers 21
2.2.2 Missing observations 22
2.3 Preliminary Analysis 27
2.4 Principal Component Analysis 33
2.4.1 PCA Theory 33
2.4.2 Applications of principal component analysis 41
2.4.3 The data used in principal component analysis 46
2.5 Existing Corrosion Modeling Prediction Models 49
2.6 Growth Models 52
2.6.1 Gompertz growth model 52
2.6.2 Logistic growth model 55
2.6.3 Confined exponential growth model 56

3. METHODOLOGY 58

3.1 Data Screening Analyses 58
3.1.1 Data quality check and outlier analysis 58
3.1.2 Missing observation analysis: Neural Network Analysis 60
3.2 Dew Point Temperature 64
3.3 Correlation Analysis 65
3.4 Corrosion Severity Ranking Using Principal Component Analysis 76
3.5 Corrosion Growth Modeling Development 79
3.5.1 The GL model 81
3.5.2 The GC model 82
3.5.3 The CL model 84
3.6 Statistical Techniques for Modeling 86
3.6.1 Theory 86
3.6.2 Lack-of-fit test for assessing the fit of the model 91
3.6.3 Plotting residuals 92
3.6.4 Weighted least squares 93

4. CORROSION SEVERITY RANKING ANALYSIS 96

4.1 Data Screening Analyses 96
4.1.1 Data quality check and outlier analysis 96
4.1.2 Missing observation analysis 99
4.2 Corrosion Severity Ranking 102

5. CORROSION GROWTH ANALYSES 112

5.1 Predictive Corrosion Growth Models 112
5.2 Discussion 124

6. CONCLUSIONS AND FUTURE RESEARCH 129

6.1 Conclusions 129
6.2 Future Research 131

REFERENCES 134

APPENDIX A 140

APPENDIX B 159

LIST OF TABLES

Table		Page
1.1	Description of raw data as collected by Arinc, Inc.	4
3.1	Correlation matrix for atmospheric conditions at Hickam AFB	74
3.2	Correlation matrix for atmospheric conditions at Kadena AB	74
3.3	Correlation matrix for atmospheric conditions at Macdill AFB	74
3.4	Correlation matrix for atmospheric conditions at RAF Mildenhall	75
3.5	Correlation matrix for atmospheric conditions at Pease ANGB	75
3.6	Correlation matrix for atmospheric conditions at Seymour Johnson AFB	75
3.7	Atmospheric conditions conducive to corrosion growth	76
4.1	Summary of data quality check	97
4.2	Summary of outlier analysis	97
4.3	Summary of missing observation analysis	101
4.4	Numbers of 30-minute intervals when an atmospheric condition is conducive to corrosion growth for each atmospheric data condition	103
4.5	Compositional data set	103
4.6	Covariance matrix	104
4.7	Eigenvalues of the covariance matrix	104
4.8	Eigenvectors	105
4.9	The first two principal components	105
4.10	Test site locations of rack exposure with climate types and distance from the sea (culled from Howard et al., 1999)	111
4.11	Corrosion severity rankings by locations of the three scenarios (number 1 is the most severity)	111
5.1	Confined exponential model fitting to corrosion growth data set of Hickam AFB using SAS®	116
5.2	Lack-of-fit test for Confined Exponential model for Hickam AFB data with weighted least squares	119

5.3 Model parameter estimates for Confined Exponential model for Hickam AFB data with weighted least squares 119

5.4 Lack-of-fit test for Power Law model for Hickam AFB data with weighted least squares 120

5.5 Model parameter estimates for Power Law model for Hickam AFB data with weighted least squares 120

5.6 Lack-of-fit test for GL model for Hickam AFB data with weighted least squares 121

5.7 Model parameter estimates for GL model for Hickam AFB data with weighted least squares 121

5.8 Lack-of-fit test for GC model for Hickam AFB data with weighted least squares 122

5.9 Model parameter estimates for GC model for Hickam AFB data with weighted least squares 122

5.10 Lack-of-fit test for CL model for Hickam AFB data with weighted least squares 123

5.11 Model parameter estimates for CL model for Hickam AFB data with weighted least squares 123

5.12 Summary of predictive corrosion growth models for the four bases 126

LIST OF FIGURES

Figure		Page
1.1	Aircraft lap joint with damaging moisture and corrosion	2
1.2	A C/KC-135 refueling a Fighter	3
1.3	Overall methodology tasks	11
2.1	The system of atmospheric corrosion	14
2.2	Cause-and-effect diagram influencing corrosion growth	15
2.3	A neural network architecture	26
3.1	Overall methodology for obtaining corrosion severity ranking	59
3.2	A flow chart for detecting outliers	60
3.3	A matrix of scatter plots of Hickam AFB	68
3.4	A matrix of scatter plots of Kadena AB	69
3.5	A matrix of scatter plots of Macdill AFB	70
3.6	A matrix of scatter plots of RAF Mildenhall	71
3.7	A matrix of scatter plots of Pease ANGB	72
3.8	A matrix of scatter plots of Seymour Johnson	73
3.9	Procedure for transforming the original data set into a compositional data set	78
4.1	A snap-shot of 30-minute interval recordings of air temperature at Hickam AFB	98
4.2	Neural network experiment screen	100
4.3	Data points predicted from the method of neural network	101
4.4	Scree plot	104
4.5	Scatter plot of the first two principal components for the six air bases	106
5.1	Corrosion growth data sets obtained from the six operational air force bases	113
5.2	Residual analysis from the results of the five models using OLS for Hickam AFB	117

5.3 Model adequacy checking for Confined Exponential model for Hickam AFB data with weighted least squares 119

5.4 Model adequacy checking for Power Law model for Hickam AFB data with weighted least squares 120

5.5 Model adequacy checking for GL model for Hickam AFB data with weighted least squares 121

5.6 Model adequacy checking for GC model for Hickam AFB data with weighted least squares 122

5.7 Model adequacy checking for CL model for Hickam AFB data with weighted least squares 123

5.8 Corrosion growth predictive models for Hickam AFB 127

5.9 Corrosion growth predictive models for Kadena AB 127

5.10 Corrosion growth predictive models for RAF Mildenhall 128

5.11 Corrosion growth predictive models for Seymour Johnson AFB 128

ABSTRACT

This dissertation presents a new methodology for defining corrosion severity ranking by location for six operational air force bases, Hickam Air Force Base (AFB), Kadena Air Base (AB), Macdill AFB, Royal Air Force (RAF) Mildenhall, Pease Air National Guard Base (ANGB), and Seymour Johnson AFB. Three new corrosion growth predictive models are also presented so that a foundation for establishing a corrosion maintenance and inspection schedule of the C/KC-135 aircraft can be developed. The corrosion severity ranking scheme and the predictive growth models for the six operational air force bases will allow the United States Air Force (USAF) to concentrate their efforts on proactively inspecting aircraft for corrosion when deployed and operated at bases deemed as highly severe corrosion sites.

The method of principal component analysis (PCA) is used for the first time to analyze compositional data sets of atmospheric conditions (or thresholds) for defining corrosion severity ranking by locations (air force bases). The results show that the ranking for the six operational air bases from the most severe site to the least severe site is Hickam AFB, Kadena AB, Macdill AFB, Seymour Johnson AFB, RAF Mildenhall, and Pease ANGB.

In order to develop a more accurate corrosion growth predictive model, three corrosion growth predictive models are developed by modifying and combining the following existing growth models: the Gompertz growth and the logistic growth models (GL model), the Gompertz growth and the confined exponential growth models (GC model), and the logistic growth and the confined exponential growth models (CL model). The confined exponential growth model, the power law equation, and the three new

models (i.e., GL, GC, and CL models) are compared through lack-of-fit tests and model adequacy checking (after performing weighted least square analysis). Corrosion growth data sets from four operational air bases (Hickam AFB, Kadena AB, RAF Mildenhall, and Seymour Johnson AFB) are used to perform the statistical tests. The results showed that the CL model provides the best fit for all corrosion growth data sets of the four operational air bases and dominates the other models in terms of weighted mean square error. The CL model also reveals that Hickam AFB is the most severe corrosion site and supports the results of the PCA analysis on corrosion severity ranking. Although other corrosion growth models exists, this research represents the first models based on corrosion growth data of alloys obtained from operational C/KC-135 aircraft.

A CORROSION SEVERITY RANKING METHODOLOGY AND A PREDICTIVE MODEL FOR CORROSION GROWTH BASED ON ENVIRONMENTAL AND CORROSION GROWTH DATA

CHAPTER 1

INTRODUCTION

The goal of the United States Air Force (USAF) Aging Aircraft Program is to predict corrosion growth so that aircraft maintenance schedules can be developed to prevent aircraft structural failures due to corrosion. At issue is the ability to model and predict corrosion growth during the operational life of an aircraft.

In 1992, the Oklahoma City Air Logistic Center (OC-ALC) and Arinc, Inc. of Oklahoma City began a program of aging aircraft disassembly and hidden corrosion detection in order to investigate the influence of aging aircraft corrosion on the USAF C/KC-135 fleet. The C/KC-135 is an air refueling tanker and special purpose aircraft. Figure 1.1 illustrates the lap joint construction showing damaging moisture and corrosion around the rivet area. This type of corrosion is mostly found in aircraft lap joints constructed from aluminum alloy sheets (e.g., 2024-T3) with aluminum pop rivets. The lap joint is a region of particular concern to the USAF because corrosion is known to occur extensively in this area and is difficult to detect. To build a comprehensive program for analyzing and detecting corrosion for the C/KC-135 fleet, Arinc, Inc. had a three program attack: 1) evaluate and identify non-destructive inspection/testing (NDI/NDT) equipment for hidden corrosion detection and quantification of the C/KC-135 fuselage and wings, 2) invasively disassemble a complete C/KC-135, and 3) conduct aircraft structural corrosion data gathering.

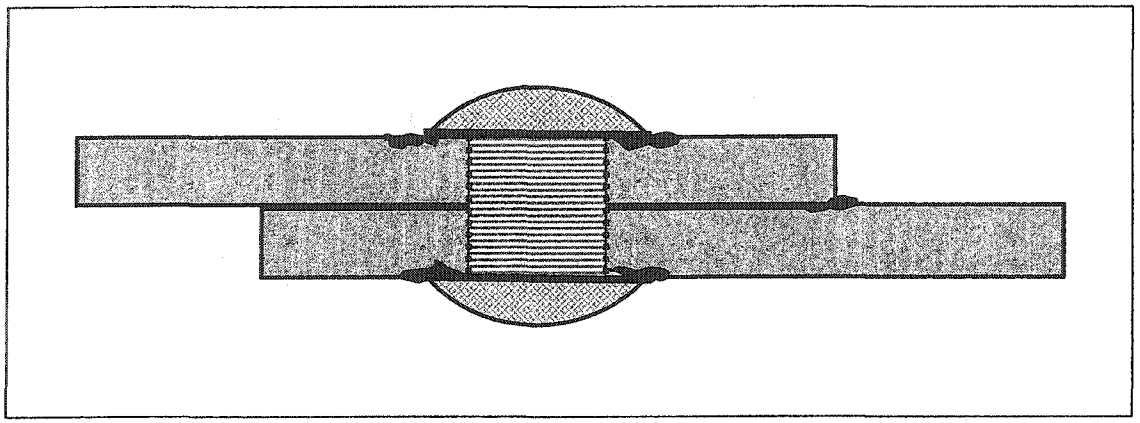


Figure 1.1: Aircraft lap joint with damaging moisture and corrosion

Note that the fleet of C/KC-135 has operated over 600 individual aircrafts around the world and was built in the mid-1950s, but is now expected to operate until the year 2040 (Groner and Nieser, 1996). They are currently the oldest aircraft in the USAF fleet (Ferrer and Kelly, 2002). Figure 1.2 shows the C/KC-135 aircraft. The primary mission of this aircraft is to aerial refuel compatible USAF, Navy, Marine Corps, and US-allied aircraft. The C/KC-135 is equipped with a flying boom for fuel transfer. A special drogue can be attached to the flying boom on the ground so that it can refuel probe-equipped aircraft. The flying boom is controlled by an operator stationed in the rear of the airplane. In addition, this aircraft can hold passengers and cargo with a deck above the fuselage-mounted tanks.

The C/KC-135 fuselage construction employed spot-welded doublers and lap joints, which promote hidden corrosion (Groner and Nieser, 1996). In addition to the fuselage lap joint skins, the wing skin fastener area is also considered as a source of corrosion in aircraft, which is not easy to detect. The Arinc, Inc. program, thus, dealt with detecting hidden corrosion and evaluating the ability of NDI/NDT equipment for detecting hidden corrosion in lap joints and wing skins. After evaluation and inspection

by NDI/NDT techniques, lap joint inspection and wing skin fastener areas were cut from a retired C/KC-135 and subjected to disassembly in an effort to find and quantify the hidden or inaccessible corrosion. The corrosion detection results from the NDI/NDT techniques were compared to the actual corrosion in order to evaluate the NDI/NDT techniques. Note that NDI/NDT techniques consist of eddy current, ultrasonic, acoustic emission, thermal imaging, shearography, and enhanced visual inspection (more details of these techniques can be found in Hagemaier et al.,1985).

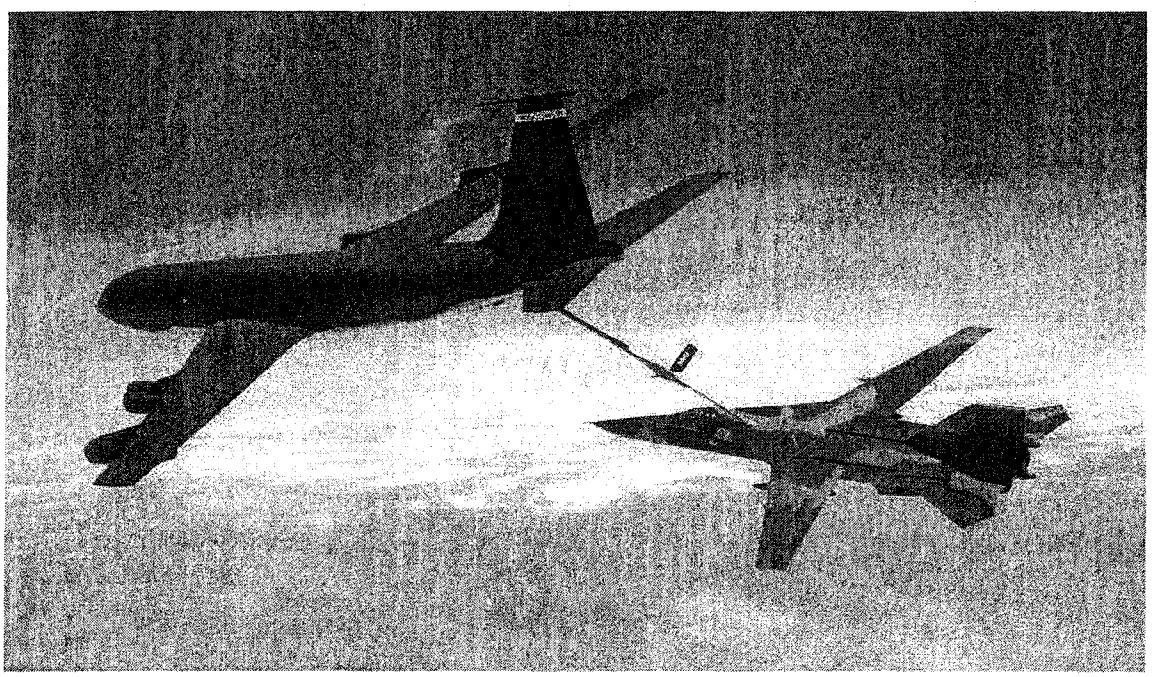


Figure 1.2: a C/KC-135 refueling a Fighter
Source: "KC-135 Stratotanker" [online].
Available: http://www.fas.org/nuke/guide/usa/bomber/kc-135.htm

However, the program of aging aircraft disassembly and hidden corrosion detection at Arinc, Inc. in 1992, did not include consideration of what atmospheric conditions can lead to hidden corrosion or how corrosion grows over the time. In 1996, the OC-ALC, in conjunction with Arinc, Inc., began collecting atmospheric condition data and corrosion growth data from exposure racks of the C/KC-135 faying surfaces specimens (lap joints and wing skin fastener area coupons) (Howard et al., 1999). A description of the atmospheric data as collected by Arinc, Inc. is presented in Table 1.1.

Table 1.1: *Description of raw data as collected by Arinc, Inc.*

Data	Description	Acronym	Unit of measurement
Relative humidity	The amount of water vapor in the air compared with the amount of vapor needed to make the air saturated at the air's current temperature (range 0-100%RH)	RH	%RH
Air temperature	A measure of the warmth or coldness of the air captured at each operational air base	AT	°F
Rain pH	The acidity and sulphate content of rainfall (range 0.0-14.0)	pH	—
Rainfall	Water that provides moisture on the metal surfaces	RF	inch
Time-of-wetness 1	Length of time that moisture is present on the metal surface. The TOW1 sensor is used to detect light dew (range 0-1800 seconds)	TOW1	second
Time-of-wetness 2	Length of time that moisture is present on the metal surface. The TOW2 sensor is used to detect rain and heavier liquid condensation. (range 0-1800 seconds)	TOW2	second
Surface temperature	A measure of the warmth or coldness of the metal surface captured at each operational air base	ST	°F

Note that a coupon is a corrosion-monitoring specimen of material exposed to the environment on a rack for a given duration and removed for analysis. Sixty coupons were installed on an exposure rack at each of six operational air force bases around the world. The six operational air force bases chosen for the experiment included Hickam Air Force Base (AFB) in Hawaii, Kadena Air Base (AB) in Japan, Macdill AFB in Florida, Royal Air Force (RAF) Mildenhall in England, Pease Air National Guard Base (ANGB) in New Hampshire, and Seymour Johnson AFB in North Carolina. These bases were chosen to represent the range of atmospheric conditions C/KC-135 aircraft could be exposed to over their operational life. The sixty coupons on fabricated racks at each location consist of forty-five lap joint specimens and fifteen wing skin specimens. The materials used for the forty-five lap joint specimens were new 2024-T3 Alclad[1] aluminum, used 2024-T3 Alclad aluminum, and used 2024-T3 with the Alclad removed. Each material was used for fifteen of the lap joint specimens. The material of the fifteen wing skin specimens was 7178-T6 aluminum upper wing skin. Note that these aluminum alloys were actual aircraft construction material. All the coupons were nominal 0.04 inch thick (1.016 millimeter). Thus, two hundred and seventy lap joint specimens and ninety wing skin specimens were installed by Arinc, Inc. and were being monitored at the six rack locations while forty-two specimens randomly sampled were returned yearly for inspection of corrosion growth. The specimens were investigated at the test sites and selected specimens were returned periodically for data collection in order to determine the corrosion growth.

---

[1] Alclad is a pure aluminum coating that is highly resistant to corrosive attack while clad is a high strength aluminum alloy sheet coated.

Atmospheric conditions are clearly important because environment has long been identified as the root-cause of corrosion problems (Feinberg et al., 1994). The atmospheric condition data sets from the six operational air bases had been collected at 30-minute intervals between the years 1996 (as the initial year of exposure) and 1999. The atmospheric condition data captured include air temperature, relative humidity (RH), rain pH, rainfall, time-of-wetness (TOW), and surface temperature. Note that the moisture present on the surface of a metal was measured from two moisture sensors. That is, TOW can be separated into TOW1 and TOW2. The TOW1 sensor detects light dew whereas the TOW2 sensor detects rain and heavier liquid condensation.

The atmospheric condition data collections from the six air force bases were monitored by a Solus computer system (Solus Systems, 1994). The reason for using the Solus system in this program is that this system requires low current (148 mA) while it has versatility in recording data. This system involves a general purpose computer that can monitor conditions through sensors, detect events through inputs, and log data acquired through the sensors and the control devices. All sensors in the Solus computer system operate using wet cell battery power maintained by solar panels (since all rack sites are located in remote site areas in which line power is not available). The Solus computer is connected to the host through modems and the telecommunications system (e.g., telephone lines, satellites) with a modified serial connection cable. The main terminal box provides terminal connections for the environmental monitoring sensors and routing circuitry to the Solus computer system. Air temperature, surface temperature, relative humidity, rainfall, rain pH, and time-of-wetness (TOW1 and TOW2), are also monitored through sensors. However, Arinc, Inc. experienced several data collection and

transmittal downtimes during the course of this study which led to errors and gaps in atmospheric data collection.

Corrosion growth data sets collected from the six air bases have been collected annually between the year 1996 (as the initial year of exposure) and the year 1999 and the thickness loss of material has been measured from each coupon after each one-year exposure. It is important to note that the material thickness loss is currently considered the most significant corrosion characteristic that can be measured and used to evaluate corrosion.

After one year of the program's study, Arinc, Inc. used visual examinations to quantify corrosion areas of the lap joints and wing skins, and used eddy current inspection to evaluate corrosion severity by location. Four levels of corrosion (i.e., uncorroded, light corroded, corroded, and destructively corroded) were used to subjectively categorize corrosion severity (Howard et al., 1999).

Besides the four levels of corrosion used to subjectively categorize corrosion severity by location, Arinc, Inc. used quantification of corrosion areas from lap joint specimens. The quantity of corrosion areas was attributable to percent metal loss (Howard et al., 1999). Arinc, Inc. used eddy current inspection to detect the percent metal loss. However, Howard et al. stated that eddy current inspection could not always accurately quantify corrosion in percent metal loss.

Arinc, Inc. then generated a composite evaluation of the six bases' corrosion using visual (subjective) and quantified corrosion (percent metal loss) at each of the six locations. With the most corrosive location given the lowest number, the six locations were ranked by category from one to six. The objective of the Arinc, Inc. study was to provide the USAF with corrosion severity analysis at the locations.

The results of corrosion severity ranking developed by Arinc, Inc. showed that the most severe locations were Hickam AFB, Kadena AB, and RAF Mildenhall followed by Pease ANGB, Macdill AFB, and Seymour Johnson AFB. However, the result of their study was based on subjective visual evaluation and a questionable methodology for estimating percent metal loss. In addition, their evaluation was based on data for only one year. Moreover, their evaluation did not include atmospheric conditions at each base that could be used to provide data for a more comprehensive corrosion severity ranking scheme.

1.1 Problem Statement and Research Objectives

According to the first year (1996) results of the study by Arinc, Inc., the following research issues were identified:

1. The analysis of corrosion severity ranking performed by Arinc, Inc. in 1996 was too subjective in that it lacked a valid quantifiable methodology.
2. Missing data is a considerable concern that must be addressed. Arinc, Inc. experienced equipment failure and data acquisition (download) interruptions. As a result, some atmospheric data were lost.
3. The atmospheric data dependency must be addressed. For example, a large RH level can cause moisture to be present and therefore should trigger a TOW reading. Consequently, data must be examined jointly to identify conditions for corrosion growth.
4. The atmospheric data within each measurement is a time series and is highly correlated.

5. The corrosion growth data is collected on an annual basis where a set of coupons is removed from the exposure racks, each coupon's corrosion measurements are recorded and the coupon is not returned to the rack. Consequently, each year's accumulated corrosion results are not on the same set of coupons.
6. Analytical models have not been developed for predicting corrosion based on corrosion growth data sets.

1.1.1 Research objectives

Currently, the C/KC-135 fleet is at least 30 years old while the USAF wishes to continue its operational life to the year 2040. The USAF realizes that environmental conditions coupled with the nature of the airframe construction (such as joint fastening system and different material types) may lead to extensive corrosion growth and requires inspection and maintenance programs to detect corrosion. However, the USAF also recognizes that to reduce the amount of time and money spent repairing corrosion damage, corrosion should be prevented or minimized. Consequently, the research presented has the following objectives:

1. To provide a corrosion severity ranking scheme for the six operational air bases stemming from the atmospheric conditions surrounding the bases. This will allow the USAF to concentrate their efforts on proactively inspecting aircraft for corrosion when deployed and operated at highly severe corrosion sites.
2. To provide a predictive model of corrosion growth based on the corrosion data gathered from coupons collected at each base. This will allow the USAF to look at the rate of corrosion growth by site as a function of time.

At issue is the ability of this research to address the data dependency and data loss issues (as outlined in Section 1.1) and the mixed level multivariate atmospheric data (e.g., threshold values are discrete while temperature is continuous).

The research contribution of this work is that for the first time a methodology for defining corrosion severity rankings is provided that in the future can aid aircraft maintenance programs for prioritizing corrosion inspection and repairs by base. In addition, three corrosion growth predictive models are developed based on atmospheric exposure of operational aircraft alloys. Past models have been developed for other metals but not for the alloys used in C/KC-135 operational aircraft. That is, no corrosion model has been developed for C/KC-135 alloys that have been "aged" to represent actual operational wear and tear.

1.2 Research Methodology Outline

Figure 1.3 illustrates the overall methodology tasks followed in this research. The two types of data, the atmospheric conditions and the corrosion growth data, were collected by Arinc, Inc. (see further details in Chapter 2). Corrosion growth data were measured yearly in terms of corrosion thickness loss by coupon, while the atmospheric conditions were collected on a half-hour basis from the six air force bases.

To rank corrosion severity of the six air force bases, the method of principal component analysis (PCA) is used for simultaneously analyzing atmospheric conditions data sets of the bases. The time series data sets of the atmospheric conditions are translated to a compositional data set based on the percentage of 30-minute time intervals exceeding thresholds or meeting conditions conducive to corrosion growth. The compositional data set is then used to obtain a corrosion severity ranking of the bases.

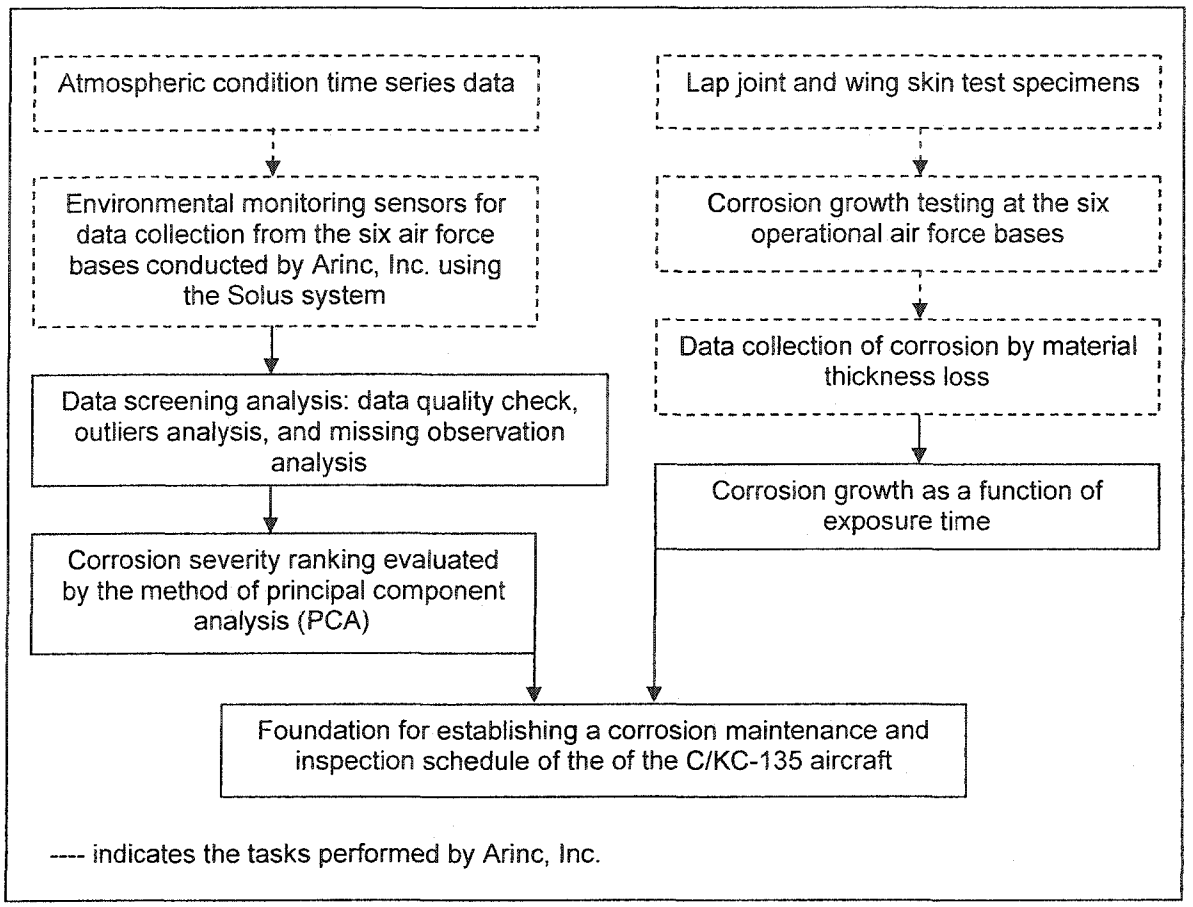


Figure 1.3: Overall methodology tasks

One of the purposes of this research is to develop a predictive model of corrosion growth using corrosion growth data sets. Based on corrosion growth phenomena, several predictive corrosion growth models (corrosion growth thickness loss as a function of time) are proposed for the first time by modifying existing growth models (i.e., the Gompertz growth model, the logistic growth model, and the confined exponential growth model). The newly proposed models are used to fit the data sets of corrosion growth for each air base. Then, a statistical comparison of the proposed models, an existing corrosion growth model (power law equation model), and an existing growth model

(confined exponential growth model) is presented, where the "best" model is identified in terms of statistical accuracy.

1.3 Research Organization

The research organization begins with Chapter 2, which provides a literature review on the background and theory of atmospheric corrosion, preliminary data analyses, PCA analysis and existing corrosion growth models. Chapter 3 gives a detailed procedure of the research methodologies used for performing corrosion severity ranking analysis and corrosion modeling analyses. Chapters 4 and 5 present the results of the corrosion severity ranking analysis and the corrosion growth modeling, respectively. Chapter 6 gives a summary of the results and outlines future research.

CHAPTER 2

LITERATURE REVIEW

This chapter presents a literature review that describes the theory, methodology, and applications used to support this research. The first section provides a theory of atmospheric corrosion for supporting corrosion severity ranking and predictive modeling analyses. The second section provides data screening techniques that are used for outlier analysis and missing observation analysis of the atmospheric condition data. The third section describes preliminary analysis of correlation among atmospheric condition factors. The fourth section describes detailed information on the method of principal component analysis which is used to perform corrosion severity ranking by location. The fifth section provides some corrosion modeling efforts as studied by prior researchers. The last section describes the theory of growth models, which is used in this research to derive new corrosion growth models.

2.1 Atmospheric Corrosion

Atmospheric corrosion is an electrochemical process involving a metal, corrosion products, a surface electrolyte, and the atmosphere (Kucera and Mattsson, 1987). Atmospheric corrosion is defined as the corrosion of materials exposed to air and its pollutants, rather than corrosion caused by immersion of the metal in a liquid (Roberge, 2000). Atmospheric corrosion develops under thin layers of adsorbed oxygen and water. An oxide layer is formed on the metal surface when the gaseous oxygen interacts with the

metal. The growth of this layer is determined by reactions at the metal-oxide interface and by the transfer of reacting particles through the oxide layer. During the ongoing corrosion process, a corrosion product (output) forms after long atmospheric exposure. Figure 2.1 illustrates the system of the atmospheric corrosion that consists of the input, the corrosion process, and the output. Atmospheric corrosion process, as related to an electrochemical reaction, depends on the time-of-wetness (TOW), the temperature, the humidity, the content and concentration of chemical impurities in the air (e.g., sulfur dioxide content, hydrogen sulfide content, chloride content), the amount of rainfall, dust, and even the position of the exposed metal (Schweitzer, 1998). Furthermore, it also depends on geographical locations, the distance from sea or salt sources, and nature of the metals. From this research, the cause-and-effect diagram in Figure 2.2 is developed and presented.

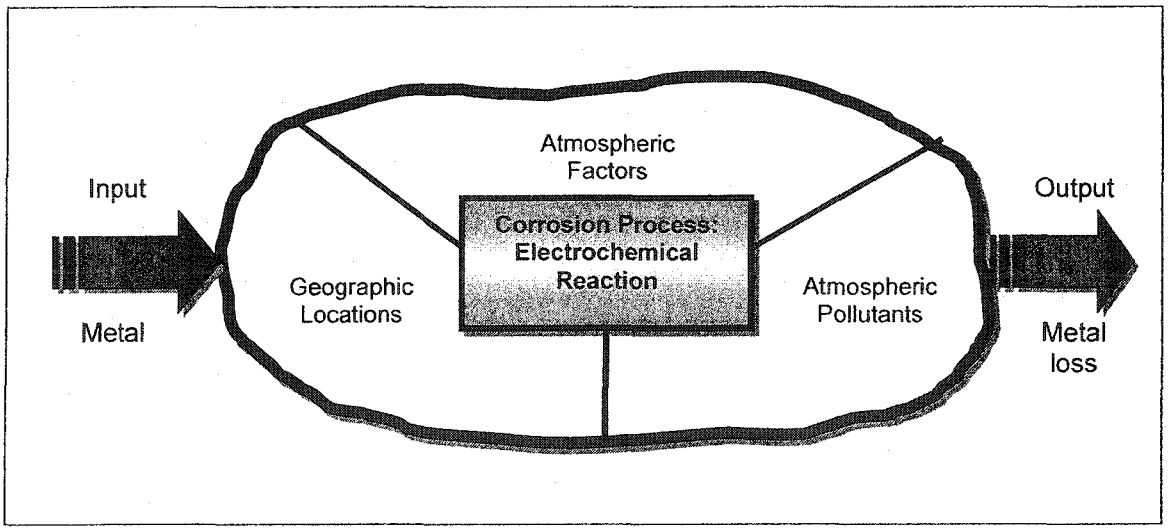


Figure 2.1: The system of atmospheric corrosion

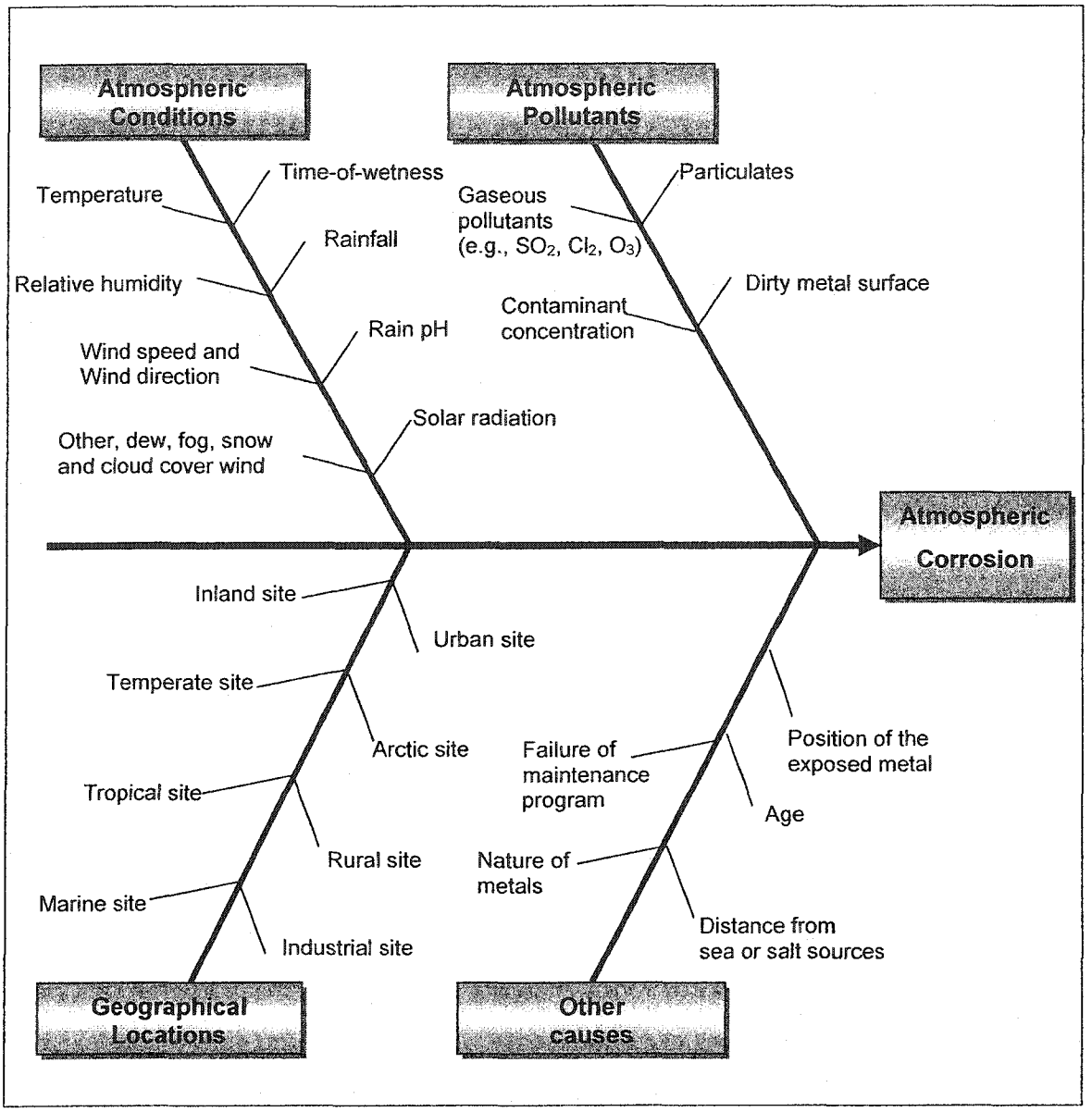


Figure 2.2: Cause-and-effect diagram influencing corrosion growth

Figure 2.2 illustrates the factors influencing atmospheric corrosion. The atmospheric conditions, the atmospheric pollutants, and the geographical locations as ascertained from literature have been identified as contributors to corrosion. Moreover, other contributors (e.g., nature of metal, distance from sea, maintenance schedule, and age of the aircraft) have also been identified to cause corrosion. The graph represents the research conducted to present a comprehensive depiction of the factors identified as promoting corrosion growth. The identification of the factors of Figure 2.2 is explained in the paragraphs that follow and the means to capture the data from the Arinc, Inc. experiment is also outlined.

Time-of-wetness (TOW), an important practical variable in atmospheric corrosion, is the duration of the electrochemical corrosion processes on the metal surface. TOW is strongly dependent on the critical relative humidity (RH), air temperature, surface (metal) temperature, duration/frequency of rain, fog, dew, wind speed, wind direction, and hours of sunshine (Lawson, 1995). In the Arinc, Inc. experiment, TOW is separated as TOW1 and TOW2 (Howard et al., 1999). A TOW1 sensor was used to detect light dew while the TOW2 sensor was used by Arinc, Inc., to detect rain and heavier liquid condensation. Hence, TOW can be obtained from the Arinc, Inc. data either by instruments that detect condensed moisture surfaces or by counting the number of hours at any specific time interval when the temperature is above 0°C and the relative humidity is greater than 80%RH (Dean, 1993; Roberge, 2000).

Temperature is another factor in atmospheric corrosion. For a constant humidity level, an increase in temperature leads to a high corrosion rate because it tends to stimulate corrosive attack by increasing the rate of electrochemical reactions and the diffusion processes (Roberge, 2000). Schweitzer (1991) stated that high temperatures

(60°F or more) increase the rate of corrosive attack on surface metals because corrosion reactions are thermally activated. Below freezing, corrosion does not occur due to poor electrolyte activity.

For the Arinc, Inc. experiment, the surface temperature is also an important factor in atmospheric corrosion since moisture formation on the aircraft skin surface can increase the rate of the corrosion reaction. It is important to note that the moisture that can condense from air and dew is formed when the ambient temperature is within ±4°F of the approximate dew point temperature (Howard et al., 1999). Thus, moisture may not condense if the surface temperature is significantly hotter than the ambient temperature. This implies that corrosion under condensing conditions (i.e., surface temperature is within ±4°F of the approximate dew point temperature) is a function of the rate of condensation and corrosion products from the metal surface.

Humidity, the percentage of water vapor in the air at a given temperature, is one of the most important factors affecting atmospheric corrosion. At high relative humidity (RH) level, the moisture film at the metal surface increases in thickness and the corrosion process becomes an electrochemical reaction. Corrosion growth occurs at high humidity levels (60%RH or more) due to the condensation of moisture in the oxide layers, or the vaporization of the corrosion products (Wallace et al., 1985). However, in addition to RH, the nature of metals and the presence of pollutants may influence moisture film formation.

Rain affects atmospheric corrosion by providing moisture on the metal surfaces. Light rain can be harmful since it is a source of moisture that resides on the surface, while heavy rain can be beneficial since it washes away and dissolves pollutant deposits from the surface. The acidity and sulphate content of rainfall also affects atmospheric

corrosion. The rain pH of either less than 4 or greater than 8.5 has been shown to be conducive to corrosion growth (Wallace et al., 1985). In conjunction with temperature and RH measurements, data obtained from a rainfall gauge is useful in determining the TOW for the Arinc, Inc.

Wind speed and wind direction affect the dispersion of air pollutants and the accumulation of particulates on the metal surfaces while solar radiation causes damage of protective coatings and contributes to exposing the underlying metal to corrosion (Roberge, 2000).

The content and concentration of chemical impurities in the air or atmospheric pollutants (e.g., sulfur dioxide [SO_2] and chloride [Cl_2]) are considered to be major contributors to atmospheric corrosion of metals. Sulfur dioxide is produced from the burning of fossil fuel (e.g., coal) for heating purposes, industrial activity, and thermal electric power generation. Chloride comes mainly from marine sources. Chloride exists in the atmosphere as particles or droplets that settle on the surfaces and provide the ionic constituents necessary to accelerate corrosion. The chloride deposition levels decrease considerably farther from the shoreline. Thus, distances from sea or salt sources and from pollutant sources are factors that accelerate atmospheric corrosion. The degree of corrosion growth is determined by the contaminant concentration, the length of time the pollutants remain on the surface, and the composition of contaminating materials.

Consequently, geological locations are also considered as factors affecting atmospheric corrosion. Industrial sites are usually highly corrosive locations and the corrosivity tends to be significantly dependent on concentrations of sulfur dioxide, chloride, phosphates, and nitrates. Marine sites are also considered as high corrosive locations. Marine sites extending some 4-5 km inland tend to have the most severe

corrosion due to the effect of windswept chlorides (Roberge, 2000). On the other hand, rural and urban sites are generally the least corrosive locations because these regions tend to have low amounts of chemical pollutants from industrial activity, domestic fuel emission, etc.

2.2 Data Screening in Time-Series Data

This section describes the theory and procedures for handling the problems of outliers and missing observations in time series data. The data collected during the Arinc, Inc. experiment can be analyzed as a time series if the conditions and assumptions of time series analysis are upheld. Fundamentals of the Box-Jenkins approach and well-known time series models, including autoregressive model (AR), moving average (MA), autoregressive moving average (ARMA), and autoregressive integrated moving average (ARIMA) are presented.

In the autoregressive model (AR), the current value of a time series is expressed as a finite, linear aggregate of previous values of the time series and an uncorrelated residual series (Box et al., 1994). This model can be defined as follows:

$$\tilde{Y}_t = \Phi_1 \tilde{Y}_{t-1} + \Phi_2 \tilde{Y}_{t-2} + \ldots + \Phi_p \tilde{Y}_{t-p} + a_t, \tag{2.1}$$

where $\tilde{Y}_t = Y_t - \mu$ is the deviate from the mean μ and a_t is the uncorrelated residual series (a zero-mean gaussian white noise process with variance σ^2) at time t. This model consists of an autoregressive process of order p (lagged dependent variable) and unknown parameters ($\mu, \Phi_1, \ldots, \Phi_p, \sigma_a^2$) that can be estimated from the data.

In the moving average (MA), the current value of the time series is expressed as a finite number of the uncorrelated residual series. This model is defined as follows:

$$\tilde{Y}_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \ldots - \theta_q a_{t-q}, \tag{2.2}$$

which is called a moving average process of order q. This model consists of unknown parameters ($\mu, \theta_1, \ldots, \theta_p, \sigma_a^2$) that can be estimated from the data.

The autoregressive moving average (ARMA) is a combination of the autoregressive and moving average models. This model is defined as follows:

$$\tilde{Y}_t = \Phi_1 \tilde{Y}_{t-1} + \Phi_2 \tilde{Y}_{t-2} + \ldots + \Phi_p \tilde{Y}_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}. \tag{2.3}$$

This model consists of an autoregressive process of order p, a moving average process of order q, and unknown parameters ($\mu, \Phi_1, \ldots, \Phi_p, \theta_1, \ldots, \theta_q, \sigma_a^2$) that can be estimated from the data.

The autoregressive integrated moving average (ARIMA) includes nonstationary behavior which does not vary about the fixed mean. Such nonstationary behavior can be represented by a model, which calls for the dth difference of the process to be stationary. This model can be defined as follows:

$$w_t = \Phi_1 w_{t-1} + \ldots + \Phi_p w_{t-p} + a_t - \theta_1 a_{t-1} - \ldots - \theta_q a_{t-q}. \tag{2.4}$$

A time series is integrated of order d if the time series becomes stationary after being first differenced d times. Note that w_t is defined by:

$$w_t = \nabla^d Y_t, \tag{2.5}$$

where ∇ is the backward difference operator, which is defined by the operator of backward shift, B. The operator ∇ can be written in terms of B as:

$$\nabla Y_t = Y_t - Y_{t-1} = (1 - B)Y_t.$$

For example, an ARIMA(1, 1, 1) process (corresponding to $p = 1$, $d = 1$, and $q = 1$) can be written in terms of ∇ and B as:

$$\nabla Y_t = \Phi_1 \nabla Y_{t-1} + a_t - \theta_1 a_{t-1}$$

or

$$(1 - \Phi_1 B)\nabla Y_t = (1 - \theta_1 B)a_t.$$

2.2.1 Outliers

One of the problems commonly encountered in time series analysis is *outliers*, which are distinct from most of the other observations. Outliers in the atmospheric data from the Arinc, Inc. experiment might result from a gross error such as a recording or typing error, or from a non-repetitive exogenous intervention such as a weather disaster. Hence some outliers might be considered as valid observations and for those cases, the model for the time series data will need to take those observations into account (Chatfield, 1984). On the other hand, outliers might be considered as freak observations and for those cases, the outliers need to be adjusted or removed from the data set before further analysis of the data. If not treated properly, outliers tend to distort the estimates of model parameters that produce unrealistic prediction errors.

In practical applications, Collett and Lewis (1976) stated that an outlier detection procedure was a subjective decision by the analyst that outliers would be searched for in the time series data set. Chatfield (1984) also stated that the treatment of outliers was a complex subject in which common sense was as important as theory.

Robinson (1979) and Ljung (1993) suggested that after knowing the locations of outliers, these outliers should be treated as missing observations in the time series data set. Then, a method of prediction for handling missing observations (as proposed in the next subsection) can be used to replace them with their predicted values.

2.2.2 Missing observations

AR and ARMA models are commonly used to model univariate time series data and to describe autocorrelated errors in regression models involving time series data (Chatfield, 1984). Most estimation methods used to determine the unknown parameters of these models are developed under the assumption that data are available at consecutive and equally spaced time intervals. However, time series data occasionally have some missing observations that impact unknown parameter estimation. If there are missing observations in time-series data sets, the results from time series modeling might be misleading and the statistical inference testing might be biased and inefficient. Many researchers have tried to solve this problem. The following paragraphs represent an overview of some works developed to deal with missing observations.

Some researchers have developed likelihood functions of various time series models to obtain parameter estimates and have claimed that the missing observations could be estimated with parameter estimates. Jones (1980) developed a form of the likelihood function for the autoregressive moving average model (ARMA) by using a Kalman filter approach for handling missing observations in time series data set. Note that the Kalman filter approach is an iterative computational algorithm used to calculate predictive values and forecast variances in time series models. Fuller (1996) demonstrated Jones's technique on a simulated time series data set from a second order

autoregressive model (AR(2)) with a few missing observations. Basu and Reinsel (1996) proposed a form of the likelihood function for the autoregressive moving average model (ARMA). They demonstrated their technique on a time series data set of monthly averages for total ozone from the Dobson spectrophotometer at Huancayo, Peru, obtained over the period January 1978-December 1991.

Ratinger (1996) proposed an *ad hoc* technique applied to a seasonal autoregressive integrated moving average (ARIMA) time-series data set with a gap of missing observations. Ratinger suggested that a linear filter representation should be added in the seasonal ARIMA model before filling in a gap of missing observations. The author also derived this *ad hoc* approach using minimum mean square error smoothing constants in the seasonal ARIMA time-series data set. The smoothing constants are the functions of the missing observations within a gap.

In addition to the techniques described thus far, meteorologists have tried to propose the techniques for handling the problem of missing observations in weather data. Kemp et al. (1983) used available data from one or more adjacent weather stations to develop a prediction equation as a basis for estimating missing observations for daily maximum and minimum temperatures. Based on the assumption that the difference between daily temperatures at the adjacent stations was equal to the difference between the monthly average temperatures of the adjacent stations, Kemp et al. (1983) performed regression analysis on available data from the adjacent stations as a basis for estimating the missing observations. DeGaetano et al. (1995) also proposed a technique to estimate missing daily maximum and minimum temperatures. Their technique used the nearest available station data to reconstruct missing temperature values. Their technique also considered the differences in observation time between the missing-data station and those

used in estimation. Only stations with observation times similar to that of the missing-data station were used. However, if a sufficient number of stations with a similar observation time could not be identified, adjustments were made to simulate the maximum or minimum temperature corresponding to the appropriate time of observation. After obtaining the sufficient number of stations, the missing daily maximum and minimum temperatures were determined by prediction and interpolation using all available data from the stations within the same climate division.

However, the techniques for handling missing observations described thus far have some restrictions and assumptions for each type of time series models. The technique of Jones (1980), for example, assumed that all observations had the same probability of missing. For the Arinc, Inc. atmospheric data, the problems of missing observations might arise from measurement errors, equipment failure, loss of power to the datalogger, or natural disasters where the probabilities of these causes are not the same. Moreover, the techniques presented are appropriate for estimating only a few missing observations or a single gap of missing observations. In the meteorological case, the data from adjacent weather stations were available for filling missing observations by applying the method of regression.

Neural networks have been widely used for time series prediction in various applications such as market predictions, meteorological, and network traffic forecasting. Feed-forward networks are one of the most often used approaches for time series prediction. Tang et al. (1991) performed time series forecasting of market sales using the feed-forward networks and the conventional Box-Jenkins time series approach. By comparing the performances of the neural networks and the conventional approach, Tang et al. found that the neural network is a better choice for long-term forecasting. In this

research, the feed-forward networks with the BP algorithm are used to predict missing observations in long-term forecasting by applying the model parameters studied by Tang et al. (1991).

A neural networks model consists of: 1) a set of inputs, 2) a set of weights or connecting links, 3) an adder for summing the input signals and weights by the respective weights of the neural, 4) an activation function to deliver an output, and 5) a set of desired outputs (Haykin, 1999).

Figure 2.3 illustrates a neural network architecture. A set of input-output pairs is referred to as a set of training data or training sample. The weight of each connection in the network is a function of a learning rate and a momentum value. The larger learning rate makes the network become unstable (i.e., oscillatory) even though the network increases the rate of learning. To alleviate this problem, a momentum value is included in each network revision of weight. Note that the learning rate is the rate of network convergence, which is evaluated by the mean square errors of each iteration. The momentum value is a positive constant, which is used to control the feedback loop. Tang et al. (1991) performed simulation to investigate the effect of training parameters (i.e., momentum value and learning rate). They found that the network converged very quickly at a low learning rate (e.g., 0.1) and high momentum (e.g., 0.9).

Activation function such as threshold function, piecewise linear function, and sigmoid function, is defined for limiting the amplitude range of the output signal to some finite value of a neuron.

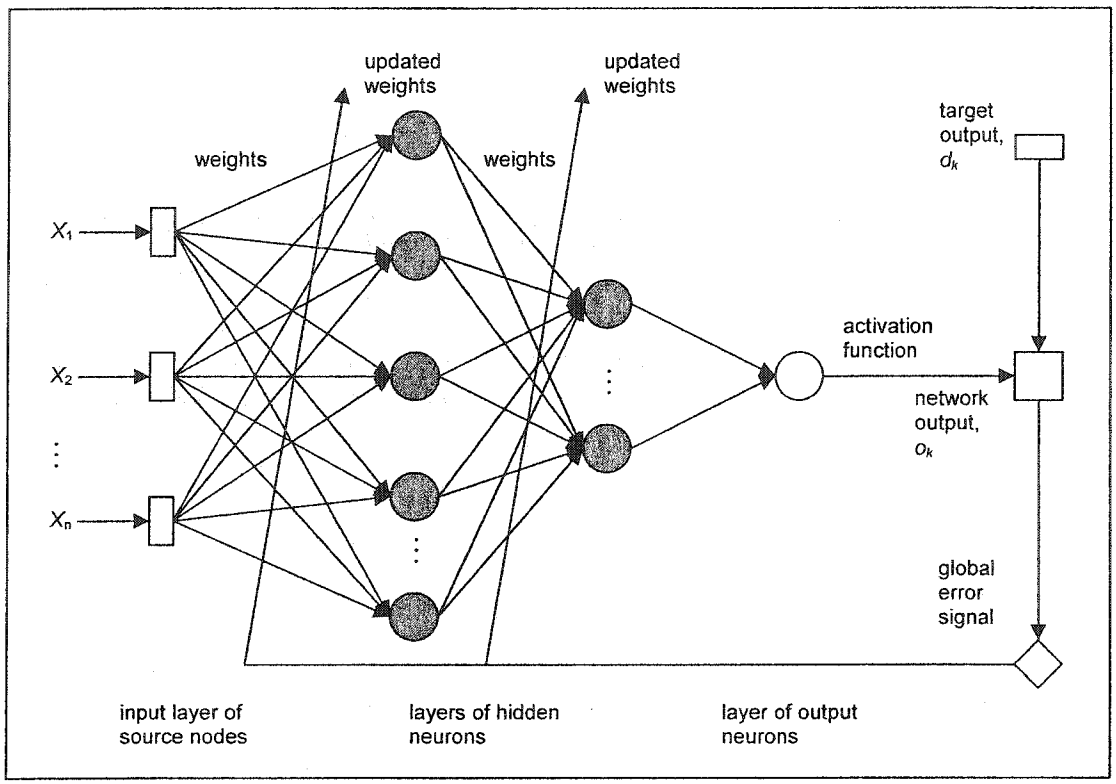


Figure 2.3: A neural network architecture

The number of neurons in the input and output layers are determined by the number of input and output variable(s) whereas the number of neurons in the hidden layer(s) can be established by randomly setting up the initial values. The global error representing the difference between the network outputs and the desired outputs is used as a criterion for deciding the number of neurons in the hidden layer(s). Thus, trial and error is performed to determine the number of neurons of the hidden layer(s) during the training process.

An important class of neural networks, feed-forward networks, has been applied successfully to solve some difficult and diverse problems using a popular algorithm known as the back propagation (BP) algorithm. The feed-forward networks with the BP

algorithm works by propagating the input signal through the network in a forward direction, on a layer-by-layer basis (Haykin, 1999). Generally, neural networks obtain knowledge of a process by training with input and output data. Each neuron (or node) receives information from several input data sources by summing input data with a connection weight and using an activation (or transfer) function to deliver an output. Similarly, each neuron of the hidden layers also receives information from the outputs of all input layer nodes. In the output layer, each neuron uses an activation function (e.g., sigmoid function) to produce an output. The outputs from the network and the target outputs from actual data are used to calculate a global error. The objective of the training process in the BP algorithm is to adjust the weight of each connection in the network to minimize a global error function. The training process is maintained on an iteration-by-iteration (or epoch-by-epoch) basis until the weight of each connection in the network stabilizes and the global error function converges to some minimum value based on a pre-established stopping criteria. The performance of the trained network is tested with data that the network has not seen before and then, assessed by comparing the actual data and the predicted values from the trained neural networks.

2.3 Preliminary Analysis

This section provides multivariate preliminary analysis for uncovering multivariate relationships. The data obtained from the Arinc, Inc. experiment is highly dependent on each other and exhibits several multivariate relationships. The main tools used for multivariate preliminary analysis are scatterplot matrix and correlation analysis (Weihs, 1993). In a real application of climatological data analysis, Wilks (1995) used a scatterplot matrix to reveal the relationship among the climatological variables (i.e.,

precipitation, temperature) at Ithaca and Canadaigua, New York, for January 1987. Kajiyama and Koyama (1997) investigated correlation between the maximum corrosion depth and 21 environmental factors measured along a 159 meter bare, ductile cast iron pipeline route at intervals of 1 meter. Soil resistivity, corrosion rate of the probe rod, corrosion potential, and pipe-to-soil potential were obtained by in-situ measurements at ground level in the field before excavation. Other environmental factors (e.g., water content, pH, content of ferrous sulfide [FeS], hydrogen peroxide [H_2O_2]) were measured in the laboratory by examining soil samples. Upon completion of surveys of the environmental factors, the maximum corrosion depth measurements were conducted along the 159-meter pipeline route. Kajiyama and Koyama showed the results of the correlation analysis among the maximum corrosion depth and the environmental factors. They concluded that the environmental factors correlating with the maximum corrosion depth were pipe-to-soil potential, specific gravity, and the content of ferrous sulfide.

Plotting many variables against each other in the preliminary stages of data analysis is a good practice that allows examination of the data in a straightforward fashion. The scatterplot matrix is one of the statistical analysis tools used to interpret data by graphically displaying the relationship among variables (Weihs, 1993). A scatterplot matrix of p variables is a matrix of graphs with a scatterplot of the ith variable against the jth variable as the (i, j) entry of the matrix ($i \neq j$) and without the diagonal elements in the positions.

The scatterplot matrix can be interpreted from the data patterns as positive, negative, or no relationship. A positive relationship between the two variables is displayed by an ellipse of points that slopes upward, demonstrating that an increase in the cause variable also increases the effect variable. On the other hand, a negative

relationship between the two variables is displayed by an ellipse of points that slopes downward, demonstrating that an increase in the cause variable results in a decrease in the effect variable. A plot indicates that there is no relationship between the two variables if a cluster of points is difficult or if it is impossible to determine whether the trend is upward sloping or downward sloping. Furthermore, data patterns, whether in a positive or negative direction, should be interpreted for strength by examining the tightness of the clustered points. Note that the more the points are clustered to look like a straight line, the stronger the relationship.

When the scatterplot matrix is performed, the next step is to measure the inter-relationship among p variables using correlation analysis. Note that the two variables are considered to be random variables. The output of this measurement is called the correlation coefficient and it ranges between –1.0 and +1.0. A correlation coefficient of +1.0 is a perfect positive correlation whereas a correlation coefficient of -1.0 is a perfect negative correlation. A correlation matrix is useful in obtaining a preliminary impression of the interrelationships among the variables. Moreover, a correlation matrix can be used to check for multicollinearity (i.e., interrelationship among the independent variables). The diagonal terms of the correlation matrix will always be 1 since each variable is perfectly correlated with itself whereas the off-diagonal terms are limited to the range –1.0 to +1.0. Note that the correlation matrix is symmetrical. Consequently, the ease with using the correlation matrix among random variables as a measure of interrelationship is that it is dimensionless, which makes its interpretation convenient. When using variables measured in either different units or the same units with large different values among variables, a correlation matrix should be used rather than a covariance matrix. However, the correlation matrix can be determined from a covariance matrix.

Let us consider the p independent random variables $x_1, x_2, \ldots, x_p$ as constituting a random vector

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}. \tag{2.6}$$

The mean vector is

$$E(\mathbf{x}) = E\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{bmatrix} = \boldsymbol{\mu}. \tag{2.7}$$

The covariance matrix of $\mathbf{x}$ denoted by Σ is

$$\boldsymbol{\Sigma} = E(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})'. \tag{2.8}$$

The i^{th} diagonal element of the covariance matrix, σ_{ii}, is the variance of the i^{th} component of $\mathbf{x}$ so that we can denote this by σ^2_i. We can define the correlation coefficient denoted by ρ_{ij} between the two random variables, x_i, and x_j, as follows:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}, \qquad i, j = 1, 2, ..., p. \tag{2.9}$$

Note that $\rho_{ij} = \rho_{ji}$. The covariance matrix may be written as,

$$\boldsymbol{\Sigma} = E\begin{bmatrix} (x_1-\mu_1)^2 & (x_1-\mu_1)(x_2-\mu_2) & \cdots & (x_1-\mu_1)(x_p-\mu_p) \\ (x_2-\mu_2)(x_1-\mu_1) & (x_2-\mu_2)^2 & \cdots & (x_2-\mu_2)(x_p-\mu_p) \\ \vdots & \vdots & \ddots & \vdots \\ (x_p-\mu_p)(x_1-\mu_1) & (x_p-\mu_p)(x_2-\mu_2) & \cdots & (x_p-\mu_p)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pp} \end{bmatrix} = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \cdots & \sigma_1\sigma_p\rho_{1p} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2^2 & \cdots & \sigma_2\sigma_p\rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_p\sigma_1\rho_{p1} & \sigma_p\sigma_2\rho_{p2} & \cdots & \sigma_p^2 \end{bmatrix}. \tag{2.10}$$

If Equation 2.10 is divided by $\sigma_i\sigma_j$, $i, j = 1, 2, \ldots, p$, we can obtain a correlation matrix denoted by ρ as,

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1p} \\ \rho_{21} & 1 & \cdots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & 1 \end{bmatrix}. \tag{2.11}$$

However, the parameters described thus far are point estimates of the population quantities. With the sample of n observations, let the sample mean vector, $\bar{\mathbf{x}}$ be

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^{n}\mathbf{x}_k = \begin{bmatrix} \frac{1}{n}\sum_{k=1}^{n}x_{1k} \\ \frac{1}{n}\sum_{k=1}^{n}x_{2k} \\ \vdots \\ \frac{1}{n}\sum_{k=1}^{n}x_{pk} \end{bmatrix} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} \tag{2.12}$$

and the sample covariance matrix, S, be

$$\mathbf{S} = \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{bmatrix}, \tag{2.13}$$

where s_i^2 is the sample variance and the sample covariance is

$$s_{ij} = \frac{n\sum_{k=1}^{n} x_{ik}x_{jk} - \sum_{k=1}^{n} x_{ik}x_{jk}}{n(n-1)}, \qquad i,\ j = 1,2,\ldots,p. \tag{2.14}$$

Note that $s_{ij} = s_{ji}$. A sample correlation matrix denoted by $\mathbf{R}$ is

$$\mathbf{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{bmatrix}, \tag{2.15}$$

where r_{ij}, $i, j = 1, 2, \ldots, p$ is the sample correlation defined by

$$\begin{aligned} r_{ij} &= \frac{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik} - \bar{x}_i)^2}\sqrt{\sum_{k=1}^{n}(x_{jk} - \bar{x}_j)^2}}, \qquad i,\ j = 1, 2,\ldots,p \\ &= \frac{\sum_{k=1}^{n} x_{ik}x_{jk} - n\bar{x}_i\bar{x}_j}{\sqrt{\sum_{k=1}^{n} x_{ik}^2 - n\bar{x}_i^2}\sqrt{\sum_{k=1}^{n} x_{jk}^2 - n\bar{x}_j^2}}, \qquad i,\ j = 1, 2,\ldots,p. \end{aligned} \tag{2.16}$$

Note that once again, $r_{ij} = r_{ji}$. The computation of these correlations defined in Equation 2.16 can be expressed in matrix notation as

$$\mathbf{R} = \mathbf{D}^{-1}\mathbf{S}\mathbf{D}^{-1}, \tag{2.17}$$

where **D** is a $p \times p$ diagonal matrix whose diagonal elements are the sample standard deviations of the p variables (i.e., the square roots of the corresponding diagonal elements of **S**).

2.4 Principal Component Analysis

Principal component analysis (PCA) is an approach used for summarizing the data so that no dependent variable exists (Afifi and Clark, 1990). The summary variables, called principal components, are computed from all of the original independent variables. Another definition of PCA is a multivariate approach in which a number of related variables are linearly transformed to set of uncorrelated variables (Jackson, 1980, 1991). Furthermore, this technique is used to reduce the number of variables without losing much of the information (Afifi and Clark, 1990).

2.4.1 PCA Theory

The procedure of PCA is to transform a vector of the original correlated variables, $[x_{11}\ x_{12}\ \ldots\ x_{1p}]$, into a vector of the new variables, $[z_{11}\ z_{12}\ \ldots\ z_{1p}]$, that are uncorrelated with each other. Note that these linear combinations represent the selection of a new coordinate system obtained by rotating the original system with x_{11}, x_{12}, ..., x_{1p} as the coordinate axes. The logic of rotating the original coordinate system is to maximize the variance of the new coordinate system. After the first coordinate axis on which the variance is maximal, there remains some variability around this coordinate axis. In principal component analysis, after the first coordinate axis has been extracted (i.e., after

the first coordinate axis has been drawn through the data), another coordinate axis will be extracted in order to maximize the remaining variability, and so on. In this manner, consecutive coordinate axes are extracted. Since each consecutive coordinate axis is defined to maximize the variability that is not captured by the preceding coordinate axis, consecutive coordinate axes are independent of each other. This implies that consecutive coordinate axes are uncorrelated or orthogonal to each other. The original data can be transformed into the new coordinate axes by this manner. The general form of orthogonal transformation is defined as

$$\mathbf{Z} = \mathbf{XU}. \tag{2.18}$$

This model can be written as

$$\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \cdots & z_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ z_{n1} & z_{n2} & \cdots & z_{np} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}. \tag{2.19}$$

Let $\mathbf{x}_k$ be a vector of the original independent variables for each observation, k defined as $[x_{k1} \quad x_{k2} \ldots x_{kp}]$, $k = 1, 2, \ldots, n$. Let $\mathbf{z}_k$ be a vector of principal component scores of $\mathbf{x}_k$ for observation k defined as $[z_{k1} \quad z_{k2} \ldots z_{kp}]$, $k = 1, 2, \ldots, n$ where z_{kp} is a principal component score of $\mathbf{z}_k$. Let $\mathbf{U}$ define as $[\mathbf{u}_1 \quad \mathbf{u}_2 \ldots \mathbf{u}_p]$. Thus, the orthogonal transformation for each observation, k, is defined as follows

$$\mathbf{z}_k = \mathbf{x}_k \mathbf{U}, \quad k = 1,2,\ldots,n. \tag{2.20}$$

Note that $\mathbf{z}_k$ scores are uncorrelated among the principal component scores. Conversely, $\mathbf{x}_k$ for observation k can be determined by

$$\mathbf{x}_k = \mathbf{z}_k \mathbf{U}', \quad k = 1,2,...,n. \tag{2.21}$$

Since $\mathbf{U}$ is an orthogonal matrix $p \times p$ (i.e., $\mathbf{U}'\mathbf{U} = \mathbf{I}$ and $\mathbf{U}\mathbf{U}' = \mathbf{I}$, where $\mathbf{I}$ is identity matrix), the transpose of $\mathbf{U}$ is equal to its inverse. Thus, $\mathbf{U}'$ in Equation 2.24 can be replaced by $\mathbf{U}^{-1}$. The vectors of matrix $\mathbf{U}$ are of unit length are defined in Equation 2.22 below:

$$\begin{aligned} \left(\mathbf{u}_i{}'\mathbf{u}_i\right)^{\frac{1}{2}} &= 1, \quad i=1, 2,..., p \\ \left(\mathbf{u}_j{}'\mathbf{u}_j\right)^{\frac{1}{2}} &= 1, \quad j=1, 2,..., p. \end{aligned} \tag{2.22}$$

The coordinate axes of the uncorrelated variables are described by the vectors $\mathbf{u}_i$ that make up the matrix $\mathbf{U}$ of direction cosines. The columns of $\mathbf{U}$ are called *characteristic vectors* or *eigenvectors*. In this research, eigenvectors will be employed rather than characteristic vectors. A first column vector of $\mathbf{U}$ defined as $\mathbf{u}_1$ or $[u_{11}\ u_{21}\ \ldots\ u_{p1}]'$ represents the coefficients of the first principal component (pc) called the first eigenvector of the covariance or correlation matrix, a second vector $\mathbf{u}_2$ or $[u_{12}\ u_{22}\ \ldots\ u_{p2}]'$ represents the coefficients of the second pc called the second eigenvector, and the last vector $\mathbf{u}_p$ or $[u_{1p}\ u_{2p}\ \ldots\ u_{pp}]'$ represents the coefficients of the p pc called the p eigenvector. Thus, principal components are particular linear combinations of the p original random variables with coefficients equal to the eigenvectors of the covariance or correlation matrix. This implies that in Equation 2.20 each principal component (i.e., an element of $\mathbf{z}_k$ score) is a linear combination of an observation of the original variables (i.e., $\mathbf{x}_k$) with coefficients (i.e., $\mathbf{u}_i$ that is equal to the eigenvectors of the correlation or covariance matrix) as expressed in Equation 2.23

$$\begin{bmatrix} z_{11} & z_{12} & \cdots & z_{1p} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} & \cdots & u_{1p} \\ u_{21} & u_{22} & \cdots & u_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p1} & u_{p2} & \cdots & u_{pp} \end{bmatrix}. \tag{2.23}$$

A $p \times p$ covariance matrix **S** can be reduced to a diagonal matrix **L** by premultiplying and postmultiplying it by a particular orthogonal matrix **U**. Therefore,

$$\mathbf{L} = \mathbf{U}'\mathbf{S}\mathbf{U}, \tag{2.24}$$

where

$$\mathbf{L} = \begin{bmatrix} l_1 & 0 & \cdots & 0 \\ 0 & l_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l_p \end{bmatrix}.$$

Here **S** is the covariance matrix of the original variables while matrix **L** is the covariance matrix of the principal components. The diagonal elements of matrix **L** are called *characteristic roots* or *eigenvalues* that are correspondingly equal to the variances of **Z** when all observations of the original variables (i.e., $\mathbf{x}_k$, $k = 1, 2, \ldots, n$) are transformed into the new variables (i.e., $\mathbf{z}_k$, $k = 1, 2, \ldots, n$). In this research, again, eigenvalues are used rather than characteristic roots. Note that matrix **S** will be replaced by matrix **R** if eigenvalues are obtained from correlation among the variables.

It is worth noting that the principal components are sorted by descending order of eigenvalues (i.e., from the largest variance of any linear combination of the original variables to the smallest variance of any linear combination of the original variables). This implies that the first principal component accounts for as much variation in the data

as possible while each succeeding principal component accounts for as much of the variation unaccounted for by preceding principal components as possible. Consequently, PCA can be used for identifying the atmospheric conditions most conducive to corrosion growth.

Note also that the off-diagonals of the matrix **L** representing the covariance are zero. This means that z_{11}, z_{12},..., z_{1p} are uncorrelated. Furthermore, the determinants of the two matrices, **S** and **L**, are the same. The trace of the covariance matrix **L** (i.e., the sum of eigenvalues) is equal to the sum of the original variances; this is known as one of the important properties of PCA. According to this property, the percentage or proportion of the total variability is accounted for by each pc as defined as follows

$$\%\mathrm{pc}_i = \frac{l_i}{\sum_{i=1}^{p} l_i} \times 100, \tag{2.25}$$

where $\%\mathrm{pc}_i$ is the percentage of the total variability accounted for by each pc. This value is important for pc interpretations. Eigenvalues can be determined from the solution of the determinental equation, called the characteristic equation, as follows (Jackson, 1980 and 1991):

$$|\mathbf{S} - \mathit{\Pi}| = 0$$

$$\left| \begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{bmatrix} - \begin{bmatrix} l & 0 & \cdots & 0 \\ 0 & l & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l \end{bmatrix} \right| = 0$$

$$\left|\begin{bmatrix} s_1^2 - l & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 - l & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 - l \end{bmatrix}\right| = 0, \tag{2.26}$$

where **I** is the identity matrix. The values $l_1, l_2, \ldots, l_p$ are found by this equation to produce a pth degree polynomial in l. Note that the matrix **S** will be replaced by the matrix **R** if eigenvalues are obtained from the correlation matrix.

If $\mathbf{u}_i \neq \mathbf{0}$ is a p-dimensional vector defined in Equation 2.22, it will be an eigenvector of **S** corresponding to eigenvalue l_i if

$$\begin{aligned} &\mathbf{S}\mathbf{u}_i = l_i\mathbf{u}_i \qquad \text{or} \\ &[\mathbf{S} - l_i\mathbf{I}]\mathbf{u}_i = \mathbf{0}. \end{aligned} \tag{2.27}$$

There are many methods and algorithms to solve Equation 2.27 in order to obtain the corresponding eigenvectors. A set of homogeneous linear equations is the one used to determine the eigenvectors by replacing $\mathbf{u}_i$ with $\mathbf{t}_i$ as defined in Equation 2.28. This vector, $\mathbf{t}_i$ must be normalized to unit length in order to obtain an eigenvector $\mathbf{u}_i$ from Equation 2.29.

$$\begin{aligned} &[\mathbf{S} - l\mathbf{I}]\,\mathbf{t}_i = \mathbf{0} \\ &\left(\begin{bmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{bmatrix} - \begin{bmatrix} l & 0 & \cdots & 0 \\ 0 & l & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & l \end{bmatrix}\right)\begin{bmatrix} t_{11} \\ t_{21} \\ \vdots \\ t_{p1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \\ &\begin{bmatrix} t_{11}(s_1^2 - l) + t_{21}s_{12} + \ldots + t_{p1}s_{1p} \\ t_{11}s_{21} + t_{21}(s_2^2 - l) + \ldots + t_{p1}s_{2p} \\ \vdots \\ t_{11}s_{p1} + t_{21}s_{p2} + \ldots + t_{p1}(s_p^2 - l) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \end{aligned} \tag{2.28}$$

$$\mathbf{u}_i = \frac{\mathbf{t}_i}{\sqrt{\mathbf{t}_i' \mathbf{t}_i}}, \quad i = 1,2,...,p. \tag{2.29}$$

It is worth noting that $\mathbf{u}_i$ in Equation 2.29 makes up the matrix **U**. The matrix **S**, once again, will be replaced by the matrix **R** if eigenvalues are obtained from the correlation matrix.

However, diagnostic analyses should be checked for precise transformation by determining the correlations between the principal components and the original variables as follows:

$$r_{ij} = \frac{u_{ji}\sqrt{l_i}}{s_j}, \tag{2.30}$$

where r_{ij} is the correlation coefficients between the i^{th} pc and the j^{th} original variables and u_{ji} is an element of **U** matrix. l_i is an eigenvalue determined from matrix **S**. The value s_j in matrix **S** will be replaced by the value r_{ij} of the original variable matrix **R**. The matrix $\mathbf{R}_{zx}$, defined for the correlation coefficients between the i^{th} pc and the j^{th} original variables, will be

$$\mathbf{R}_{zx} = \begin{array}{c} \begin{matrix} x_{i=1} & x_{i=2} & \cdots & x_{i=p} \end{matrix} \\ \begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ r_{21} & r_{22} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pp} \end{bmatrix} \end{array} \begin{matrix} z_{i=1} \\ z_{i=2} \\ \vdots \\ z_{i=p}. \end{matrix} \tag{2.31}$$

It is clear to say that the first pc (i.e., $z_{i=1}$ or $[r_{11}\ r_{12}\ \ldots\ r_{1p}]$) is more highly correlated with the original variables than the other ones since the first pc explains more variability than the other ones. It is worth noting that

$$\sum_{i=1}^{p} r_{ij}^{\ 2} = 1, \quad j=1,2,...,p. \tag{2.32}$$

Stopping rules are the criteria used to determine the number of retained variables. As mentioned before, PCA is used to reduce the number of variables without losing much of the information by retaining the first few principal components based on the stopping rules. Based on generally accepted rules of thumb, the following criterions should be followed when determining how many principal components should be considered (Jackson, 1991):

1. Continue to estimate principal components until the cumulative eigenvalues contribution is significant at the chosen level (e.g., 80%)
2. Continue to estimate principal components until the variance (eigenvalue) exceeds one (i.e., the variance of one of the original variables)
3. Discontinue calculating principal components immediately prior to an abrupt decrease in the magnitude of the characteristic value.

Cattell (1966) developed a graphical technique for determining the number of retained variables, named scree. This method is concerned only with a plot of eigenvalues of a covariance matrix or a correlation matrix and the number of components. The scree plot can be helpful in deciding the appropriate number of retained variables. As mentioned before, eigenvalues are ordered from largest to smallest with the first few explaining most of the variability. Generally, the scree plot illustrates a steep drop over the first few components followed by a leveling off for the rest of the components. The criterion of this test is to plot a graph of eigenvalues as the function of the number of variables and try to draw a straight line connecting all eigenvalues. The

number of retained variables can be found if all components have eigenvalues above the straight line.

2.4.2 Applications of principal component analysis

Throughout this section, some applications of PCA are described in the fields of meteorology and criminology, but the ideas may readily be applied to the analysis of corrosion severity ranking with the atmospheric condition data sets. Note that none of the works addressed any assumptions about the data in terms of its applicability to PCA.

Principal component analysis has been used in meteorological fields for studying the interrelationships among several meteorological variables. The concept of eigenvectors in principal component analysis is very useful and interpretable for the multivariate data. Eigenvectors are derived from the data being studied and strongly resemble the important features of the data. Maximum variance is accounted for by choosing in order the eigenvectors associated with the largest eigenvalues of the approximate covariance matrix. Stidd (1967) described the use of eigenvectors for climatic estimates from the sets of 12 mean-monthly precipitation values for 60 stations in Nevada. As expected from climatological considerations, the data pattern is high in winter and low in summer. However, the variability among stations in Nevada is large so that it is difficult to determine precisely the month-to-month differences in the data pattern. The first three eigenvectors for the mean-monthly precipitation data of the 60 Nevada stations account for 93% of the variation, which has been deemed sufficient accuracy. The author stated that the first eigenvector resembled the annual cycle of winter storms. The second eigenvector represented the annual cycle of summertime, convective precipitation, while the third eigenvector represented the effect of late spring

and early fall rainfall observed at the most stations. This showed that the eigenvectors were strongly related to the influences of several separate natural processes. Stidd performed the vectors of pc scores (i.e., the linear combinations of the first three eigenvectors and the 12 monthly values for the 60 Nevada stations) and plotted the first three vectors of pc scores in the map of the state of Nevada. The author found that the spring and fall peaks of precipitation were most prominent in Northeast Nevada (i.e., the indication of some positive signs in the third vector of pc scores). Furthermore, the author found that the three maps and the corresponding eigenvectors supply all the information needed to estimate the mean monthly precipitation for any area in the state of Nevada (Stidd, 1967).

Similarly, Kutzbach (1967) implemented the methods of eigenvectors and PCA for the three climatic variables (i.e., monthly mean sea-level pressure [SLP], surface temperature, and precipitation). The author found that the patterns of the first several eigenvectors of the three variables illustrated realistically the covariance structure of the three variables that was consistent. From the first vector of pc scores plotted in the map of North America, it was clear that the center of high SLP variability was in the center of the continent. Moreover, the results showed that the eigenvectors could be of considerable descriptive or diagnostic value.

The methods of eigenvectors and principal component analysis are general and useful technique of statistically summarizing the variability of winds. Since the wind velocity observations are not decomposed into direction and speed, vector-based PCA is the most appropriate method for investigating wind fields. Klink and Willmont (1989) used the methods of eigenvectors and PCA to analyze and quantify the characteristic of the winds. They used the data of the surface winds for 1975, derived from 68 stations

across the United States. Each station collected wind direction and wind speed recorded every three hours. For each station, they obtained the time series of the three-hourly wind direction and wind speed observations. These time-series data sets then were interpolated to a regular 2° of latitude by 2° of longitude using the spherical interpolation routine and grid-point interpolation and contouring.

A detailed description on the applications of eigenvectors and PCA in meteorological fields can be found in Wilks (1995).

Besides PCA applications on meteorological data, Ahamad (1968) performed an analysis of crimes by the method of principal component analysis. The data consisted of frequencies of occurrence of eighteen types of crime (e.g., homicide, assault, homosexual offence) for fourteen years (1950-1963), in England and Wales. The eighteen types of crime represented variables whereas the fourteen years represented observations. The objective of his study was to investigate the relationships among several different crimes and to determine to what extent the variation in the frequencies of occurrence of types of crime from year to year could be explained by a small number of uncorrelated variables. The analysis showed that the first three principal components accounted for 92% of the total variance. These components suggested that much of the increase in the crime rate could be explained by the rapid increase in population.

An application of PCA can be found in the study of the Olympic track record ranking by Dawkins (1989). Dawkins used PCA to perform national track record ranking study of the fifty-five countries. The national records in track events for women included 100, 200, 400, 800, 1500, 3000 meters, and marathon whereas the national records for men included 100, 200, 400, 800, 1500, 5000, 10000 meters, and marathon. The women's and the men's sets were treated separately. Since the time units of these data

sets were different (i.e., the time unit for 100, 200, and 400 was seconds whereas the time unit for the remaining events was minutes), each data set needed to scale to give mean zero and unit standard deviation. If the original data were analyzed using the same time units, the marathon data would swamp the effect of the other events and also would be weighted excessively in the analysis by PCA. The first principal component appeared to be interpretable as the overall athletic excellence among the nations whereas the second principal component represented a contrast between the related times of the short and long distances. Since the first principal component accounted for maximum variation, it was reasonable to use the first principal component for ranking the countries. The comparisons between the results from PCA and from the Olympic rank were given for both women's and men's events.

Naik and Khattree (1996) revisited the Olympic track record ranking where principal component analysis was utilized. In order to compare the athletic performances of the nations, Naik and Khattree stated that the appropriate variables were the speeds rather than the total time taken because these variables succeeded in retaining the possibility of having different degrees of variability in different variables. These variables were defined as the distances (in meters) covered per second for the various track events, which were in the same unit. For example, the women USA's athletic runner speed for 100 meters was calculated as 100 meters/10.81 seconds. Thus, unlike the starting point of PCA with the correlation matrix used in Dawkins's study, Naik and Khattree used the covariance matrix to perform national track record ranking study of the fifty-five countries. The PCA results showed that the first principal component represented a weighted average of all the speeds in the various events, which measured the overall athletic excellence among the nations whereas the second principal

component appeared to be interpretable as the measure of differential achievement. Based on the first principal component, the results of rankings of nations were compared to Dawkins's analysis. The nations in the top ten lists for men were the same as those given by Dawkins, except that Kenya and France were switched from their previously assigned ranks as ninth and eighth, respectively. On the other hand, the nations in the top ten lists for women showed more contrasting rankings.

Another application of PCA can be found in the SAS® (1990) handbook, showing a criminal rate study of the fifty states of the United States. The data consisted of crime rates per 100,000 people in seven categories for each of the fifty states. The seven categories represented variables including murder, rape, robbery, assault, burglary, larceny, and auto, while the fifty states represented observations. The objective of this study was to investigate regional trend of crime rate using the method of PCA. The analysis showed that the first three principal components accounted for 87% of the total variance. The interpretation of the first principal component was a measure of overall crime rate whereas the interpretation of the second principal component was a measure of the prevalence of property crime (e.g., robbery, burglary) over violence crime (e.g., murder, rape). However, the interpretation of the third principal component was not obvious. The plot of the first two principal components showed the trend of crime rate of the fifty states. The states with high overall crime rates were indicated at the extreme right of the plot whereas the states with low overall crime rate were indicated at the extreme left of the plot.

The idea of meteorology, national track records, and criminology described thus far can be applied to the analysis of corrosion severity ranking by locations (i.e., operational air force bases) with atmospheric condition data sets. An objective of this

research is to rank corrosion severity by locations based on the atmospheric condition data sets captured from the six operational air force bases. The procedure of this analysis will be described in more detail in Chapter 3.

2.4.3 The data used in principal component analysis

Since none of the previous work addressed any underlying assumptions for the data before PCA can be applied, the following represents work that has been done with PCA on various data set types. PCA was used in situations where the variables were measured at a variety of scale levels such as nominal, ordinal, or interval (Young et al., 1978). Young et al. proposed a technique for handling these situations. Their procedure is to scale all of the variables in standard units and obtain a solution that minimizes the sum of squares by iterative procedures. They gave an illustrative example of a cylinder problem in which the data specify 12 physical characteristics of 30 cylinders. The 12 physical characteristics include aspects of their height, volume, electrical resistance, moment of inertia, etc. The objective of their study was to demonstrate their procedures on mixed measurement level multivariate data. The results showed that the minimum sum of squares was 0.0079 and the variance accounted for by each variable was 1 after 30 iterations. Their results are identical to the true underlying structure that was lending credence to their procedure.

PCA was also used to analyze real number data sets (e.g., continuous, discrete, proportion, and percentage) and complex data sets. Jackson (1991) gave many examples using PCA with continuous and discrete data sets. Audiometric data set is a continuous data set, which measures hearing loss from 100 subject males, all aged 39, who presumably had no indication of noise exposure or hearing disorders. An instrument

called an audiometer was used to measure the subject's hearing in their left ear and right ear at frequencies 500, 1000, 2000, and 4000. The limits of the instrument are –10 to 99 decibels. The covariance matrix for these 100 observations and corresponding correlation matrix were used with the original data for the starting point in PCA. The objective of his study was to distinguish between normal hearing and hearing loss people. Moreover, this study demonstrated what differences occurred if the starting point was the original data sets using the covariance matrix and the correlation matrix. Jackson found that the first principle component represented the overall hearing level of a respondent. This implies that individual suffering hearing loss at certain frequencies was suffering at the other frequencies as well. The second principal component indicated the contrast between the high frequencies and the low frequencies. The results also showed that there were different eigenvalues, eigenvectors, and principal component scores when the starting point was the original data sets using the covariance matrix and the correlation matrix.

Jackson (1991) gave an example of PCA applied to discrete data. The data set consisted of reports of personal assaults in England and Wales for the year 1878-1887 broken down by the quarters of the year. The objective of the author's study was to identify the effect of season on the incidence of crime. The study showed that the incidence of the crimes increased steadily throughout the 10-year period and that there was a higher incidence in the warmer months.

PCA was used to analyze compositional data sets such as vectors of percentages or proportions of various chemical compounds or ingredients. Note that the definition of compositional data is that the sum of these percentages or proportions must be equal to unity. Due to the awkward constraint that the components of each vector must sum to

unity, Aitchison (1983) stated that compositional data are difficult to perform statistically. Aitchison introduced transformation techniques for handling such data. Aitchison's procedure is to transform the original data to new logarithmic data of the following form:

$$c_{ij}^{*} = \ln(c_{ij}) - \frac{1}{p}\sum_{j=1}^{p}\ln(c_{ij}), \quad i = 1,2,...,n \quad j = 1,2,...,p, \tag{2.33}$$

where c_{ij}^{*} is a logarithmic transformed data point, c_{ij} is the original data, p is the number of variables, and n is the number of observations. Then, the logarithmic transformed data sets are used in the PCA procedure. Aitchison gave the illustrative examples on steroid metabolites and Aphyric Skye lavas data sets. Jackson (1991) gave an illustrative example on U.S. budget data from 1967 to 1986. Khattree and Naik (2001) illustrated an example using geology data regarding the proportions of elements in a specimen found in two or more sources.

PCA was also used to analyze complex data sets. The most likely field of application for complex PCA is in time series analysis (Jackson, 1991). In this case, the covariance matrix is a Hermitian matrix. Note that a Hermitian matrix is a unique matrix made up of complex numbers such that the diagonal elements are real and the pair of off-diagonal elements are complex conjugates of each other. The Hermitian matrix is also made up of the sum of the real part that is symmetric and an imaginary part that is skew-symmetric.

For this research, the data are transformed into compositional data based on relationships indicating corrosion growth. The compositional data is then analyzed via PCA to obtain corrosion severity ranking.

2.5 Existing Corrosion Modeling Prediction Models

Predictive models for corrosion growth have been developed by several researchers. Sereda (1960) proposed a multiple regression model that used the sulphate (SO_2) rate and surface temperature to predict the log corrosion rate per day for low carbon plain steel. Haynie et al. (1978) identified sulphate (SO_2), nitrate (NO_2), ozone (O_3), and relative humidity as major factors influencing corrosion growth. They developed a least squares fit model for predicting corrosion measured in terms of weight loss on weathering steel and galvanized steel using a two-level factorial design. Power law equation was used to model accumulated corrosion growth as a function of time. This equation was expressed as

$$P = Kt^m, \tag{2.34}$$

where P is the metal loss or the corrosion penetration depth after t years, K is a constant, and m is an exponent. The power law function is intrinsically linear because it can be transformed to a straight line by a logarithmic transformation as follows:

$$P^* = K^* + mt^*, \tag{2.35}$$

where P^* represents $\log P$, K^* represents $\log K$, and t^* denotes $\log t$.

Pourbaix (1982) used the linear bilogarithmic laws (e.g., power law equation) to extrapolate long-term corrosion up to 20-30 years from four years of tests. The linear bilogarithmic laws expressed as the relationships between time and corrosion penetration depths, mean corrosion rate, and instantaneous corrosion rate, were defined as follows:

$$\log P = K^* + m\log t$$
$$\log\frac{P}{t} = K^* + (m-1)\log t \qquad (2.36)$$
$$\log\frac{dP}{dt} = (K^* + \log m) + (m-1)\log t,$$

where *P/t* is the mean corrosion rate, and *dP/dt* is the instantaneous corrosion rate. Note that this equation is valid and reliable for predicting long-term corrosion of different types of atmospheres and for a number of materials. The author demonstrated the effect of the exposure time on corrosion penetration, mean corrosion rate, and instantaneous corrosion rate of carbon steel, copper steel, weathering steel, galvanized steel, and aluminized steel exposed in industrial climate and marine climate at the locations in the United States and the European countries. With different types of materials exposed at different types of atmospheres, the results showed that the corrosion penetration increased as a function of exposure time whereas the corrosion rates and the instantaneous corrosion rates decreased as a function of exposure time.

It is worth noting that corrosion growth is not linear with time (Roberge, 2000). According to the power law equation, if *m* is equal to 1, this equation will be expressed as a linear model. Moreover, if *m* is larger than 1, this equation will be expressed as a convex curve. This implies that the exponent *m* impacts the characteristics of the power law equation. However, corrosion growth rates do not follow the power law equation if the exponent *m* is larger or equal to 1.

Predictive models of corrosion growth are useful for explaining the durability of metallic structures, determining the economic costs of damages associated with the degradation of metals, and obtaining experience about the effect of atmospheric parameters on corrosion kinetics (Feliu et al., 1993a). Feliu et al. (1993a, 1993b)

developed a linear regression model for predicting corrosion growth in steel, zinc, copper, and aluminum as a function of seven environmental parameters--time-of-wetness, relative humidity, number of rain days per year, temperature, sulfate concentration, chloride concentration, and marine atmospheric quality. In their study, the power law equation was used to develop the predictive models of such materials. They proposed that m is the function of the meteorological and pollutant data worldwide. Thus, the exponent m for different metals is expressed in different values.

Multiple linear regression model was used to investigate the relationship between corrosion growth and atmospheric conditions. In 1993, Bhattacharjee et al. conducted an experiment on atmospheric corrosion of mild steel. The experimental sites included 17 locations in India comprised of industrial areas, coastal areas, and combinations of both. Bhattacharjee et al. developed a multiple linear regression model for predicting corrosion loss as a function of the atmospheric conditions. The rate of atmospheric corrosion was the dependent variable whereas the atmospheric conditions, namely temperature, relative humidity, rainfall, number of rainy days, sulphur dioxide (SO_2), and sodium chloride concentrations (NaCl), were the independent variables. Goodness of the fit, F- and t-tests were used to identify statistically significant parameters. Bhattacharjee et al. found that the major factors influencing corrosion growth were SO_2 for industrial sites and NaCl for coastal sites. After SO_2 and NaCl, temperature was identified to be the next significant factor contributing to atmospheric corrosion.

Although the existing corrosion models have been developed by several researchers, none of them address corrosion growth on operational aircraft. The experiments were designed for steel (not aircraft's aluminum parts) or based on lab-

simulated corrosion. This research develops predictive corrosion growth models for corrosion growth on operational aircraft.

2.6 Growth Models

Growth models are applied in many fields such as biology, botany, forestry, zoology, and ecology (Banks, 1994; Draper and Smith, 1998). Generally, a growth model can be obtained by making assumptions about the type of growth, formulating a differential equation that represent these assumptions, and solving the differential equation. Note that a growth model is a nonlinear model.

2.6.1 Gompertz growth model

This model was developed from an exponential growth model which is defined as

$$\frac{dP}{dt} = aP, \tag{2.37}$$

where P is the magnitude (accumulated value) of the growing quantity and t is time.

At this point it is postulated that the growth coefficient, a, changes with time according to the relationship (Banks, 1994):

$$\frac{da}{dt} = -ka, \tag{2.38}$$

where k is the decay coefficient of the growth coefficient. With the initial condition $a(0) = a_0$, the solution to the differential equation of 2.38 is:

$$\frac{da}{dt} = -ka$$
$$\int \frac{da}{a} = -k \int dt$$
$$\ln a = -kt + c,$$

where c is the constant of integration. The initial condition $a = a_0$ when $t = 0$ gives $\ln a_0 = c$. Hence,

$$\ln a = -kt + \ln a_0$$
$$a = a_0 e^{-kt}.$$

Thus, the principal phenomena of the Gompertz model is the incorporation of an exponentially decreasing growth coefficient. Substituting the solution of this differential equation into Equation 2.37 obtains

$$\frac{dP}{dt} = a_0 e^{-kt} P. \tag{2.39}$$

The solution to 2.39 is:

$$\int \frac{dP}{P} = a_0 \int e^{-kt} dt$$
$$\ln P = -\frac{1}{k} a_0 e^{-kt} + c,$$

where c is the constant of integration. The initial condition $P = P_0$ when $t = 0$ gives

$$\ln P_0 = -\frac{1}{k} a_0 + c \quad \text{or} \quad c = \ln P_0 + \frac{1}{k} a_0.$$

Hence,

$$\ln P = -\frac{a_0}{k} e^{-kt} + \ln P_0 + \frac{a_0}{k}$$
$$P = P_0 \exp\left[\frac{a_0}{k}(1 - \exp(-kt))\right].$$

With initial condition $P(0) = P_0$, the Gompertz model can then be written as:

$$P = P_0 \exp\left[\frac{a_0}{k}(1 - \exp(-kt))\right]. \tag{2.40}$$

In Equation 2.40, it is clear that as $t \to \infty$,

$$P = P_* = P_0 \exp\left(\frac{a_0}{k}\right), \tag{2.41}$$

where P_* is the ultimate limiting value of the growing quantity. Substituting this result into Equation 2.40 obtains an alternative form of the Gompertz model

$$P = P_* \exp\left(-\frac{a_0}{k}\exp(-kt)\right). \tag{2.42}$$

Researchers used the Gompertz growth model extensively in population studies and to represent the course of animal growth. In a biological study, Richards (1959) used the Gompertz growth model to fit the growth data of the length of the hypocotyls of *Cucumis melo* when grown in darkness at different temperatures over thirty days. Richards stated that the magnitudes of the model parameters might be used to assess the importance in growth of experimentally controllable factors. In his study, Richards also gave detail descriptions regarding interpretation of the model parameters in terms of biological implication. Applications of the Gompertz growth model can also be found in Banks (1994). *However, researchers have not used the Gompertz model for fitting corrosion growth data.*

2.6.2 Logistic growth model

The logistic growth model is based on the change, *dP/dt*, that is assumed to be proportional to (i) the current quantitative level, P; and (ii) a condition $1-(P/P_*)$ that makes $dP/dt \to 0$ when $P \to P_*$, where P_* is the ultimate limiting value of the growing quantity.

Hence,

$$\frac{dP}{dt} = aP\left(1-\frac{P}{P_*}\right), \tag{2.43}$$

where a is a growth coefficient. The solution of 2.43 is:

$$\int \frac{dP}{P(1-P/P_*)} = \int a dt.$$

Partial fractioning the integrand on the left hand side gives:

$$\int\left[\frac{1}{P}+\frac{1/P_*}{1-(P/P_*)}\right]dP = at + c,$$

where c is the constant of integration. Hence,

$$\ln P - \ln\left(1-\frac{P}{P_*}\right) = at + c.$$

The initial condition $P = P_0$ when $t = 0$ gives

$$c = \ln\left[P_0/(1-(P_0/P_*))\right]$$

Hence,

$$\ln\frac{P}{(1-(P/P_*))} = at + \ln\left[P_0/(1-(P_0/P_*))\right]$$
$$P = P_*\left[1+((P_*/P_0)-1)\exp(-at)\right]^{-1}.$$

Thus, the logistic growth model is expressed as:

$$P = P_*\left[1+\left(\frac{P_*}{P_0}-1\right)\exp(-at)\right]^{-1}. \tag{2.44}$$

Like the Gompertz growth model, the logistic growth model has been used to fit data of animal growth. Yano et al. (1998) applied the logistic growth model to the data of bacteria growth in a pharmaceutical study. They developed a new pharmacodynamic model for the analysis of in vitro bactericidal kinetics with the bacterial phases divided into two compartments. The model equations of the bacterial growth were expressed as nonlinear simultaneous differential equations and solved by both the simulation and the least squares curve-fitting procedure. *Like the Gompertz growth model, researchers have not used the logistic growth model for fitting data for corrosion growth modeling.*

2.6.3 Confined exponential growth model

The differential equation for the confined exponential growth model is:

$$\frac{dP}{dt} = a_*(P_* - P), \tag{2.45}$$

where a_* is a growth coefficient and P_* is the ultimate limiting value of the growing quantity. Note that a_* and P_* are assumed to be constants. The solution to Equation 2.45 is:

$$\int\frac{dP}{P_* - P} = \int a_* dt$$
$$-\ln(P_* - P) = a_* t + c,$$

where c is the constant of integration. The initial condition $P = P_0$ when $t = 0$ gives

$$c = -\ln(P_* - P_0).$$

Hence,

$$-\ln(P_* - P) = a_* t - \ln(P_* - P_0)$$
$$P_* - P = (P_* - P_0)e^{-a_* t}.$$

Thus, the confined exponential growth model is defined as:

$$P = P_* - (P_* - P_0)e^{-a_* t}. \tag{2.46}$$

If the initial value is zero (i.e., $P(0) = P_0 = 0$), the solution to Equation 2.46 is:

$$P = P_*\left(1 - e^{-a_* t}\right) \tag{2.47}$$

Researchers have used the confined exponential growth model for various applications in physical sciences and engineering in phenomena involving heat transfer and mass transfer. This model has also found applications in social sciences, geography, and agriculture (e.g., tree growth study, crop yields study) (Bank, 1994). *However, once again, researchers have not used the confined exponential growth model for fitting data for corrosion growth modeling.*

CHAPTER 3

METHODOLOGY

This chapter provides the methodology to be used in this research. The flowchart of Figure 3.1 depicts the overall methodology followed. The first section gives a technique of data screening analyses for the atmospheric time series data. In this section, the techniques for handling the problems of outliers and missing observations are given. The second section gives a detailed description of how dew point temperature is generated. The third section provides a detailed information of correlation analysis for time series data sets of the eight atmospheric conditions and justifies the use of PCA. The forth section describes a procedure of the method of PCA for corrosion severity ranking analysis. The fifth section gives a detailed description of corrosion growth modeling development. The last section provides the nonlinear regression method used in this research for corrosion growth modeling.

3.1 Data Screening Analyses

This section provides some data screening techniques for preliminary analysis in time series data. These screening techniques consist of data quality check and outlier analysis and missing observation analysis.

3.1.1 Data quality check and outlier analysis

The first stage of the outlier detection is essentially a data quality check that compares data to known limits for each atmospheric condition (Collett and Lewis, 1976).

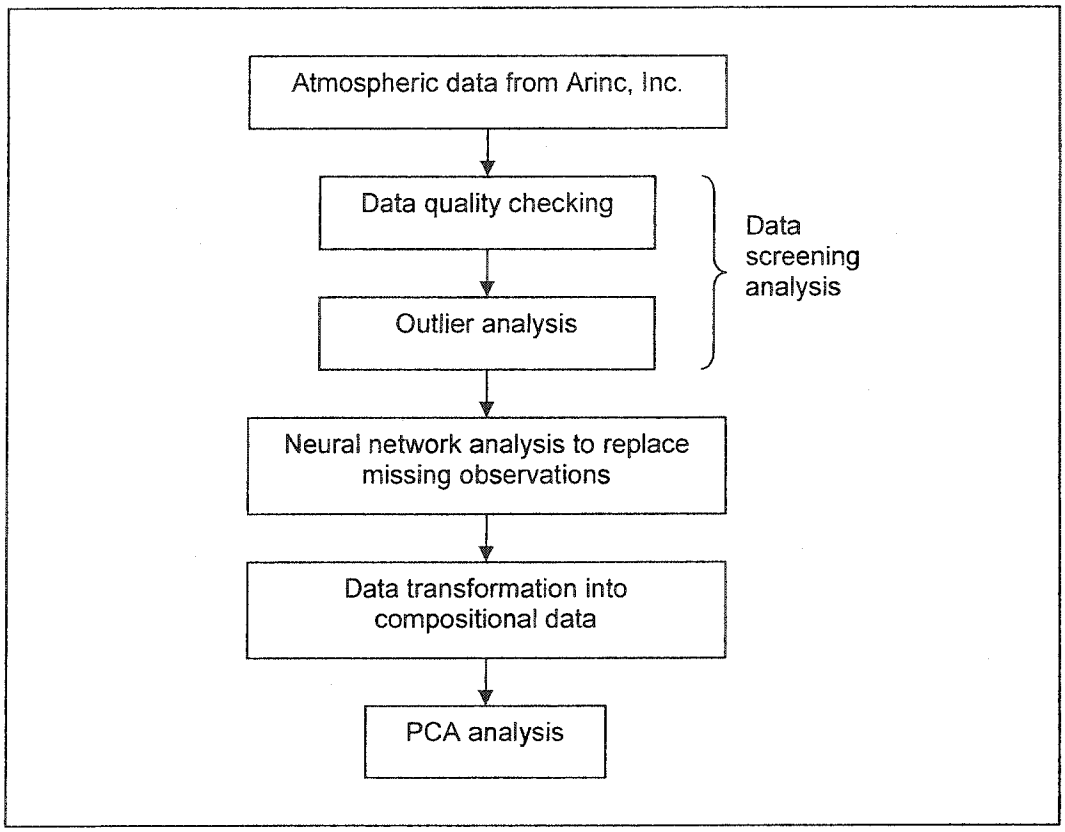


Figure 3.1: Overall methodology for obtaining corrosion severity ranking

For example, the maximum value of relative humidity (RH) must not exceed 100% while the maximum and minimum values of pH must be not higher than 14 and lower than 0, respectively. Since data quality checks are not sufficient for detecting outliers, it is necessary to perform a second pass through the data to identify outliers.

The flowchart of Figure 3.2 used the methodology in this research to detect outliers in the atmospheric data. Recall that the data is recorded as a time series where outlier analysis begins with examining the first two consecutive observations (x_t, x_{t-1}) in relationship to upper and lower bounds (as explained in Chapter 2). If the absolute value of x_t - x_{t-1} exceeds c, x_t is considered an outlier and is then removed from the data set. This procedure continues until the last value of the data set has been analyzed.

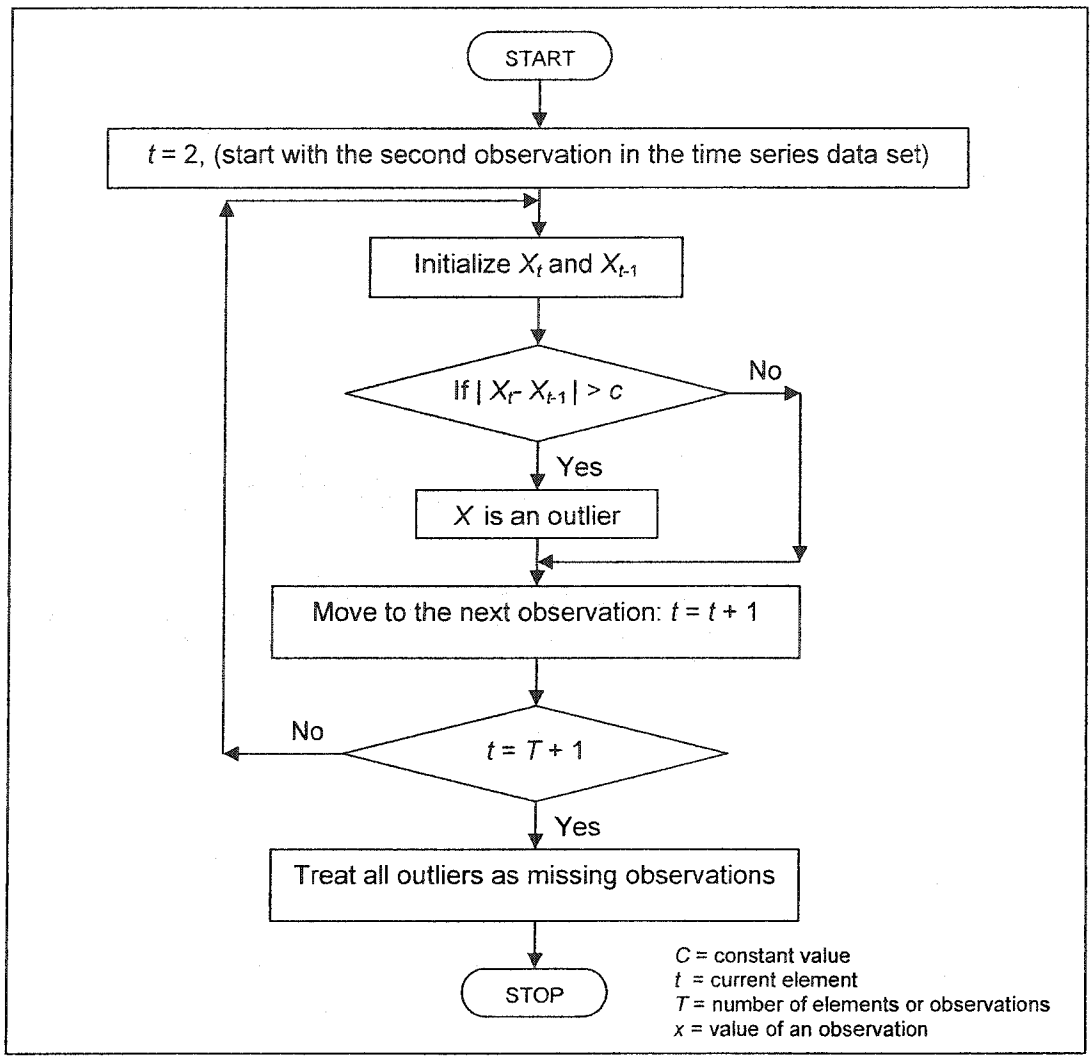


Figure 3.2: A flow chart for detecting outliers

3.1.2 Missing observation analysis: Neural Network Analysis

Missing observations in a data set can increase the chances of obtaining unreliable results, which are not meaningful. Missing observations also affect the interpretations of statistical analysis. If a time series predictive model (e.g., AR, MA, ARMA, or ARIMA) is fitted to a time series data set with missing observations, the model parameters might be biased and thus, using the time series for forecasting analysis will be inaccurate. Based on Robinson (1979), if the locations of the outliers in a data set are known, one can

treat them as missing observations and a procedure can be implemented to replace the missing observations. Neural networks are used to replace missing observations resulting from the data quality check, the outlier analysis and the data gaps caused by faulty equipment or transmittal errors.

Time series data sets obtained from Arinc, Inc. of atmospheric conditions consist of air temperature, relative humidity, rain pH, rainfall, time of wetness, and the coupon's surface temperature. These time series data sets have many gaps of missing observations which are not easy to replace by the techniques described in the literature review due to the restrictive assumptions of those techniques (e.g., data acquisition from one or more adjacent sites is not possible). Section 2.2.2 gives a more detailed description of neural networks and provides a literature review in time series forecasting.

The method to form a neural network structure on training and forecasting the atmospheric condition data sets is as follows:

Let $\mathbf{x} = [x_1, x_2, ..., x_{a-1}, x_a, x_{a+1}, ..., x_{m-1}, x_m]'$ denote a vector of time series data with m observations, which is then transformed to a new matrix $\mathbf{Z}$ with:

$$\mathbf{Z} = \left[\begin{array}{ccccc|c} x_1 & x_2 & \cdots & x_{q-1} & x_q & x_{q+1} \\ x_2 & x_3 & \cdots & x_q & x_{q+1} & x_{q+2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{a-1} & x_a & \cdots & x_{a+q-3} & x_{a+q-2} & x_{a+q-1} \\ x_a & x_{a+1} & \cdots & x_{a+q-2} & x_{a+q-1} & x_{a+q} \\ \hline \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ x_{m-q-1} & x_{m-q} & \cdots & x_{m-3} & x_{m-2} & x_{m-1} \\ x_{m-q} & x_{m-q+1} & \cdots & x_{m-2} & x_{m-1} & x_m \end{array}\right], \tag{3.1}$$

where q denotes the number of time lags and a represents the number of data for training. The number of patterns of this data set is equal to m-q, the number of rows of matrix $\mathbf{Z}$.

Each value in the last column of the matrix **Z** represents a target output corresponding to each pattern. Feed-forward networks with the back propagation (BP) algorithm work by propagating the patterns of time lags (the first q columns) through the network in a forward direction. As described in Section 2.2.2, neural networks obtain knowledge of a process by training with the first q columns and the last column data of matrix **Z**. Each neuron receives information from several input data sources by summing the input data with connection weights and using the hyperbolic tangent sigmoid transfer function as the activation (or transfer) function to deliver an output. Similarly, each neuron of the hidden layers also receives information from the outputs of all input layer nodes. In the output layer, a neuron uses the hyperbolic tangent sigmoid transfer function as the activation function to produce the final output in output layer. The training outputs from the networks and the target outputs from the last column data of matrix **Z** are used to calculate a global error. A global error function is minimized with the gradient descent convergent criterion. The training process is maintained on an epoch-by-epoch basis until the weight of each connection in the network stabilizes and the global error function converges to some minimum value based on stopping criteria. After training the networks with the first a patterns, the number of patterns, m-q-a, is used to test the performance of the networks. The missing observations are predicted by using this methodology.

To demonstrate the methodology for using a neural network structure on training and forecasting the atmospheric condition data sets, a vector of time series data set with 15 observations of air temperature for Hickam AFB is defined by:

$\mathbf{x}$ = [84.3, 83.9, 83.9, 82.9, 84.9, 84.9, 87.6, 83.9, 79.9, 77.2, 75.2, 75.6, 74.2, 74.2, 73.9].

The vector $\mathbf{x}$ is transformed to a matrix **Z** as defined in Equation 3.1:

$$\mathbf{Z} = \left[\begin{array}{cccccc} 84.3 & 83.9 & 83.9 & 82.9 & 84.9 & 84.9 \\ 83.9 & 83.9 & 82.9 & 84.9 & 84.9 & 87.6 \\ 83.9 & 82.9 & 84.9 & 84.9 & 87.6 & 83.9 \\ 82.9 & 84.9 & 84.9 & 87.6 & 83.9 & 79.9 \\ 84.9 & 84.9 & 87.6 & 83.9 & 79.9 & 77.2 \\ 84.9 & 87.6 & 83.9 & 79.9 & 77.2 & 75.2 \\ \hline 87.6 & 83.9 & 79.9 & 77.2 & 75.2 & 75.6 \\ 83.9 & 79.9 & 77.2 & 75.2 & 75.6 & 74.2 \\ 79.9 & 77.2 & 75.2 & 75.6 & 74.2 & 74.2 \\ 77.2 & 75.2 & 75.6 & 74.2 & 74.2 & 73.9 \end{array}\right].$$

The matrix **Z** is formed from the number of observations (i.e., $m = 15$) and the number of time lags (i.e., $q = 5$). The number of rows of the matrix **Z** is 10 (i.e., m-q) and the number of columns of the matrix **Z** is 6 (i.e., q+1). A value 84.9 in the last column of the first row represents a target output corresponding to the first pattern (row). The first six patterns (rows) of the matrix **Z** are used for training by feed-forward networks with the BP algorithm. The training outputs from the networks (6 values) and the target outputs from the last columns of the matrix **Z** (i.e., 84.9, 87.6, 83.9, 79.9, 77.2, 75.2) corresponding to each patterns of the matrix **Z** are used to calculate a global error. The training process continues until the network stabilizes and the global error function converges to a minimum value based on stopping criteria. After training the networks with the first 6 patterns, the remaining patterns (i.e., 4 patterns) are used to test the performance of the networks. Hence, the missing observations of the time series data set are forecasted by using this methodology. Note that Matlab® neural network toolbox (2000) is used for filling missing observations of the atmospheric conditions for all operational air bases.

3.2 Dew Point Temperature

Dew point or saturation temperature is the temperature at which a given mixture of water vapor and air is saturated (Perry and Green, 1997). Dew point temperature plays an important role in atmospheric corrosion. If the aircraft's surface temperature is within the approximate dew point temperature ±4°F, corrosion tends to occur (Howard et al., 1999). For example, given a dew point temperature of 76.1°F and the aircraft's surface temperature of 79.3°F, the approximate dew point temperature is estimated as 76.1±4°F, or a range of 72.1°F to 80.1°F. Thus, aircraft's surface temperature (i.e., 79.3°F) is within the approximate dew point temperature ±4°F (i.e, 72.1°F and 80.1°F). This recording indicates the conditions for corrosion growth and can be counted as a 30-minute interval exceeding or falling within a threshold for promoting corrosion growth. Consequently, dew point temperature can also be used with the Arinc, Inc. data when performing corrosion severity ranking analysis by PCA.

Since dew point temperature was not recorded by Arinc, Inc., the dew point temperature can be estimated as a function of air temperature and RH where the relationship between air temperature and RH is provided through the Antoine equation (Perry and Green, 1997). Given an air temperature, T (°C), and the RH level recorded at T, the Antoine equation determines the saturation pressure, P_{sat} (mmHg), as:

$$\log P_{sat} = A - \frac{B}{T + C}, \tag{3.2}$$

where A, B, and C are the Antoine equation constants. For water, the constants of A, B, and C are 8.10765, 1750.286, and 235.0, respectively (Dean, 1979). Since the Arinc, Inc.

data is collected from an environment that is not "controlled", water is always present in the atmosphere. That is, the aircraft were not placed in hangers that had relative humidity kept to a low level and thus, moisture in the air was always present. Air pressure contained moisture, P_{H_2O}, is a function of P_{sat} and defined as:

$$P_{H_2O} = P_{sat} \frac{\text{RH}}{100}. \tag{3.3}$$

By converting Equation 3.2, the dew point temperature, T_{dp}, can be obtained by

$$T_{dp} = \frac{B}{A - \log P_{H_2O}} - C. \tag{3.4}$$

For example, given an air temperature of 20°C an RH of 80%, and the Antoine equation constants for water, Equation 3.2 can be used to obtain P_{sat}, (or 17.53006 mmHg). Then, the air pressure contained moisture, P_{H_2O}, can be estimated from Equation 3.3, which is equal to 14.02405 mmHg. Hence Equation 3.4 can estimate the dew point temperature, which is equal to 16.45°C or 61.61°F.

This procedure is used to obtain the dew point temperature as one of the atmospheric conditions used for performing corrosion severity ranking.

3.3 Correlation Analysis

The interrelationships among the atmospheric conditions are explored because a large amount of information about the joint behavior of the data can play an important role for corrosion severity ranking analysis.

To illustrate the interrelationships among p variables, a (p-1 × p-1) matrix of scatter plots can be a useful way to represent the identified correlations between various variables of multivariate data on a single two-dimensional display. In this study, p is equal to eight variables representing RH, air temperature, dew point temperature, rainfall, pH, TOW1, TOW2, and surface temperature.

According to the matrices of scatter plots shown in Figures 3.3-3.8, the plotted relationships between air temperature and dew point temperature of the six operational air bases illustrate high positive correlations and the plotted relationships between RH and dew point temperature indicate positive correlations. This confirms that the dew point temperature is a function of the RH and the air temperature as described before. The relationships between RH and either TOW1 or TOW2 at all bases show that the average RH must be greater than 40% for TOW1 and TOW2 to be triggered. This implies that as RH increases, TOW1 and TOW2 increase. In addition, the relationships between rainfall and RH also indicate that the average RH must be greater than 30% for rainfall to be recorded at Seymour Johnson AFB, greater than 40% for rainfall to be recorded at Hickam AFB, greater than 50% for rainfall to be recorded at Macdill AFB, and greater than 60% for rainfall to be recorded at the remaining AFBs. On average, TOW2 is triggered at higher RH levels than TOW1 because TOW1 is geared towards dew point readings while TOW2 is geared towards rain conditions. Clearly, the relationship between TOW1 and TOW2 at Hickam AFB, Kadena AB, and RAF Mildenhall, seems to be a highly positively correlated. The correlations of TOW1 and TOW2 of Macdill AFB, Pease ANGB, and Seymour Johnson AFB are not clearly positive or negative.

Focusing on the plots of air temperature and surface temperature, it is apparent that the two variables seem to have positive correlation for Hickam AFB, Macdill AFB,

RAF Mildenhall, Pease ANGB, and Seymour Johnson AFB. Moreover, the plots of dew point temperature and surface temperature for Hickam AFB, Macdill AFB, RAF Mildenhall, Pease ANGB, and Seymour Johnson AFB also indicate positive correlation. On the other hand, the plots of air temperature and TOW1 for Pease ANGB and Seymour Johnson AFB tend to be negatively correlated. Notice that it is not easy to indicate which scatter plot is positive, negative, or without relationship if the plot is randomly scattered in the single two-dimensional display. However, correlation matrices can be used to specify the values of coefficients among variables as shown in Tables 3.1-3.6. As expected the results of Tables 3.1-3.6 reveal that the matrices of scatter plots and the correlation matrices show that the atmospheric variables seem to be correlated with each other. Hence it is necessary to transform the correlated atmospheric conditions into uncorrelated atmospheric conditions in order to perform corrosion severity ranking analysis.

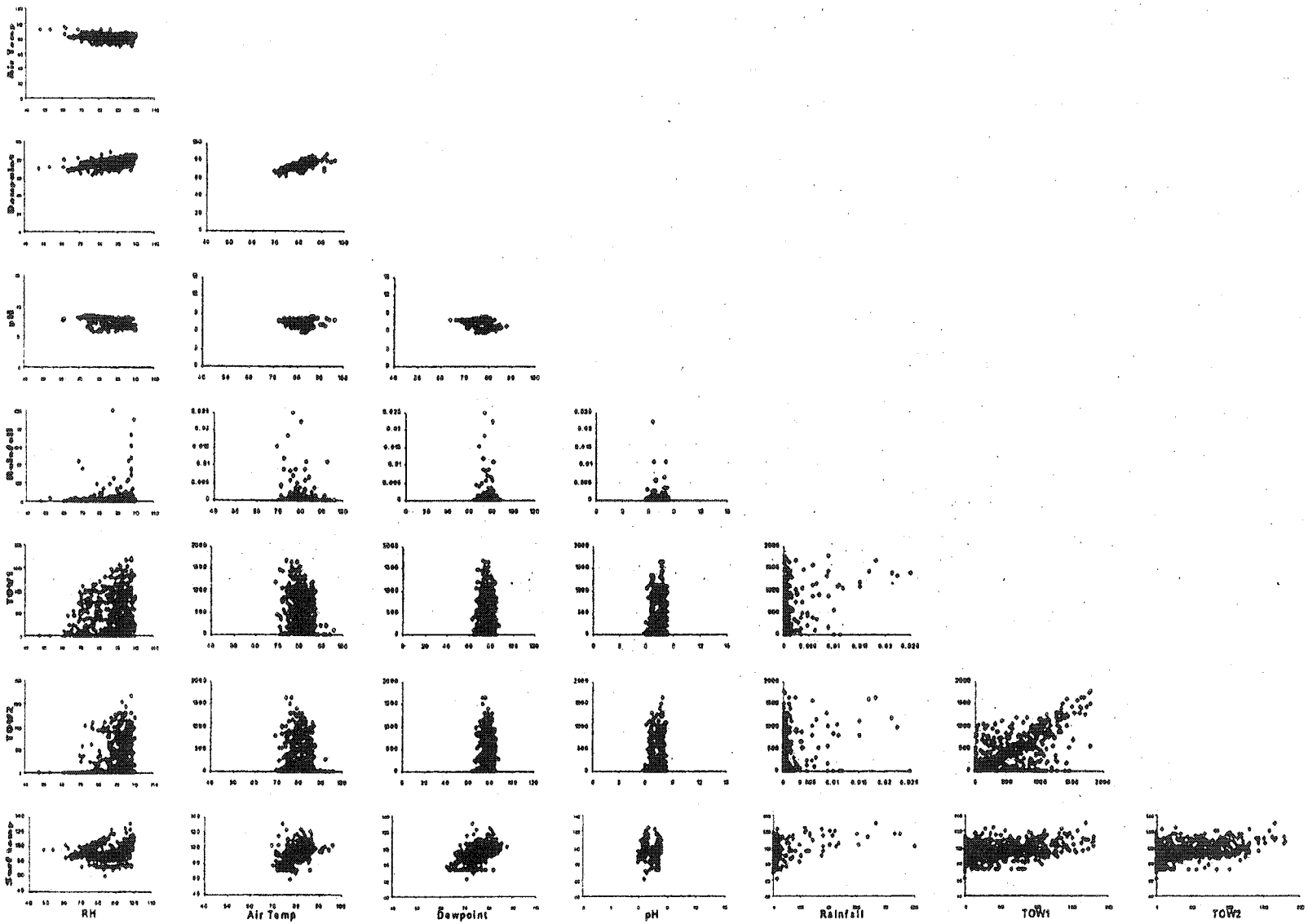


Figure 3.3: A matrix of scatter plots of Hickam AFB

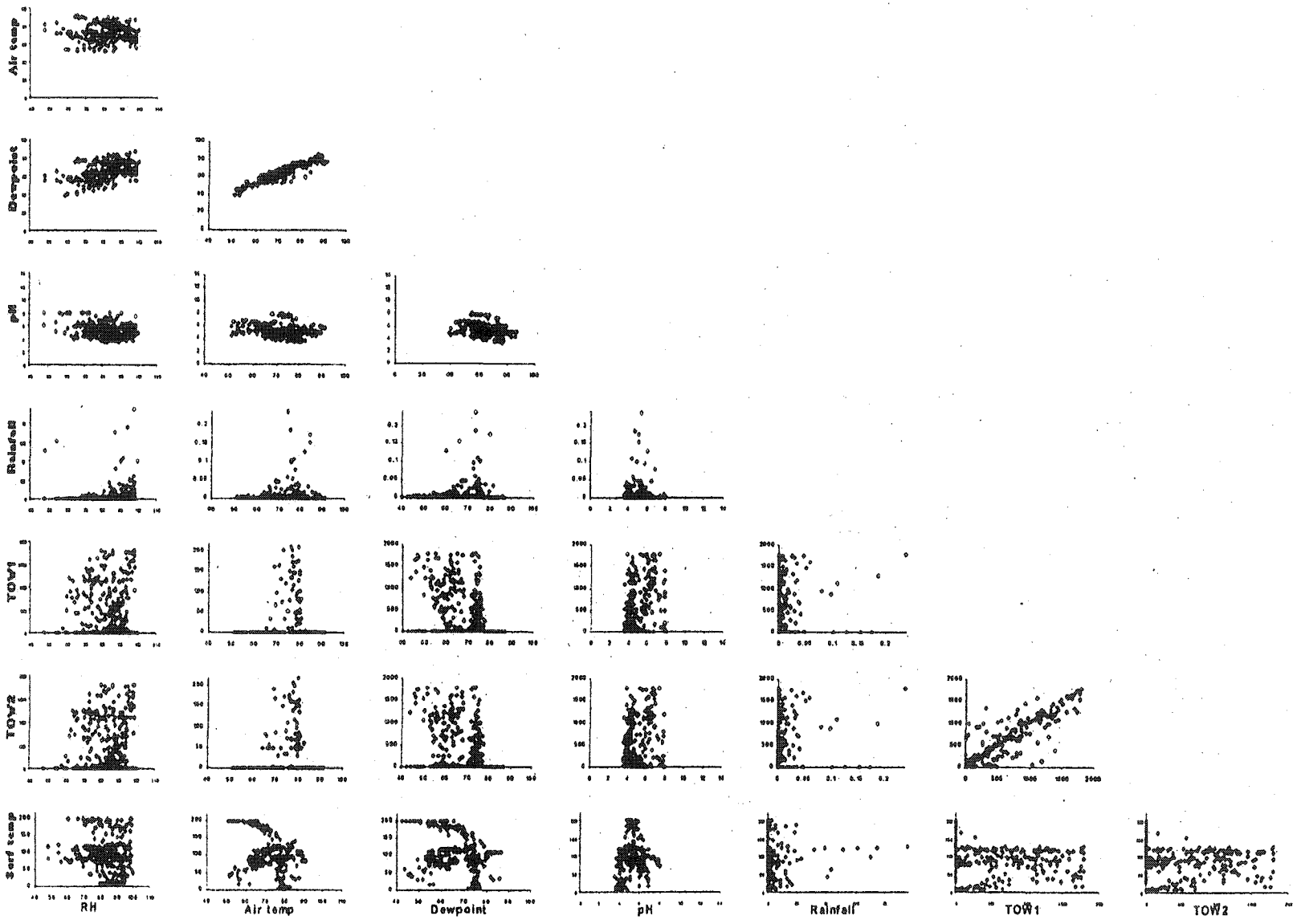


Figure 3.4: A matrix of scatter plots of Kadena AB

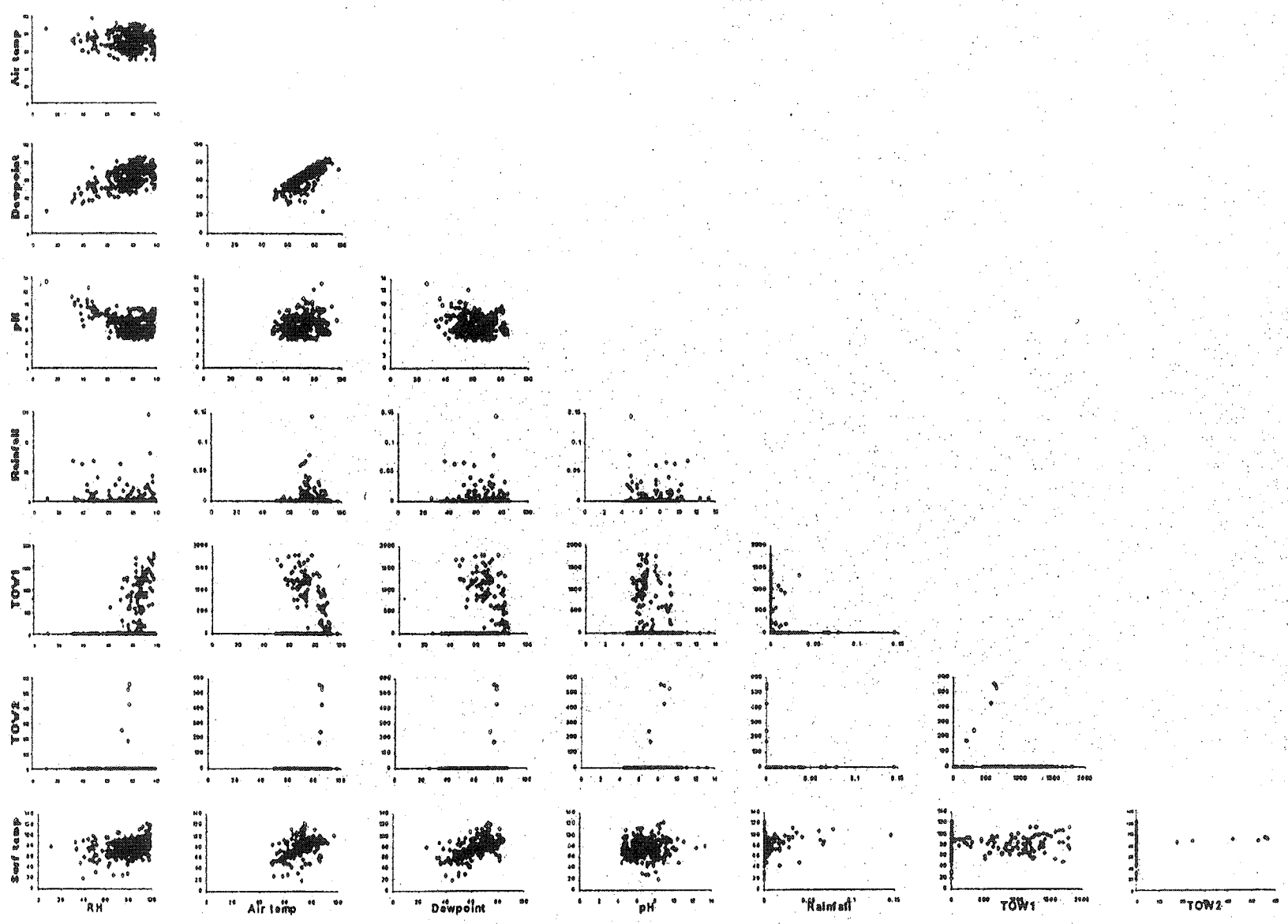


Figure 3.5: A matrix of scatter plots of Macdill AFB

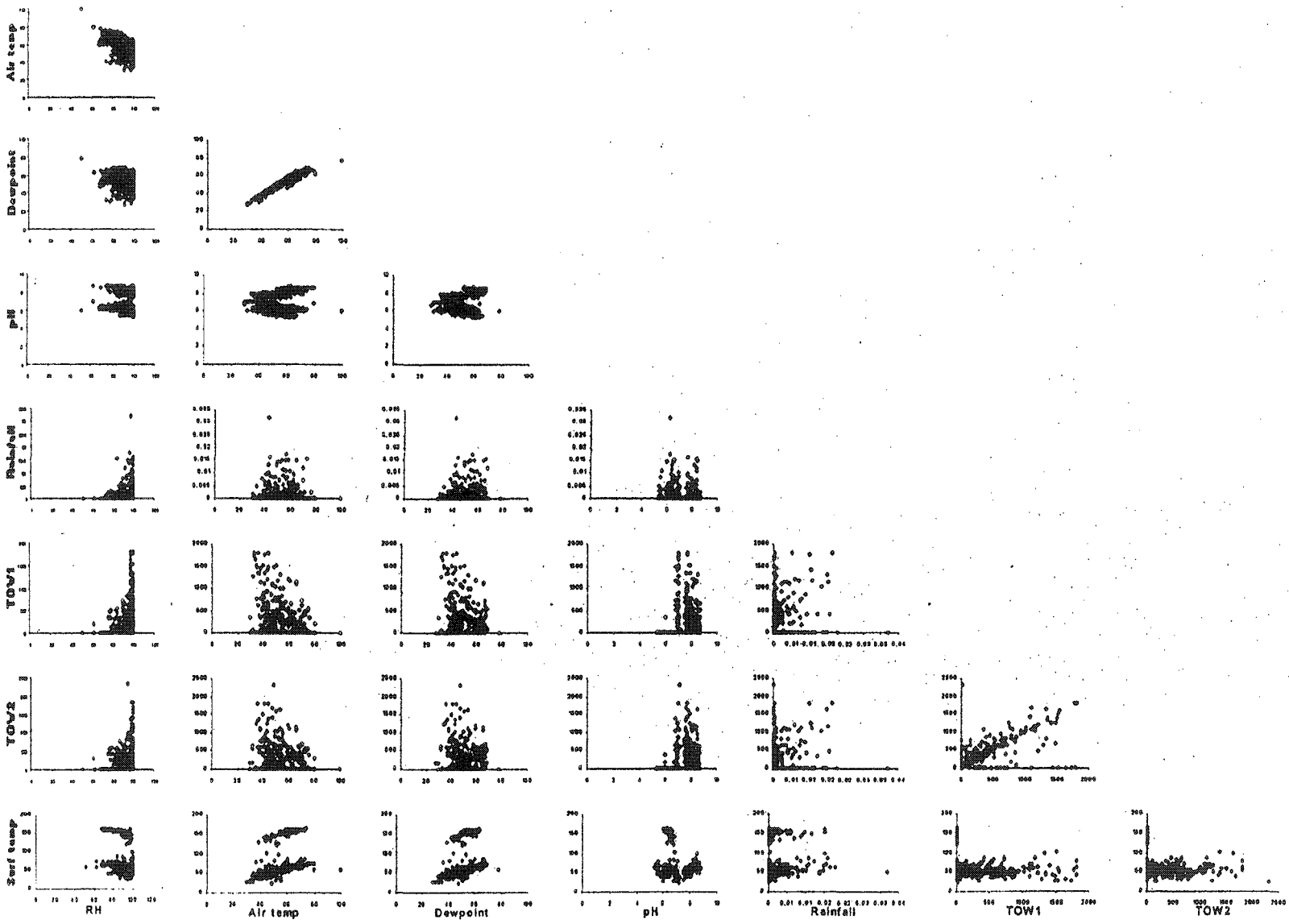


Figure 3.6: A matrix of scatter plots of RAF Mildenhall

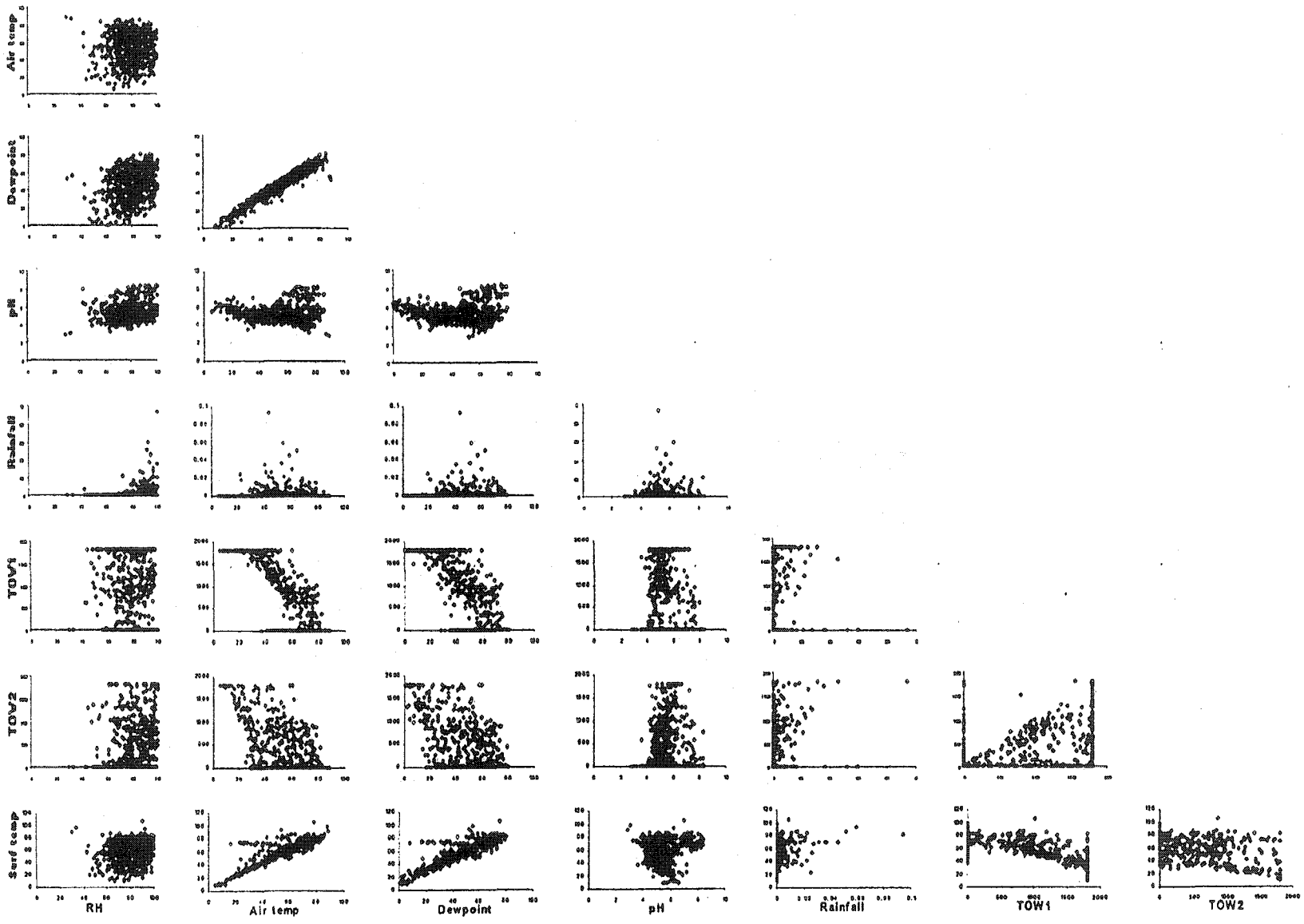


Figure 3.7: A matrix of scatter plots of Pease ANGB

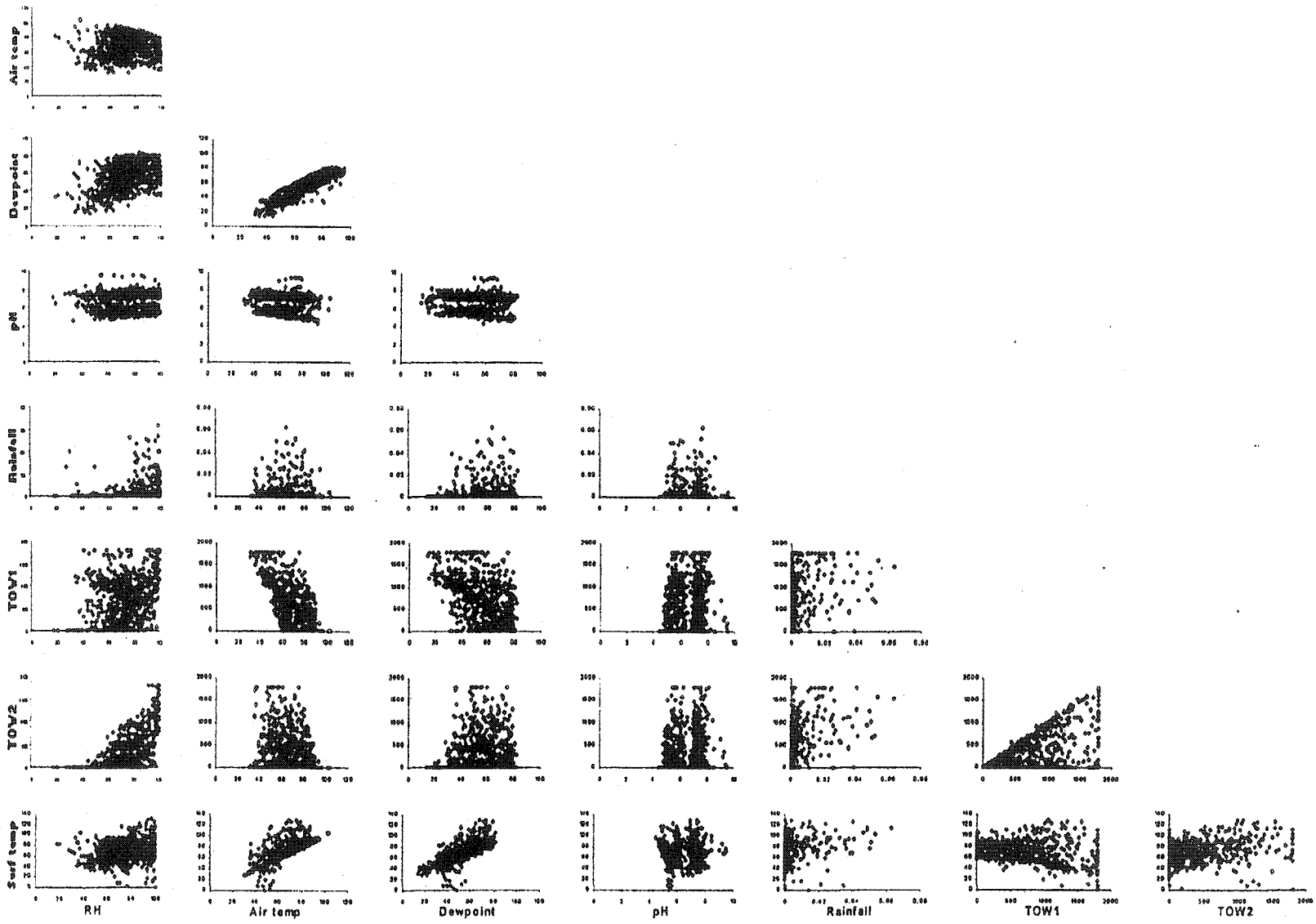


Figure 3.8: A matrix of scatter plots of Seymour Johnson AFB

Table 3.1: Correlation matrix for atmospheric conditions at Hickam AFB

	RH	AT	DP	pH	RF	TOW1	TOW2	ST
RH	1.0000							
AT	-0.6907	1.0000						
DP	0.1017	0.6460	1.0000					
pH	-0.2440	-0.1136	-0.4093	1.0000				
RF	0.0392	-0.0355	-0.0095	-0.0069	1.0000			
TOW1	0.3347	-0.3412	-0.1244	0.0783	0.0635	1.0000		
TOW2	0.3622	-0.3032	-0.0496	-0.0629	0.0654	0.7203	1.0000	
ST	0.0861	0.2414	0.4076	-0.1540	0.0677	0.1963	0.2063	1.0000

Note: RH: relative humidity, AT: air temperature, DP: dew point temperature, pH: rain pH, RF: rainfall, TOW1: time-of-wetness used to detect light dew, TOW2: time-of-wetness while used to detect rain and heavier liquid condensation, and ST: surface temperature measured on the material

Table 3.2: Correlation matrix for atmospheric conditions at Kadena AB

	RH	AT	DP	pH	RF	TOW1	TOW2	ST
RH	1.0000							
AT	-0.1123	1.0000						
DP	0.4140	0.8560	1.0000					
pH	-0.2540	-0.2387	-0.3614	1.0000				
RF	0.0915	-0.0015	0.0410	-0.0006	1.0000			
TOW1	0.2822	-0.3677	-0.1988	0.1094	0.1015	1.0000		
TOW2	0.2761	-0.3498	-0.1841	0.0157	0.1024	0.8501	1.0000	
ST	-0.1102	-0.3204	-0.3555	0.1761	0.0346	-0.1174	-0.0943	1.0000

Table 3.3: Correlation matrix for atmospheric conditions at Macdill AFB

	RH	AT	DP	pH	RF	TOW1	TOW2	ST
RH	1.0000							
AT	-0.3643	1.0000						
DP	0.3695	0.7181	1.0000					
pH	-0.2947	0.1439	-0.1019	1.0000				
RF	-0.0077	-0.0015	-0.0107	0.0329	1.0000			
TOW1	0.3433	-0.2054	0.0209	-0.0620	-0.0247	1.0000		
TOW2	0.0590	0.0115	0.0497	0.0472	-0.0053	0.1494	1.0000	
ST	-0.0594	0.5846	0.5185	0.0332	0.1139	0.0381	0.0502	1.0000

Table 3.4: Correlation matrix for atmospheric conditions at RAF Mildenhall

	RH	AT	DP	pH	RF	TOW1	TOW2	ST
RH	1.0000							
AT	-0.5937	1.0000						
DP	-0.2265	0.9164	1.0000					
pH	0.0662	0.0563	0.0935	1.0000				
RF	0.0692	-0.0297	-0.0033	0.0089	1.0000			
TOW1	0.2042	-0.2303	-0.1872	0.2574	0.1114	1.0000		
TOW2	0.2083	-0.1895	-0.1359	0.3058	0.0918	0.7665	1.0000	
ST	-0.3401	0.4379	0.3659	-0.2540	0.0307	-0.1570	-0.1513	1.0000

Note: RH: relative humidity, AT: air temperature, DP: dew point temperature, pH: rain pH, RF: rainfall, TOW1: time-of-wetness used to detect light dew, TOW2: time-of-wetness while used to detect rain and heavier liquid condensation, and ST: surface temperature measured on the material

Table 3.5: Correlation matrix for atmospheric conditions at Pease ANGB

	RH	AT	DP	pH	RF	TOW1	TOW2	ST
RH	1.0000							
AT	-0.2283	1.0000						
DP	0.1561	0.9237	1.0000					
pH	0.1656	0.0041	0.0713	1.0000				
RF	0.1031	-0.0026	0.0351	0.0179	1.0000			
TOW1	0.2014	-0.7248	-0.6589	-0.0528	0.0246	1.0000		
TOW2	0.2849	-0.4473	-0.3535	0.1279	0.1452	0.4941	1.0000	
ST	-0.1009	0.8065	0.7744	0.0093	0.1046	-0.5166	-0.2422	1.0000

Table 3.6: Correlation matrix for atmospheric conditions at Seymour Johnson AFB

	RH	AT	DP	pH	RF	TOW1	TOW2	ST
RH	1.0000							
AT	-0.3699	1.0000						
DP	0.3616	0.7205	1.0000					
pH	0.0663	-0.0306	0.0191	1.0000				
RF	0.1108	-0.0277	0.0445	-0.0201	1.0000			
TOW1	0.4709	-0.6213	-0.2999	0.0056	0.1335	1.0000		
TOW2	0.4510	-0.1992	0.0916	-0.0007	0.2343	0.5971	1.0000	
ST	-0.1238	0.7082	0.5900	-0.0139	0.1667	-0.2501	0.2272	1.0000

3.4 Corrosion Severity Ranking Using Principal Component Analysis

Principal component analysis (PCA) is used to transform correlated variables to uncorrelated variables (Jackson, 1991). This technique is also used to reduce the number of variables without losing much of the information (Afifi and Clark, 1990). With the atmospheric condition data sets, PCA can be applied for corrosion severity ranking of the six air force bases. Recall that one objective of this research is to rank corrosion severity by locations based on the atmospheric condition data sets captured from the six operational air force bases.

The proposed methodology for corrosion severity ranking is to transform the original data set into a compositional data set based on the percentage of 30-minute intervals that an atmospheric variable has met a condition conducive to corrosion growth. Table 3.7 summarizes the conditions for each atmospheric variable. The reason that the percentage of 30-minute intervals is used for corrosion severity ranking analysis of the six air force bases is that atmospheric corrosion depends on the length of time that moisture is present on the metal's surface. In addition, utilizing the percentage of time allows each atmospheric variable to be measured against the same scale.

Table 3.7: Atmospheric conditions conducive to corrosion growth

Atmospheric variable	Atmospheric condition promoting corrosion growth
Relative humidity (RH)	RH > 60%
Air temperature (AT)	AT > 60°F
Dew point temperature (DT)	DT > 46°F
Rainfall (RF)	RF > 0 inch
Rain pH pH)	pH < 4 or pH > 8.5
Time-of-wetness1 (TOW1)	TOW1 > 0 second
Time-of-wetness2 (TOW2)	TOW2 > 0 second
Surface temperature (ST)	ST $\in$ DT $\pm$ 4°F

Since this data transformation places all data against a common scale, PCA analysis can be used to perform corrosion severity ranking. For the PCA analysis, the percentage of 30-minute intervals for each atmospheric condition now represents a variable, whereas the six air force base locations represent the observations. Recall from the literature review, PCA can be applied to compositional data sets such as vectors of percentages or proportions. Figure 3.9 illustrates the procedure to obtain the compositional data set from the original data set. First, the number of 30-minute intervals that an atmospheric condition is conducive to corrosion growth is transformed into a ratio as follows:

$$r_{ij} = \frac{t_{ij} \times T_j}{o_i}, \; i = 1,2,...,n \;\; j = 1,2,...,p, \tag{3.5}$$

where r_{ij} denotes the ratio of the number of 30-minute intervals when an atmospheric condition is conducive to corrosion growth to the number of 30-minute intervals of the ith site and the jth atmospheric condition, t_{ij} represents the number of 30-minute intervals when an atmospheric condition is conducive to corrosion growth, T_j denotes total number of 30-minute intervals when an atmospheric condition is conducive to corrosion growth of each atmospheric condition, o_i denotes the number of 30-minute intervals collected at ith site, n is the number of observations (i.e., air force bases), and p is the number of variables (i.e., the atmospheric conditions).

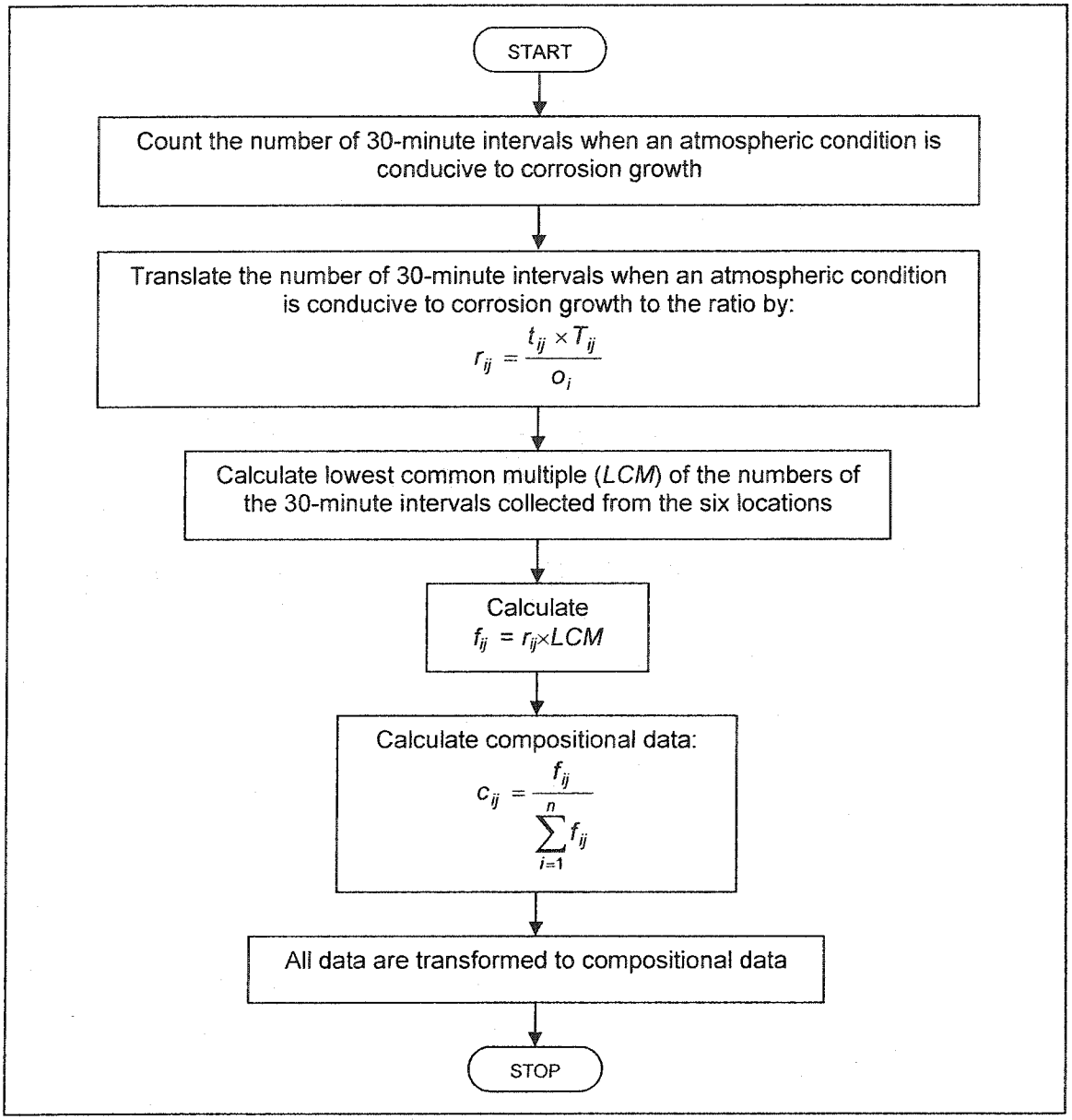


Figure 3.9: Procedure for transforming the original data set into a compositional data set

To calculate the compositional data, it is necessary to determine the lowest common multiple (*LCM*) of the number of 30-minute intervals between the bases. The *LCM* will allow PCA to rank bases (observations) by their atmospheric conditions (variables). Note that *LCM* is the smallest multiple that is exactly divisible by every member of a set of numbers. For example, the *LCM* of 12 and 18 is 36. Then, a value (f_{ij}) is calculated as follows:

$$f_{ij} = r_{ij} \times LCM. \tag{3.6}$$

Thus, compositional data can be estimated as follows:

$$c_{ij} = \frac{f_{ij}}{\sum_{i=1}^{n} f_{ij}}, \quad j = 1,2,...,p, \tag{3.7}$$

where c_{ij} denotes a compositional data of the *i*th site and the *j*th atmospheric condition. Note that $\sum_{i=1}^{n} c_{ij}$, for $j = 1,2,...,p$ should be equal to one (i.e., the definition of compositional data). After obtaining the compositional data, the method of PCA is then used to analyze corrosion severity ranking.

3.5 Corrosion Growth Modeling Development

The foundation for the corrosion growth models developed in this research is based on the known behavior of corrosion growth. Mikhailovsky (1982) states that when the gaseous oxygen interacts with the metal, an oxide is formed on the metal surface. The oxide forms the surface layer, which protects the metal from further oxidation. If the surface layer reaches a certain thickness, the oxide ceases to grow and the metal

passivates. Consequently, while corrosion growth is free initially, it reaches an ultimate limiting value. That is, it cannot corrode more than what amount of metal is initially present or it is stopped by the oxidation process itself. So any model proposed for predicting corrosion growth must uphold this phenomena.

To satisfy the growth phenomena of corrosion, the following three models are modified in order to develop a more accurate predictive model of corrosion growth and are explained in the following sections:

- the Gompertz growth model and the logistic growth model or GL model
- the Gompertz growth model and the confined exponential growth model or GC model
- the logistic growth model and the confined exponential growth model or CL model

Note that although the Gompertz, the logistic, and the confined exponential models are typical types of growth models that have been applied in many fields such as biology, botany, forestry, zoology, and ecology (Banks, 1994), these modifications have not been explored. The three modifications are proposed since the existing growth models (the Gompertz and the logistic) have not satisfied the first growth phenomena of corrosion (i.e., corrosion growth is free initial). Recall Equation 2.42 (i.e., the Gompertz model), the initial condition when $t = 0$ gives:

$$P = P_* \exp\left(-\frac{a_0}{k}\exp(-k(0))\right)$$

$$P = P_* \exp\left(-\frac{a_0}{k}\right).$$

This shows that P is not zero when $t = 0$. Recall Equation 2.44 (i.e., the logistic model), the initial condition when $t = 0$ gives:

$$P = P_*\left[1+\left(\frac{P_*}{P_0}-1\right)\exp\left(-a(0)\right)\right]^{-1}$$

$$P = P_*\left(1+\left(\frac{P_*}{P_0}-1\right)\right)^{-1}$$

$$P = P_0.$$

This also shows that P is not zero when $t = 0$. However, modifying these models may address this issue.

3.5.1 The GL model

Since the rates of change of the Gompertz and the logistic models are defined by:

$$\frac{dP}{dt} = ae^{-kt}P$$

and

$$\frac{dP}{dt} = aP\left(1-\frac{P}{P_*}\right),$$

respectively, the rate of change of corrosion growth of a developed model can be defined as:

$$\frac{dP}{dt} = ae^{-kt}\left(1-\frac{P}{P_*}\right), \tag{3.8}$$

where P and P_* are corrosion growth and the ultimate limiting value of corrosion growth, respectively, and a and k are the growth coefficient and the decay coefficient. The solution to Equation 3.8 is

$$\int \frac{dP}{P_* - P} = \int \frac{ae^{-kt}}{P_*} dt$$

$$-\ln(P_* - P) = -\frac{ae^{-kt}}{kP_*} + c,$$

where c is the constant of integration. The initial condition $P(0) = P_0 = 0$ when $t = 0$ gives:

$$c = -\ln P_* + \frac{a}{kP_*}.$$

Hence,

$$-\ln(P_* - P) = -\frac{ae^{-kt}}{kP_*} - \ln P_* + \frac{a}{kP_*} \text{ or}$$

$$\frac{P_*}{P_* - P} = \exp\left(-\frac{ae^{-kt}}{kP_*} + \frac{a}{kP_*}\right).$$

The GL corrosion growth model is then

$$P = P_* \left(1 - \frac{1}{\exp\left(\frac{a}{kP_*}\left(1 - e^{-kt}\right)\right)}\right) \tag{3.9}$$

3.5.2 The GC model

Since the rates of change of the Gompertz and the confined exponential models are defined by:

$$\frac{dP}{dt} = ae^{-kt} P$$

and

$$\frac{dP}{dt} = a(P_* - P),$$

respectively, the rate of change of corrosion growth for the developed model can be defined as:

$$\frac{dP}{dt} = a(P_* - P)e^{-kt}, \tag{3.10}$$

where P and P_* are corrosion growth and the ultimate limiting value of corrosion growth, respectively, and a and k are the growth coefficient and the decay coefficient. The solution to Equation 3.10 is

$$\int \frac{dP}{P_* - P} = \int a e^{-kt} dt$$

$$-\ln(P_* - P) = -\frac{a}{k} e^{-kt} + c,$$

where c is the constant of integration. The initial condition $P(0) = P_0 = 0$ when $t = 0$ gives:

$$c = \frac{a}{k} - \ln P_*.$$

Hence,

$$-\ln(P_* - P) = -\frac{a}{k} e^{-kt} + \frac{a}{k} - \ln P_* \text{ or}$$

$$\frac{P_*}{P_* - P} = \exp\left(\frac{a}{k}\left(1 - e^{-kt}\right)\right).$$

The GC corrosion growth model is then

$$P = P_* - \frac{P_*}{\exp\left(\frac{a}{k}\left(1 - e^{-kt}\right)\right)} \tag{3.11}$$

3.5.3 The CL model

Since the rates of change of the logistic and the confined exponential models are defined by:

$$\frac{dP}{dt} = aP\left(1 - \frac{P}{P_*}\right)$$

and

$$\frac{dP}{dt} = a(P_* - P),$$

respectively, the rate of change of corrosion growth for the developed model can be defined as:

$$\frac{dP}{dt} = a(P_* - P)\left(1 - \frac{P}{P_*}\right), \tag{3.12}$$

where P and P_* are corrosion growth and the ultimate limiting value of corrosion growth while a is the growth coefficient. The solution to this equation is

$$\frac{dP}{dt} = a\frac{(P_* - P)^2}{P_*}$$

$$\int \frac{dP}{(P_* - P)^2} = \int \frac{a}{P_*} dt$$

$$\frac{1}{P_* - P} = \frac{a}{P_*} t + c,$$

where c is the constant of integration. The initial condition $P(0) = P_0 = 0$ when $t = 0$ gives:

$$c = \frac{1}{P_*}.$$

Hence,

$$\frac{1}{P_* - P} = \frac{a}{P_*}t + \frac{1}{P_*}.$$

Thus, the CL corrosion growth model is defined as:

$$P = \frac{P_* at}{1 + at} \tag{3.13}$$

Note that the three proposed models are nonlinear models (at least one of the derivatives of the expectation function with respect to the parameters depends on at least one of the other parameters (Bates and Watts, 1988)). This can be shown as follows (for the CL model):

$$f(t, a, P^*) = \frac{P^* at}{1 + at}$$
$$\frac{\partial f}{\partial a} = \frac{P^* t}{(1 + at)^2}$$
$$\frac{\partial f}{\partial P^*} = \frac{at}{1 + at}.$$

In this research, the confined exponential growth model (i.e., Equation 2.47), the power law equation (i.e., Equation 2.34), and the three new models (i.e., Equations 3.9, 3.11, and 3.13) are statistically tested for fit. Specifically, SAS® software is used to perform nonlinear regression with an iterative estimation method (Marquardt method) for parameter estimates to fit these models to the corrosion growth data sets. A lack-of-fit test is then used to check whether or not the data indicate a nonlinear growth phenomena. Model adequacy checking techniques are used to assess the model fitting. A residual analysis test (consisting of normality and constant variance tests) is used to check

whether or not the model is adequate. If the residuals indicate a non-constant variance, weighted least squares method is employed. Models passing the model adequacy checking are compared with each other using error sum of squares criteria and the best model is then identified.

3.6 Statistical Techniques for Modeling

Since corrosion growth follows a nonlinear growth pattern, this section describes the theory regarding the least squares method and iterative methods for nonlinear parameter estimation. Furthermore, techniques of model adequacy checking for nonlinear models are also provided.

3.6.1 Theory

Generally, a nonlinear model can be expressed as:

$$Y = f(\mathbf{x}, \boldsymbol{\theta}) + \varepsilon \tag{3.14}$$

or

$$E(Y) = f(\mathbf{x}, \boldsymbol{\theta})$$

if $E(\varepsilon) = 0$ and $V(\varepsilon) = \sigma^2$ are assumed where Y is a response variable, f is the expectation function, $\mathbf{x}$ is a vector of nonlinear predictor variables defined as $(x_1, x_2, ..., x_k)'$, $\boldsymbol{\theta}$ is a vector of nonlinear parameters defined as $(\theta_1, \theta_2, ..., \theta_p)'$, and ε is an error term distributed as $\varepsilon \sim N(0, \sigma^2)$ (Draper and Smith, 1998). The errors are uncorrelated and independent.

If there are n observations, Equation 3.14 can be defined as:

$$Y_u = f(\mathbf{x}_u, \boldsymbol{\theta}) + \varepsilon_u, \quad \text{for } u=1,2,...,n, \tag{3.15}$$

where $\mathbf{x}_u$ is a vector of nonlinear predictor variables defined as $(x_{1u}, x_{2u}, ..., x_{ku})'$. The assumption of normality and independence of the errors can be written as $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \mathbf{I}\sigma^2)$, where $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)'$ and $\mathbf{0}$ is a vector of zeros and $\mathbf{I}$ is an identity matrix.

For nonlinear models, at least one of the derivatives of the expectation function with respect to the parameters depends on at least one of the parameters (Bates and Watts, 1988). To demonstrate the nonlinear growth in the proposed corrosion growth model, the nonlinear growth model can be expressed as:

$$P_u = f(t, \theta) + \varepsilon_u, \quad \text{for } u=1,2,...,n, \tag{3.16}$$

where P is corrosion growth (thickness loss) measured in millimeter and t is exposure time measured in year. The confined exponential model (see Section 2.6.3) is defined as:

$$P_u = P^*(1 - e^{-\theta t}) + \varepsilon_u, \quad \text{for } u=1,2,...,n, \tag{3.17}$$

where $P^*(1 - e^{-\theta t}) = f(t, \theta, P^*)$, P^* is the accumulated maximum growth when $t \to \infty$. Since

$$\frac{\partial f}{\partial \theta} = P^* t e^{-\theta t} \tag{3.18}$$

and

$$\frac{\partial f}{\partial P^*} = 1 - e^{-\theta t}$$

depend on the parameters θ and P^*, this model is nonlinear.

To find the normal equation for obtaining the least squares estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$, the error sum of squares for the nonlinear model and the given data is defined as:

$$S(\boldsymbol{\theta}) = \sum_{u=1}^{n} \left(Y_u - \hat{Y}_u\right)^2, \tag{3.19}$$

where $\hat{Y}_u = f(\mathbf{x}_u, \boldsymbol{\theta})$. To obtain the p normal equations, Equation 3.19 will be differentiated with respect to $\boldsymbol{\theta}$ and set equal to zero:

$$\sum_{u=1}^{n} \left\{Y_u - f\left(\mathbf{x}_u, \hat{\boldsymbol{\theta}}\right)\right\}\left[\frac{\partial f(\mathbf{x}_u, \boldsymbol{\theta})}{\partial \theta_i}\right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = 0. \tag{3.20}$$

To illustrate how Equation 3.20 can be used to solve the least squares estimate of the parameter, the corrosion growth model defined in Equation 3.17 is demonstrated. To find the normal equation for obtaining the least squares estimate $\hat{\theta}$ of θ, one needs to take the derivative with respect to θ. Hence,

$$\frac{\partial f}{\partial \theta} = P^* t e^{-\theta t}. \tag{3.21}$$

According to Equation (3.20), one obtains

$$\sum_{u=1}^{n} \left\{P_u - P^*\left(1 - e^{-\hat{\theta} t_u}\right)\right\}\left[P^* t e^{-\hat{\theta} t_u}\right] = 0$$

or

$$P^* \sum_{u=1}^{n} P_u t_u e^{-\hat{\theta} t_u} - P^{*2} \sum_{u=1}^{n} t_u e^{-\hat{\theta} t_u} + P^{*2} \sum_{u=1}^{n} t_u e^{-2\hat{\theta} t_u} = 0. \tag{3.22}$$

Clearly, the normal equation is not easy to solve although it is a simple nonlinear model with one parameter. Thus, iterative estimation methods can be employed to deal with this problem.

Several iterative estimation methods used to obtain the parameter estimates include, steepest descent, Gauss-Newton method, and Marquardt method. Each iterative estimation method requires a starting value for each parameter. To improve the fit of the curve to the data by minimizing the residual sum of squares, the nonlinear regression procedure iteratively moves along the surface of the curve by adjusting the values of the parameters. These iterations continue until improvement occurs (i.e., the residual sum of squares is minimized).

The method of steepest descent is a gradient method, which requires calculating derivatives for obtaining search directions. This iterative estimation technique starts at an arbitrary value moving along the direction of steepest descent with an arbitrary step size. Then, the derivative is calculated in the new spot and the procedure is repeated. This iterative procedure continues (in a zig-zig manner, where the new search direction is orthogonal to the previous) until convergence is achieved. This technique is emphasized in initial steps with large step sizes that make it quickly approach the minimum value of the residual sum of squares. However, there is no guarantee that this method will reach the minimum value of the residual sum of squares after performing some iterations.

The method of Gauss-Newton uses Taylor series expansion to develop a method for iterative parameter estimation. The operation of pre-multiplying the steepest descent direction by the inverse of the first derivative matrix is performed to find a suitable direction for the quadratic approximation to the function, rather than by finding a linear approximation to the function as in the steepest descent procedure. The new parameter

estimates are solved and replaced as the initial values. This iterative procedure continues until convergence is achieved. It is worth noting that this method works well when the values of parameter estimates are close to the optimum values. However, if the initial value is poorly selected, this can lead to go to the wrong direction and never converge to the minimum value of the residual sum of squares.

The method of Marquardt combines the advantages of both the Steepest Descent and Gauss-Newton methods. The method of Steepest Descent works well in initial iterations whereas the method of Gauss-Newton works well in later iterations. The method of Marquardt uses the method of Steepest Descent in the initial iterations and performs it until the residual sum of squares is no longer decreasing (Marquardt, 1963). Then, the method of Marquardt gradually switches over to the Gauss-Newton principle. This method is useful to fit many types of data to various types of equations.

Thus, the Marquardt is used to determine parameter estimates in the corrosion growth models of this study.

After obtaining the model parameter estimates, a confidence interval for a model parameter (θ) can also be calculated by:

$$\hat{\theta} \pm t_{\alpha, N-p} se\left(\hat{\theta}\right) \tag{3.23}$$

where $\hat{\theta}$ is a model parameter estimate, t_{N-p} is the appropriate value from a *t*-distribution with degrees of freedom *N-p*, *N* is the number of observations, *p* is the number of model parameters, and *se* is standard error of the parameter estimate. In addition, the upper and lower 95% confidence bounds of the predicted value ($\hat{Y}$) can be calculated from:

$$\hat{Y} \pm t_{N-p} se\left(\hat{Y}\right), \tag{3.24}$$

where $\hat{Y} = f\left(x_u, \hat{\theta}\right)$, $se\left(\hat{Y}\right) = s\sqrt{\mathbf{m}_u'\left(\mathbf{M}'\mathbf{M}\right)^{-1}\mathbf{m}_u}$, $\mathbf{m}_u$ is the vector of the first derivatives of f evaluated at the parameter estimates and x_u, and $\mathbf{M}$ is an $N \times p$ matrix of the first derivatives evaluated at $\hat{\theta}$.

A more detailed review of iterative estimation methods can be found in Bates and Watts (1988) and Ryan (1997). To check the adequacy of the model, some methods such as lack of fit test and plotting residuals are presented and will be followed in this research. They are described in the next section.

3.6.2 Lack-of-fit test for assessing the fit of the model

When there are two or more response values (repeat observations) corresponding to at least one independent variable, it is necessary to determine if the nonlinear model is appropriate (i.e., H_0: the nonlinear regression model is correct against H_1: the nonlinear regression model is not correct). The error sum of squares plays an important role for assessing the fit of the model when the data contain repeat observations. The error sum of squares consists of; 1) pure error and 2) lack of fit (Draper and Smith, 1998), where

$$\begin{aligned} &\text{Error sum of squares} = \text{pure error} + \text{lack of fit} \\ &SSE = SS_{PE} + SS_{LOF} \\ &SSE = \sum_{j=1}^{m}\sum_{i=1}^{n}\left(y_{ij} - \bar{y}_j\right)^2 + \sum_{j=1}^{m}\sum_{i=1}^{n}\left(\bar{y}_j - \hat{y}_{ij}\right)^2 \end{aligned} \tag{3.25}$$

and m is the number of distinct combinations of predictor values (or the number of distinct values of the predictor if there is only one predictor), while n is number of repeat

observations in each design point of a predictor variable. The test for the lack of fit can be performed by an F-test statistic

$$F_0 = \frac{SS_{LOF}/(m-p)}{SS_{PE}/(N-m)} = \frac{MS_{LOF}}{MS_{PE}}, \tag{3.26}$$

where p is the number of parameters in the nonlinear model, N is the sample size, MS_{LOF} is mean square of lack of fit, and MS_{PE} is mean square of pure error. Large values of F_0 indicate that more error is coming from lack of fit than from random variability. This implies that if H_0 is rejected, the nonlinear model is not appropriate.

3.6.3 Plotting residuals

Based on the assumption that corrosion models are nonlinear, the errors are assumed to be normally and independently distributed with mean zero and constant but unknown variance σ^2. The examination of residuals can also be investigated to check model adequacy. The diagnostic checking can be done easily by graphical analysis of residuals. If the model is adequate, the residuals should be structureless (i.e., they should contain no obvious patterns.).

Graphically, the plot of standardized residuals against fitted values and the plot of standardized residuals against the nonlinear predictor variables are the tools used to perform model adequacy checking. The residuals are the differences between the observations and the corresponding predictive values, while the standardized normal scores are the residuals divided by their standard deviation. The mean and the standard deviation of the standardized normal scores should be zero and one, respectively. This research will use the standardized normal scores for the normal probability plot.

Normal probability plots are other tools used to check the model adequacy (normality of the errors). Normal probability plots are graphical methods for determining whether observed data conform to a hypothesized model based on a subjective visual examination of the data. Normal probability plots include quantile-quantile (Q-Q) plot and a probability (P-P) plot. A Q-Q plot is a graph between the standardized residuals and the normal quantile whereas a P-P plot is a graph between the distribution of standardized residuals and the sample probability. If the hypothesized model adequately accounts for the observed data, the plotted points will fall approximately along a straight line.

To test the assumption of independence of the errors, a plot of standardized residuals against a predictor variable can be employed. If the plot shows suspicious behavior, such as runs of residuals of the same sign, the assumption of independence of the errors is not appropriate.

To check homogeneity of variance, a plot of standardized residuals against the fitted values is useful. If the plot indicates suspicious behavior, such as a wedge-shaped pattern or megaphone pattern, the assumption of constant variance is not appropriate. When the data includes replications at some or all of the design points, one can plot the variances or standard deviations for the replicated responses against the averages. If there is a relationship, the variance is not constant. Then, the method of weighted least squares should be employed to alleviate the problem of heterogeneity.

3.6.4 Weighted least squares

A technique called weighted least squares can be used to correct for unequal variances and still maintain the regression relationship between the dependent and

independent variables (Carroll and Ruppert, 1988; Ryan, 1997). This subsection provides the method of weighted least squares used to eliminate heterogeneity variances.

Based on Equation 3.15 (with one dependent variable and one parameter, θ), usual nonlinear regression modeling makes the following basic assumptions:

$$E(Y_u) = \text{expected value of } Y = f(x_u, \theta),\ u = 1, 2, ..., n \tag{3.27}$$

$$Y_u - f(x_u, \theta) = \varepsilon_u,\ \text{Variance}(Y_u) = \text{Variance}(\varepsilon) = \sigma^2. \tag{3.28}$$

However, if the heterogeneity of variance arises in the nonlinear regression model, these assumptions are changed to:

$$E(Y_u) = \mu_u(\theta) = f(x_u, \theta), \text{where } \mu \text{ is the mean} \tag{3.29}$$

$$\text{Variance}(Y_u) = \sigma^2 g^2(\mu_u(\theta), \theta) = \sigma^2 / w_u, \tag{3.30}$$

where g is the variance function and w_u is the true weights. According to Equation 3.30, the weights are defined as:

$$w_u = \frac{1}{g^2(\mu_u(\theta), \theta)}. \tag{3.31}$$

This implies that the weights are the inverses of the variances. If Equation 3.15 is multiplied by $\sqrt{w_u}$, the nonlinear regression model now becomes:

$$\begin{aligned} &\sqrt{w_u} Y_u = \sqrt{w_u} f(x_u, \theta) + \sqrt{w_u} \varepsilon_u \ \text{ or} \\ &Y_u^* = f^*(w_u, x_u, \theta) + \varepsilon_u^*. \end{aligned} \tag{3.32}$$

The redefined responses Y_u^* have constant variances with means given by the new nonlinear function f^*. The idea is to minimize the error sum of squares in Equation 3.32 as:

$$\text{Minimize} \quad \sum_{u=1}^{n} \left[Y_u^* - f^*(w_u, x_u, \theta) \right]^2, \tag{3.33}$$

which is equivalent to

$$\text{Minimize} \quad \sum_{u=1}^{n} w_u \left[Y_u^* - f^*(x_u, \theta) \right]^2. \tag{3.34}$$

Equation 3.34 indicates that the larger the values of the weight w_u, the larger the contribution of the squared errors. This implies that the values of the weight w_u can justify the variance of the error term be constant. Hence, when a technique of weighted least squares is implemented, the assumption that the variance of the error term be constant is satisfied.

CHAPTER 4

CORROSION SEVERITY RANKING ANALYSIS

This chapter presents results from the data screening analyses and corrosion severity ranking analysis using principal component analysis. Specifically, Section 4.1 presents the data screening analyses consisting of data quality check, outlier analysis, and missing observation analysis. Section 4.2 gives the results of the corrosion severity ranking analysis performed by the method of PCA.

4.1 Data Screening Analyses

This section describes the results and the analysis conducted to improve the quality of the data. Three areas are addressed: data quality check, outlier analysis, and missing observation analysis.

4.1.1 Data quality check and outlier analysis

This section gives an example of data quality check and outlier analysis. The first stage of the outlier detection is essentially a data quality check that compares data to known limits for each atmospheric condition.

Table 4.1 shows a summary of the data quality check on all data sets of the atmospheric conditions for the six operational air bases. Percentages of data quality check (that failed after comparing data to known limits for each atmospheric condition) for the data sets of relative humidity and rain pH at each operational air base are investigated. The maximum value of relative humidity must not exceed 100% while the

maximum and minimum values of pH must be not higher than 14 and lower than 0, respectively.

After detecting outliers with data quality checks, it is necessary to perform a second pass through the data to identify outliers. The procedure of treating outliers is given in the flowchart for detecting outliers provided in Section 3.1.1, Figure 3.2. This procedure begins with the first two consecutive observations and then the second two, etc. When the difference between the current value and the prior value exceeds some constant value (c), the data point might be considered as an outlier. In this analysis, the constant value, c is determined by multiplying the standard deviation of the atmospheric condition data set by three.

Table 4.1: Summary of data quality check

Base	Percentage of data quality check (that failed after comparing data to known limits for each atmospheric condition)	
	RH	pH
Hickam AFB	0.5841	0.2351
Kadena AB	0.1852	0
Macdill AFB	0	0.1207
RAF Mildenhall	0.2211	0
Pease ANGB	0	0.1093
Seymour Johnson AFB	0.1448	0

Note: RH: relative humidity and pH: rain pH

Table 4.2: Summary of outlier analysis

Base	Percentage of outliers				
	RH	AT	DP	pH	ST
Hickam AFB	0.8633	0.2217	0	0.0313	0.8862
Kadena AB	0.9311	0.1019	0	0.0219	1.3632
Macdill AFB	1.1012	0.2336	0	0.1038	1.1211
RAF Mildenhall	0.6065	0.1581	0	0.2011	0.8974
Pease ANGB	0.4876	0.2129	0	0.0427	0.7736
Seymour Johnson AFB	0.2992	0.1977	0	0.1127	0.9103

Note: RH: relative humidity, AT: air temperature, DP: dew point temperature, pH: rain pH, and ST: surface temperature measured on the material

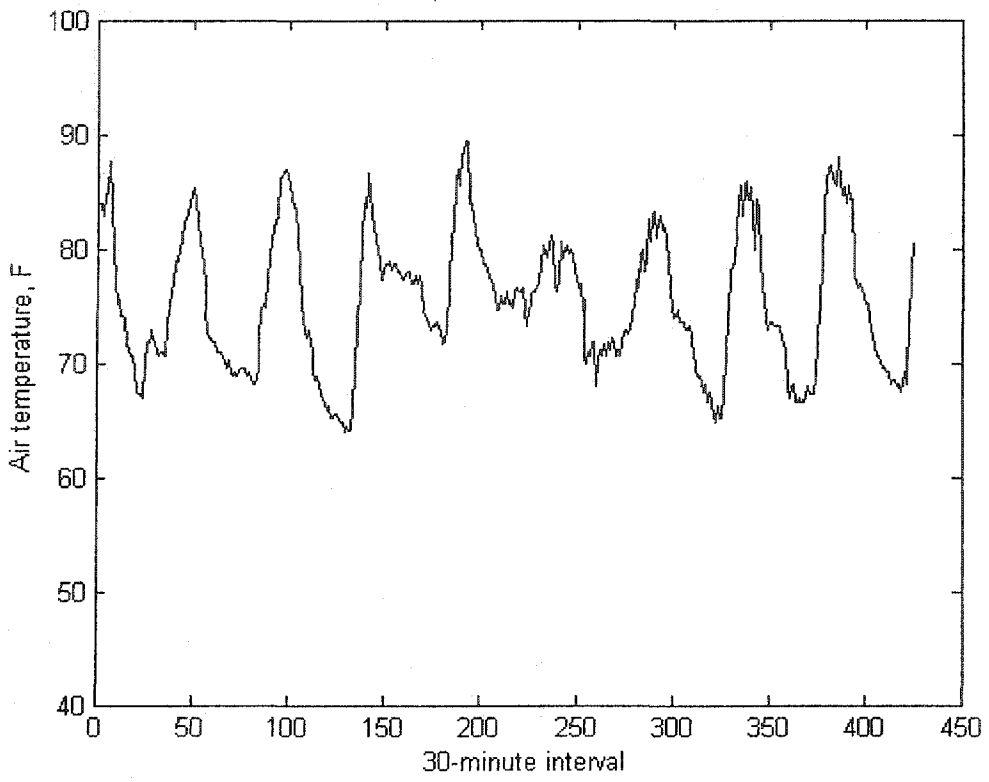


Figure 4.1: A snap-shot of 30-minute interval recordings of air temperature at Hickam AFB

Figure 4.1 depicts the data set of air temperature collected at Hickam AFB used to demonstrate outlier analysis. This data set has 425 data points with an average 75.45°F and standard deviation 6.03°F. The constant value c is equal to 3×6.03°F = 18.09°F. To demonstrate the outlier procedure, consider two consecutive air temperature data points are 84.3°F (x_{t-1}) and 78.6°F (x_t). Outlier can be detected by testing:

$$|x_t - x_{t-1}| > c,$$

where x_t and x_{t-1} are the two consecutive air temperature data points. Since the difference between the two values is less than 18.09°F, x_t (78.6°F) is not treated as an outlier.

Based on the data set of air temperature for Hickam AFB, 0.22 percent of the data set is deemed as outliers. Since each data set of atmospheric conditions for an operational air base were recorded in 30-minute time intervals with more than 20,000 realizations and many missing data gaps, it is not reasonable to demonstrate all of the data and the outlier analysis conducted. Table 4.2 shows a summary of the outlier analysis on all data sets of the atmospheric conditions for the six operational air bases. Note also that percentage of outliers for the data sets of dew point temperature at each operational air base is zero because dew point temperature is estimated after performing outlier analysis of relative humidity and air temperature.

4.1.2 Missing observation analysis

The method of neural networks is used to predict missing observations (for both outliers and non-recorded data values) in the atmospheric condition data sets. Once again, the data set of air temperature for Hickam AFB is used to demonstrate missing observation analysis. Figure 4.2 illustrates the neural network experiment implemented by the Matlab® neural network toolbox. This experiment is conducted by creating a feed forward neural network with five neurons in one hidden layer, with the hyperbolic tangent sigmoid as the transfer function in the hidden layer and output layer, and with gradient descent as the convergent criterion. Recall that as Tang et al. (1991) suggested, the learning rate for efficient learning is 0.1, the momentum is 0.9, and the appropriate number of epochs for training the networks is 3,000.

A validation test was run using simulation. The results show that the predicted values and the actual values follow the same pattern, indicating that neural networks are appropriate to predict the missing observations for long-term forecasting. To illustrate

the process and its validation, the data set of 425 data points of air temperature for Hickam AFB is used to predict consecutive missing values. Based on the data set, 96 missing data points are replaced with values generated by the neural network methodology. The graph of Figure 4.2 depicts the actual versus the predicted values and indicates the ability of neural network to replace missing data. Figure 4.3 illustrates the data set of air temperature for Hickam AFB and the 96 data points predicted from the method of neural networks.

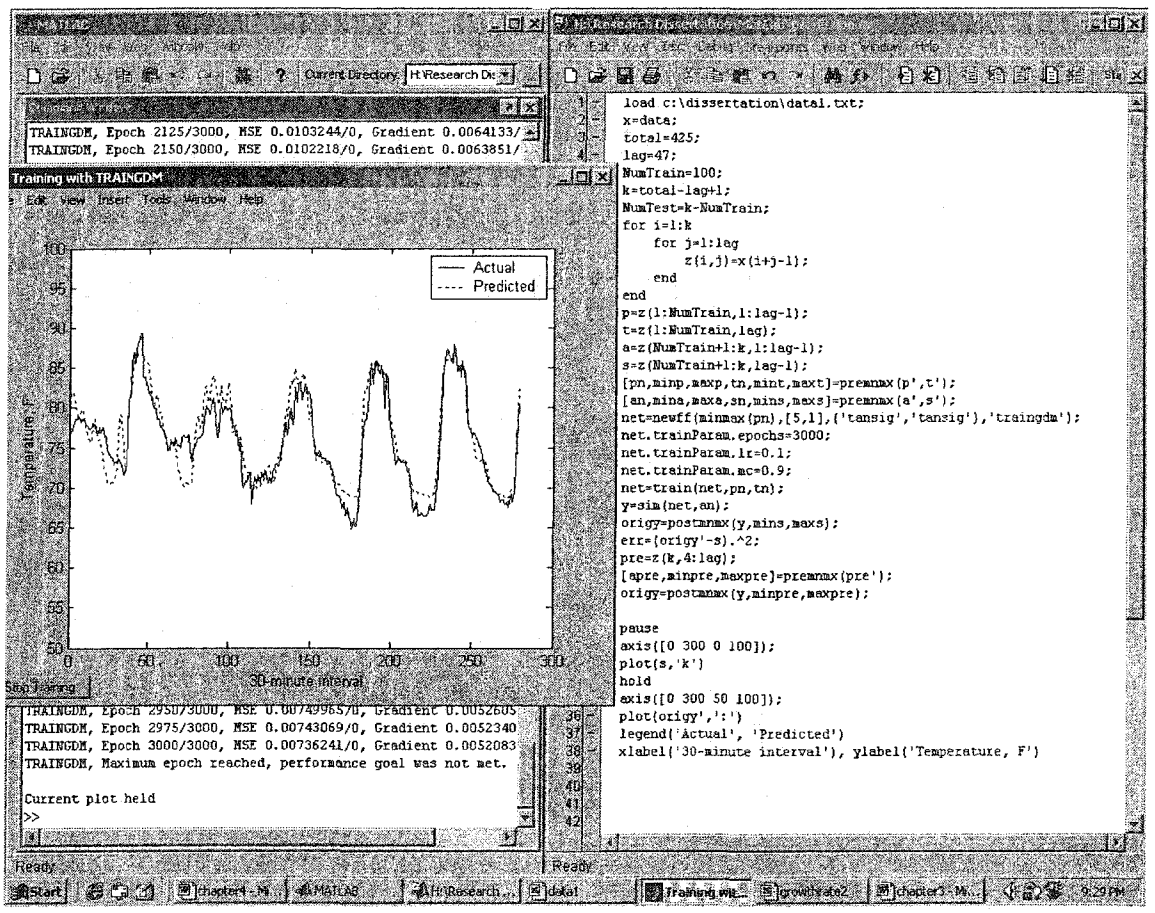


Figure 4.2: Neural network experiment screen

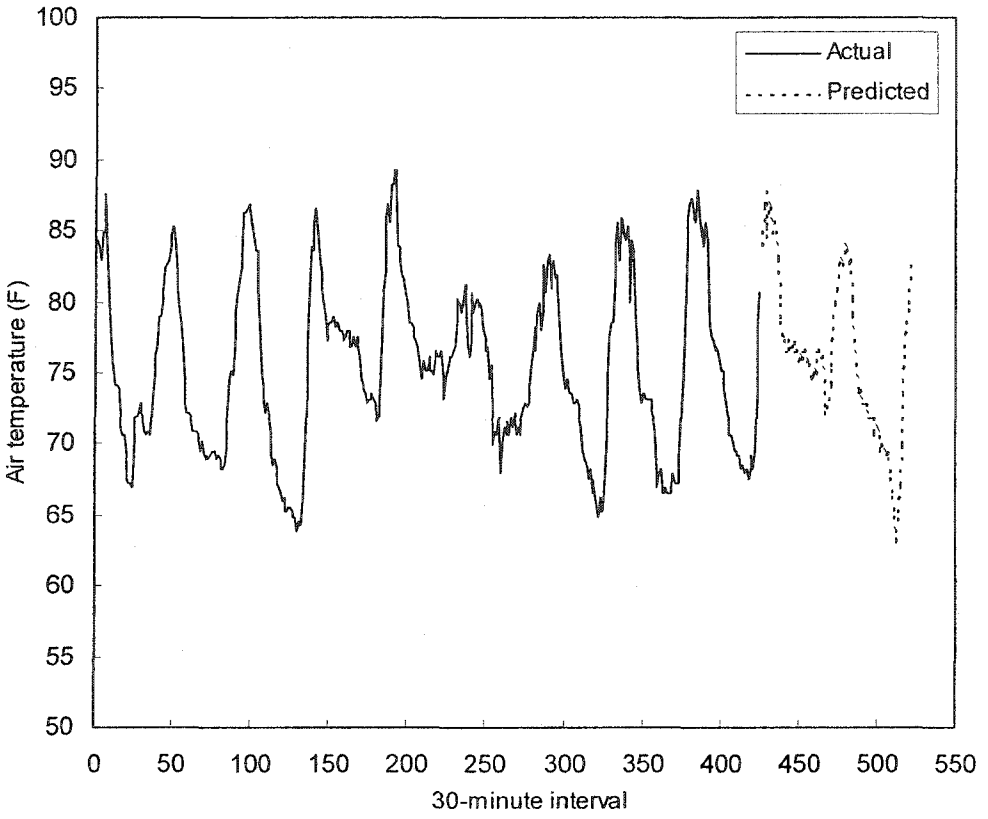


Figure 4.3: Data points predicted from the method of neural network

Table 4.3: Summary of missing observation analysis

Base	Number of gaps	Number of data points replaced in each gap	Overall percentage of the data points replaced
Hickam AFB	6	96, 89, 144, 128, 192, 336	2.04
Kadena AB	8	288, 155, 96, 144, 240, 192, 124, 144	3.54
Macdill AFB	15	144, 192, 235, 280, 91, 336, 432, 309, 288, 96, 144, 48, 96, 48, 192	8.98
RAF Mildenhall	11	192, 96, 96, 288, 240, 240, 144, 90, 127, 48, 79	3.64
Pease ANGB	18	240, 96, 240, 192, 187, 125, 96, 48, 144, 142, 94, 156, 147, 192, 144, 96, 48, 144	5.13
Seymour Johnson AFB	7	144, 96, 240, 288, 480, 432, 288	4.31

Table 4.3 summarizes the number of gaps and the data points of all atmospheric condition data sets generated by the method of neural networks of the six bases. Hickam AFB had the least number of data gaps while Macdill AFB had the highest percentage.

4.2 Corrosion Severity Ranking

One of the objectives of this research is to rank corrosion severity by base. To meet this purpose, the method of principal component analysis (PCA) is used to perform analysis on the compositional data sets rather than the raw data. The number of 30-minute intervals when an atmospheric condition is conducive to corrosion growth is derived by counting the number of times that each atmospheric condition supports a condition for corrosion growth.

Table 4.4 shows the number of 30-minute intervals for each of the eight atmospheric condition variables for the six operational air bases was deemed as supporting conditions for corrosion growth. Table 4.5 illustrates the new data set (i.e., the compositional data sets) transformed from the numbers of 30-minute intervals when an atmospheric condition is conducive to corrosion growth and represents a probability distribution for each base's atmospheric condition. Note that the total of each compositional variable is equal to one. For example, an RH′ of 0.2066 for Hickam AFB means that there is a 20.66% chance that the RH level at Hickam AFB will meet the conditions for supporting corrosion growth. The new compositional variable for this condition is represented as RH′. Similarly, the new compositional variables are AT′, DP′, pH′, RF′, TOW1′, TOW2′, and ST′ for AT, DP, pH, RF, TOW1, TOW2, and ST, respectively.

The results of the PCA reveal the covariance matrix, eigenvalues of covariance matrix (arranged in order of magnitude), eigenvectors, the first two principal components, a scree plot of eigenvalues, and a scatter plot of the first two principal components for the compositional data set. These results are illustrated in Tables 4.6-4.9 and Figures 4.4-4.5.

Table 4.4: Numbers of 30-minute intervals when an atmospheric condition is conducive to corrosion growth for each atmospheric data condition

Bases	RH	AT	DP	pH	RF	TOW1	TOW2	ST	# of intervals
Hickam	45855	48112	48212	3379	3581	18094	17340	15261	48252
Kadena	36299	27149	34067	3093	2712	17433	17029	2903	39048
Macdill	26349	27883	26140	3651	1263	12901	11025	7042	32632
Mildenhall	37044	9751	14300	2164	2183	9138	7419	16178	45115
Pease	20104	16802	19571	1560	1052	7839	6805	2074	49336
Seymour	31261	29528	26114	702	2282	17264	9816	15056	45712
Total	196912	159365	168445	11549	11073	82670	64433	48514	

Note: RH: relative humidity, AT: air temperature, DP: dew point temperature, pH: rain pH, RF: rainfall, TOW1: time-of-wetness used to detect light dew, TOW2: time-of-wetness while used to detect rain and heavier liquid condensation, and ST: surface temperature measured on the material

Table 4.5: Compositional data set

Bases	RH′	AT′	DP′	pH′	RF′	TOW1′	TOW2′	ST′
Hickam	0.2066	0.2659	0.2525	0.2074	0.2457	0.1917	0.2177	0.2367
Kadena	0.2021	0.1854	0.2204	0.2346	0.2300	0.2283	0.2642	0.0556
Macdill	0.1755	0.2279	0.2024	0.2769	0.1282	0.2021	0.2047	0.1615
Mildenhall	0.1785	0.0576	0.0801	0.1421	0.1602	0.1036	0.0996	0.2683
Pease	0.0886	0.0908	0.1002	0.0936	0.0706	0.0812	0.0836	0.0315
Seymour	0.1487	0.1723	0.1443	0.0455	0.1653	0.1931	0.1301	0.2465
Total	1	1	1	1	1	1	1	1
Mean	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667

Table 4.6: Covariance matrix

	RH′	AT′	DP′	pH′	RF′	TOW1′	TOW2′	ST′
RH	0.0019							
AT	0.0019	0.0063						
DP	0.0021	0.0052	0.0048					
pH	0.0025	0.0039	0.0042	0.0078				
RF	0.0025	0.0029	0.0032	0.0021	0.0042			
TOW1	0.0018	0.0039	0.0035	0.0027	0.0026	0.0035		
TOW2	0.0024	0.0045	0.0046	0.0051	0.0034	0.0038	0.0053	
ST	0.0018	0.0008	-0.0003	-0.0015	0.0022	0.0004	-0.0012	0.0104

Table 4.7: Eigenvalues of the covariance matrix

Number	Eigenvalue	Difference	Proportion	Cumulative
1	0.025566	0.013688	0.58	0.58
2	0.011878	0.008249	0.27	0.85
3	0.003629	0.001236	0.08	0.93
4	0.002394	0.001597	0.05	0.98
5	0.000796	0.000796	0.02	1.00
6	0.000000	0.000000	0.00	1.00
7	0.000000	0.000000	0.00	1.00
8	0.000000		0.00	1.00

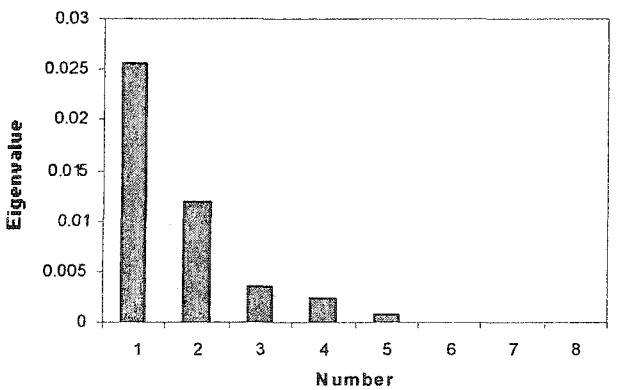


Figure 4.4: Scree plot

Table 4.8: Eigenvectors

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
RH′	0.2218	0.1562	0.2019	0.2899	0.02501	0.8950	0.0000	0.0000
AT′	0.4404	0.0629	-0.3762	-0.5678	-0.2120	0.1547	-0.0142	-0.5187
DP′	0.4176	-0.0349	-0.2246	-0.1281	-0.3430	0.0044	0.1127	0.7917
pH′	0.4359	-0.2168	0.7970	-0.1927	-0.0123	-0.1873	0.2213	-0.0803
RF′	0.3040	0.2268	-0.1301	0.6632	-0.3798	-0.2898	0.2995	-0.2854
TOW1′	0.3271	0.0403	-0.2652	0.0553	0.8041	-0.0686	0.4048	0.0541
TOW2′	0.4392	-0.0038	-0.0222	0.2428	0.1996	-0.1682	-0.8164	0.0000
ST′	0.0157	0.9260	0.2103	-0.1952	0.0770	-0.1518	-0.1342	0.1138

Table 4.9: The first two principal components

Base	Principal component 1	Principal component 2
Hickam	0.1619	0.0786
Kadena	0.1487	-0.1070
Macdill	0.1084	-0.0363
Mildenhall	-0.1427	0.1011
Pease	-0.2061	-0.1398
Seymour Johnson	-0.0702	0.1034

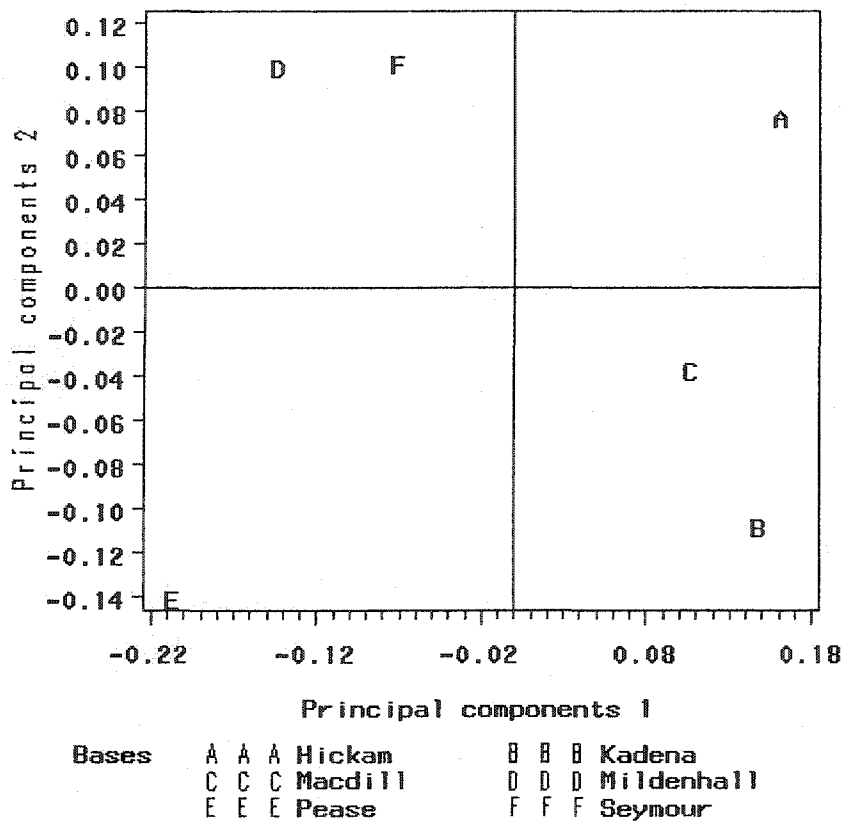


Figure 4.5: Scatter plot of the first two principal components for the six air bases

According to the compositional data in Table 4.5, the values of the eight compositional atmospheric condition variables (i.e., RH′, AT′, DP′, pH′, RF′, TOW1′, TOW2′, and ST′) of Hickam AFB are larger than the mean vector of the eight compositional atmospheric condition variables (0.1667) whereas the values of the eight compositional atmospheric condition variables of Pease ANGB are less than the mean vector of the eight compositional atmospheric condition variable (0.1667). Recall that atmospheric corrosion depends on the length of time that moisture is present on the metal's surface. This shows that more moisture is present on the aircraft's surface at Hickam AFB than on the aircraft's surface at Pease ANGB. Hence it is reasonable to

infer that Hickam AFB is the most severe site for corrosion whereas Pease ANGB is deemed to be the least severe site for corrosion.

The first two principal components are used for corrosion severity ranking analysis since the first two principal components can account for most of the variability of the data as shown in the scree plot of Figure 4.4. The scree plot involves selecting the number of important principal components based on the visual appearance of the plot. According to Figure 4.4, the number (2) of principal components to be selected is determined in such a way that the slope of the plot is steep to the left of 2 but at the same time not steep to the right. Consequently, the first two principal components are used for corrosion severity ranking analysis. The first two principal components of the six air bases can be expressed as:

$$\begin{bmatrix} z_{11} & z_{12} \\ z_{21} & z_{22} \\ \vdots & \vdots \\ z_{61} & z_{62} \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{18} \\ x_{21} & x_{22} & \cdots & x_{28} \\ \vdots & \vdots & \ddots & \vdots \\ x_{61} & x_{62} & \cdots & x_{68} \end{bmatrix} \begin{bmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \\ \vdots & \vdots \\ u_{82} & u_{82} \end{bmatrix},$$

where z_{ij} represents an element of principal component in a vector of principal component scores, x_{ij} denotes an element in a matrix of the eight atmospheric condition variables and the six air bases, and u_{ij} represents an element in an eigenvector. This implies that as much variation in the data as possible lies along the direction of the first eigenvector and the second eigenvector, consecutively. Since the eigenvectors are orthogonal, the principal components represent jointly perpendicular directions through the space of the original variables. This implies that the transformed variables are uncorrelated with each other and centered at the origin.

The first two principal component equations for each of the six air bases are given by:

$$PC1 = 0.22RH' + 0.44AT' + 0.42DP' + 0.44pH' + 0.30RF' + 0.33TOW1' + 0.44TOW2' + 0.02ST'$$

and

$$PC2 = 0.16RH' + 0.06AT' - 0.03DP' - 0.22pH' + 0.23RF' + 0.04TOW1' - 0.004TOW2' + 0.93ST',$$

where PC1 and PC2 are the first and the second principal components, RH′ is compositional relative humidity variable, AT′ is compositional air temperature variable, DP′ is compositional dew point temperature variable, pH′ is compositional rain pH variable, RF′ is compositional rainfall variable, TOW1′ is compositional time-of-wetness variable detected by light dew sensor, TOW2′ is compositional time-of-wetness variable detected by rain and heavier liquid condensation sensor, and ST′ is compositional aircraft's surface temperature variable. The first two principal component equations for each of the six air bases are the same. This shows that each principal component of each air base depends on the compositional atmospheric condition variables (i.e., the compositional data).

According to Tables 4.7 and 4.8, the first principal component with the variance 0.0256 explains about 58 percent of the variability and the eigenvector corresponding to the variance 0.0256 listed under the column $\mathbf{u}_1$ is $(0.2218, 0.4404, 0.4176, 0.4359, 0.3040, 0.3271, 0.4392, 0.0157)^t$. That is, the first principal component with all its positive coefficients seems to measure the general severity of a base. The second principal component with the variance 0.0119 explains about 27 percent of the variability and the eigenvector corresponding to the variance 0.0119 listed under the column $\mathbf{u}_2$ is

(0.1562, 0.0629, -0.0349, -0.2168, 0.2268, 0.0403, -0.0038, 0.9260)′. That is, the second principal component consists of positive and negative coefficients indicating a contrast between the first group of the compositional atmospheric condition variables (RH′, AT′, RF′, TOW1′, and ST′) and the second group of the compositional atmospheric conditions variables (DP′, pH′, and TOW2′). By examining the cumulative proportion of the variation given in Table 4.7 and scree plot depicted in Figure 4.4, at least two principal components are needed to account for 85 percent (58 + 27) of the total variability.

Based on the scatter plots of the first two principal components depicted in Figure 4.5, the horizontal line and the vertical line aligned zeros are used to separate the bases into four quadrants. The northeast quadrant is considered as the most severe zone for the compositional atmospheric conditions affecting corrosion (i.e., indicating positive signs that imply high impact affecting corrosion) while the southwest quadrant is considered as the least severe zone for the compositional atmospheric conditions affecting corrosion (i.e., indicating negative signs that imply low impact affecting corrosion). Thus, the results of the most severe and least severe zones as determined by the scatter plot of the two principal components can be used to rank the corrosion severity by locations.

Since the first principal component accounts for variability more than the second principal component, the remaining four sites should rank by considering the first principal component. Accordingly, the ranking for the six operational air bases in terms of the corrosion severity ranks from the most severe site to the least severe site: Hickam AFB, Kadena AB, Macdill AFB, Seymour Johnson AFB, RAF Mildenhall, and Pease ANGB.

General considerations of climate and distance from the sea or salt sources provide qualitative information regarding atmospheric corrosion (Brown and Masters,

1982). Marine and tropical climates are usually highly corrosive, and the corrosivity tends to be significantly dependent on the distance from the sea. Table 4.10 provides the information of the climate types and distance from the sea of the test site locations of the rack exposure (information on climate and distance culled from Howard et al., 1999). Hickam AFB has a tropical climate that is deemed a severe climate for the atmospheric corrosion and the distance from the sea is less than one mile, while Macdill AFB has a marine climate that is deemed a severe climate to the atmospheric corrosion and the test site is also nearby the sea. Since the approximate distances from the seas of RAF Mildenhall and Seymour Johnson AFB are far, the types of climate should be considered as the significant factor. Hence the least two severe corrosion test site locations would be RAF Mildenhall and Seymour Johnson.

Based on the first full year of monitoring of the environment at the six operational air bases, the results of corrosion severity ranking developed by Arinc, Inc. showed that the most severe locations were Hickam AFB, Kadena AB, and RAF Mildenhall followed by Pease ANGB, Macdill AFB, and Seymour Johnson AFB.

Based on the three scenario considerations, Table 4.11 shows the results of corrosion severity ranking of the six operational air bases. All three scenarios show that the most severe location is Hickam AFB whereas the remaining rankings are different. The PCA analysis uses threshold values for each atmospheric condition and several years worth of data collection to establish corrosion severity ranking by locations. In addition, the other two ranking schemes do not consider all atmospheric conditions that allow moisture formation. Ignoring moisture presence ignores the catalyst for corrosion growth. Only PCA provides a comprehensive analysis of factors contributing to moisture or corrosion presence.

Table 4.10: Test site locations of rack exposure with climate types and distance from the sea (culled from Howard et al., 1999)

Test site location	Climate	Distance from the sea
Hickam AFB	a tropical, maritime, and torrid climate	approximately 2,000 feet from Mamala Bay and the entrance to Ford Channel Pearl Harbor
Kadena AB	a subtropical, and maritime climate	approximately 2,000 feet from the East China Sea
Macdill AFB	a marine, and sub-tropical climate	N/A (at a peninsula in Tampa Bay)
RAF Mildenhall	a temperate maritime climate	approximately 45 miles from the North Sea
Pease ANGB	a temperate maritime, and cold climate	approximately 1.5 miles from the Great Bay, Little Bay, and the Atlantic Ocean
Seymour Johnson AFB	a temperate climate	approximately 75 miles from the Atlantic Ocean

Table 4.11: Corrosion severity rankings by locations of the three scenarios (number 1 is the most severity)

Ranking number	Arinc, Inc.	Climate types and distance from the sea	PCA
1	Hickam AFB	Hickam AFB	Hickam AFB
2	Kadena AB	Macdill AFB	Kadena AB
3	RAF Mildenhall	Kadena AB	Macdill AFB
4	Pease ANGB	Pease ANGB	Seymour Johnson AFB
5	Macdill AFB	RAF Mildenhall	RAF Mildenhall
6	Seymour Johnson AFB	Seymour Johnson AFB	Pease ANGB

CHAPTER 5

CORROSION GROWTH ANALYSES

In this chapter corrosion growth data measurements obtained from the retrieved coupons are used for developing predictive models of corrosion based on time for the bases. The models analyzed are the existing growth model (i.e., the confined exponential model), the existing corrosion growth model (i.e., power law equation), and the proposed corrosion growth models (i.e., the three proposed models, GL, GC, and CL). The first section of this chapter displays the corrosion growth data measurements obtained from the coupons at each base and utilizes the data to obtain parameter estimates for the various models. Then, the models are statistically compared by their weighted mean square errors, lack-of-fit tests, and model adequacy checking procedures. The second section provides a discussion of the statistical tests and identifies the "best" model as based the results of these tests.

5.1 Predictive Corrosion Growth Models

Figure 5.1 illustrates the corrosion growth data from the specimens at the six operational air bases expressed in thickness loss. Graphically, all data sets display a similar nature in that there is a concave growth curve with decreasing corrosion growth over time. Repeat observations by year represent the thickness loss for the coupons retrieved and thus are considered repeat observations. The repeat observations at the first year of exposure have small variation whereas the repeat observations at the last two

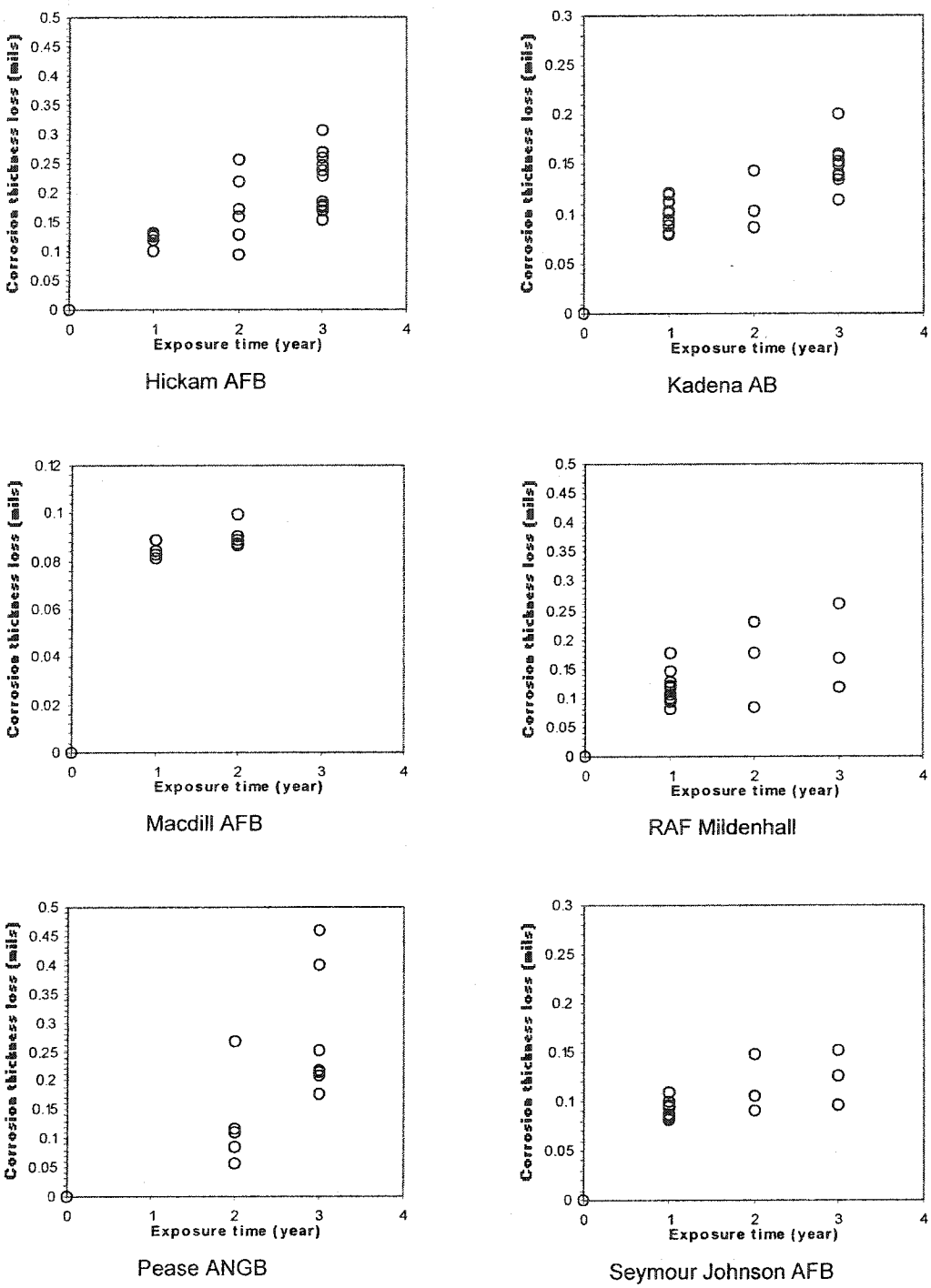


Figure 5.1: Corrosion growth data sets obtained from the six operational air force bases

years of exposure have higher variation. That is, as time goes on specimens tend to corrode at different rates. Note that Pease ANGB shows the largest variance in its data set for both years. Since the data sets of repeat observations for Macdill AFB and Pease ANGB are available only for two exposure times, performing predictive model formulation of corrosion growth is not reliable. Thus, the predictive models for corrosion growth as a function of exposure time is performed only on the remaining air bases: Hickam AFB, Kadena AB, RAF Mildenhall, and Seymour Johnson AFB.

The procedure used for fitting the confined exponential, the power law equation, and the models proposed in Section 3.6 (the GL, GC, and CL) is as follows:

1. SAS® software was used to perform nonlinear regression on the corrosion data sets. An iterative estimation method (Marquardt method) was then used to estimate the parameters for each of the five models.
2. A lack-of-fit test was used to check whether or not the fitted model displayed the behavior of growth (i.e., a concave limiting function).
3. Model adequacy checking techniques (i.e., plots of residuals versus fitted values and normality test) were used to assess model fit. If the residuals indicated non-constant variance, weighted least squares method was employed. Models passing the model adequacy checks were then compared with each other using error sum of squares criteria (detailed descriptions of these tests are outlined in Section 3.6).

Hickam AFB is used to explain the analysis and results while the analyses for Kadena AB, RAF Mildenhall, and Seymour Johnson AFB are in Appendix A. For Hickam AFB, Table 5.1 gives the SAS code for fitting the confined exponential model to the corrosion growth data set. The SAS procedure to fit nonlinear regression is PROC

NLIN with the fitting algorithm for model parameter estimates invoked by METHOD = Marquardt. The PARMS statement defines which elements of the model statement are the parameters to be estimated, (P^*), the ultimate limiting value of corrosion growth for the confined exponential model (a), the growth coefficient of the confined exponential model. The values (0.1 and 0.1) following the parameters (Pstar and a0) are their starting values. The MODEL statement defines the mathematical expression of the model, apart from the error term. To save fitted values and residuals, the OUTPUT statement is defined to save all data for the model adequacy test (i.e., residual analysis). PROC GPLOT is used to plot the residuals (the difference between the corrosion growth data and the fitted values). To check the model adequacy using normality tests of the residuals, PROC RANK is used to arrange residuals in increasing order. PROC GPLOT is also used to obtain normal probability plots. PROC RSREG is used to determine the pure error sum of squares while the lack-of-fit sum of squares is determined by subtracting pure error sum of squares from the total error sum of squares (as obtained from the procedure PROC NLIN). Pure error sum of squares, lack-of-fit sum of squares, and total error sum of squares are used for testing whether or not the data display the behavior of growth in the lack-of-fit test.

To fit the power law equation and the three new models (i.e., model GL, model GC, and model CL) to the corrosion growth data set, the MODEL statement is modified to correspond with the model equations for the power law equation, model GL, model GC, and model CL, respectively.

Generally, the ordinary least square (OLS) method is used to fit any model to data using least squares as the fit criterion. OLS has the underlying assumption that the distributions of the random errors are all normal and that these distributions all have the

same means (i.e., zero) and the same variances. If the variances are not equal, the method of weighted least squares has to be performed. Otherwise, inaccurate parameter estimates might be obtained.

Table 5.1: Confined exponential model fitting to corrosion growth data set of Hickam AFB using SAS®

```
data growth;
input year growthdepth @@;
cards;
/*Hickam data*/
0 0 1 0.1323 3 0.2471 3 0.3076;
/*lack of fit test*/
proc rsreg;
 model growthdepth = year/lackfit;
/*pure error estimation: SS(PE)*/
run;
proc nlin data = growth method = marquardt;
 parms Pstar=0.1 a0=0.1;
 model growthdepth = Pstar*(1-exp(-a0*year));
  output out = GrowthResult p = preds r = resid student = stde
   l95m = l95mean u95m = u95mean;
run;
/*set dimension plot*/
goptions hsize = 3 in vsize = 2.8 in;
proc gplot data = GrowthResult;
/*Residual plot against predictive values*/
 symbol1 v = circle c = black;
 plot stde*preds = 1/VAXIS = axis1 VMINOR = 0 VREF = 0;
 label stde = 'Standardized residuals';
 label preds = 'Fitted Value'; axis1 label = (a=90 r=0);
run;
/*Normal probability plot*/
proc rank data = GrowthResult out = normsc normal = blom;
 var stde; ranks nscore;
run;
proc gplot data = normsc;
 symbol2 v = circle c = black;
 plot stde*nscore = 2/VAXIS = axis1 VMINOR = 0;
 label nscore = 'Normal quantile'; label stde = 'Standardized residuals';
 axis1 label = (a=90 r=0);
run;
```

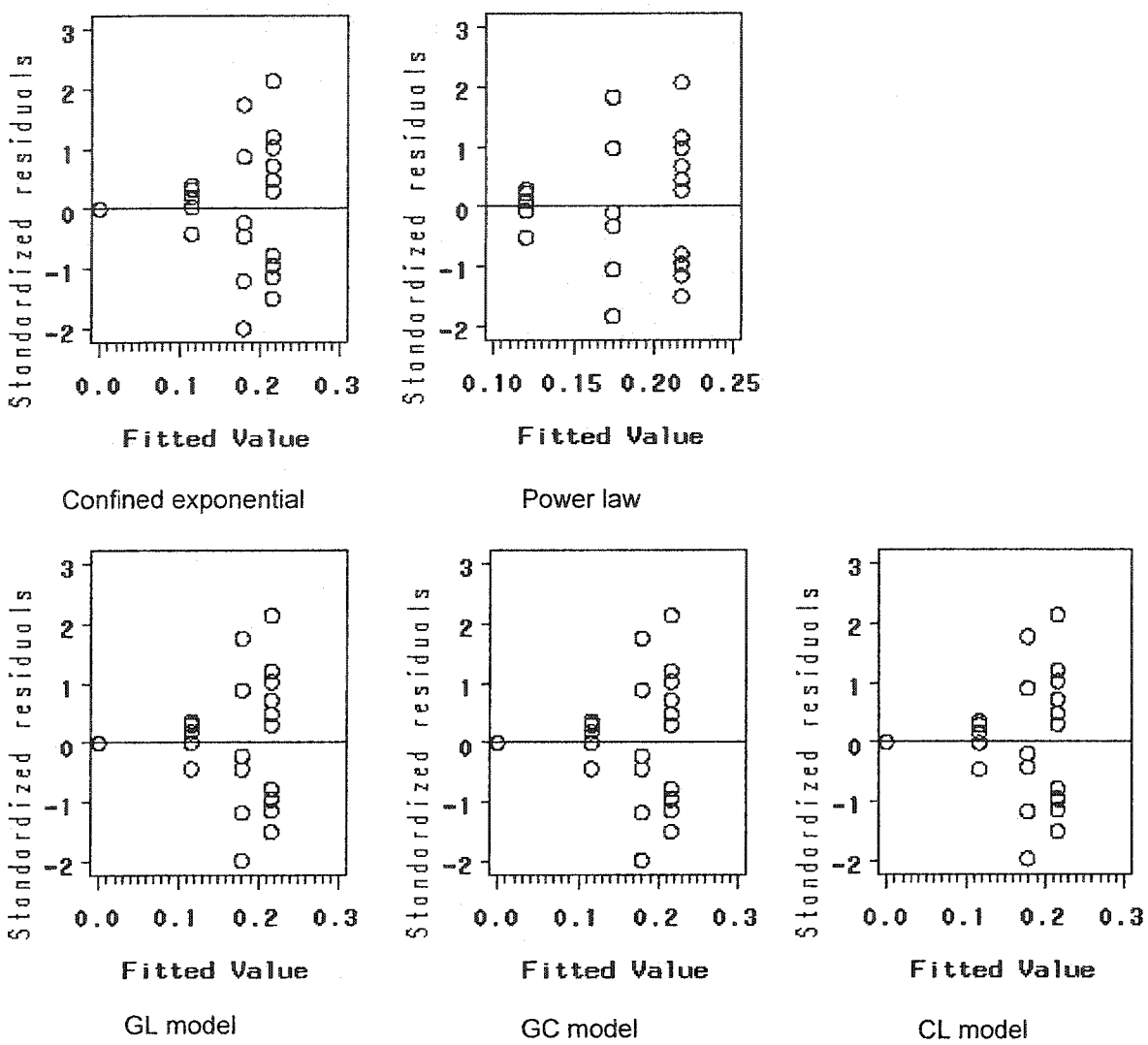


Figure 5.2: Residual analysis from the results of the five models using OLS for Hickam AFB

Figure 5.2 depicts residual analyses of the five models using the ordinary least squares for Hickam AFB. The plots show suspicious behavior, wedge-shaped pattern or a megaphone pattern, indicating the assumption of constant variance is not appropriate. Thus, the method of weighted least squares is used to correct for the non-constant variances. The weighted least square procedures and calculations for Hickam AFB, Kadena AB, RAF Mildenhall, and Seymour Johnson AFB are given in Appendix B.

After performing weighted least squares, Tables 5.2 and 5.3 show the results of the lack-of-fit test and the model parameter estimates for the confined exponential model using the Hickam AFB data. According to the results of the lack-of-fit tests for the confined exponential models for the Hickam AFB data, the F and p values indicate the lack-of-fit test not significant at a 5 percent level of significance. This implies that the model appears to be adequate for fitting the corrosion growth data set for Hickam AFB. In addition, the F-test for the overall regression of the confined exponential model indicates significance. This also implies that a nonlinear relationship exists between the average corrosion thickness loss and time, as observed during a three-year exposure time for Hickam AFB. The approximate 95% confidence limits of the final model parameter estimates for the confined exponential model exclude zero. This indicates nonzero values for all model parameters and implies that the confined exponential model is adequate for the corrosion growth data. Moreover, Figure 5.3 shows the model adequacy improves when using the weighted least square approach as indicates by the residuals and the normal probability plots.

For the power law equation and the three new models (GL, GC, and CL), the results of the lack-of-fit tests also show model significance at a 5 percent level of significance. However, the results of the model parameter estimation for the GL and GC models indicate infinity values for one of the model parameters. The GL and GC models are complicated mathematical models that require iterative parameter estimation procedures that often fail to converge.

Similarly, results for the lack-of-fit tests, model parameter estimates, and model adequacy checking of the five models for Kadena AB, RAF Mildenhall, and Seymour

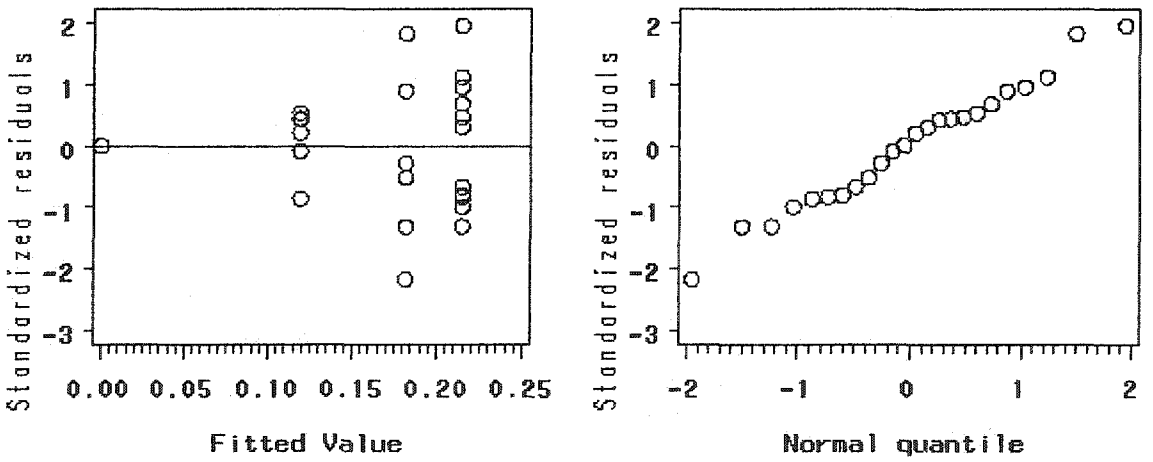
Johnson AFB data sets show the same results. Appendices A and B gives all results for Kadena AB, RAF Mildenhall, and Seymour Johnson AFB, respectively.

Table 5.2: Lack-of-fit test for Confined Exponential model for Hickam AFB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	2	388.2000	194.1000	224.8657	< 0.0001
Lack of Fit	2	0.4482	0.2241	0.2471	0.79
Pure Error	20	18.5418	0.9271		
Total Error	22	18.9916	0.8632		

Table 5.3: Model parameter estimates for Confined Exponential model for Hickam AFB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.2507	0.0368	0.1744	0.3271
a_0	0.6466	0.1890	0.2546	1.0387



Residuals against fitted values plot

Normal probability plot

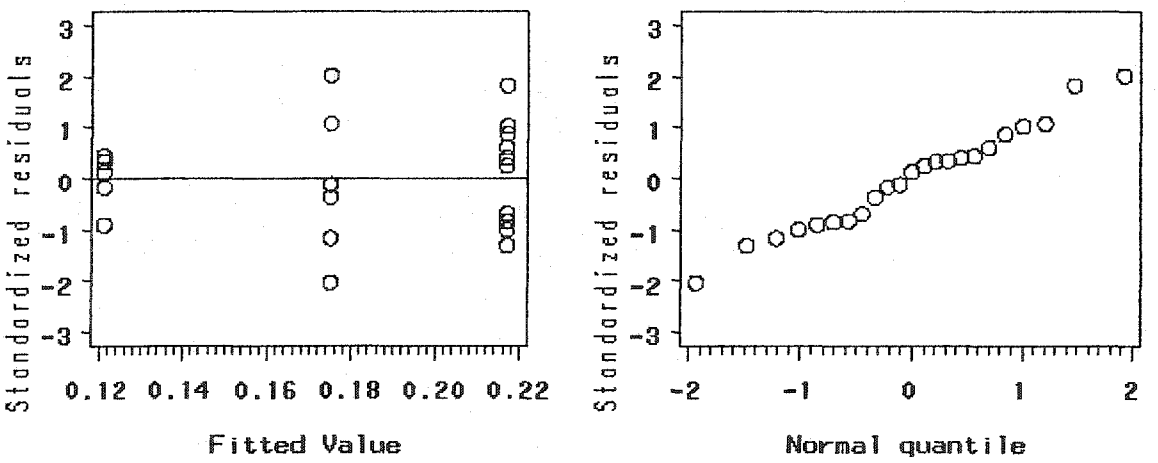
Figure 5.3: Model adequacy checking for Confined Exponential model for Hickam AFB data with weighted least squares

Table 5.4: Lack-of-fit test for Power Law model for Hickam AFB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	2	388.2000	194.1000	223.4537	< 0.0001
Lack of Fit	2	0.0597	0.0298	0.0313	0.97
Pure Error	20	19.0503	0.9525		
Total Error	22	19.1100	0.8686		

Table 5.5: Model parameter estimates for Power Law model for Hickam AFB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.1212	0.0106	0.0992	0.1432
a_0	0.5304	0.1056	0.3107	0.7501



Residuals against fitted values plot

Normal probability plot

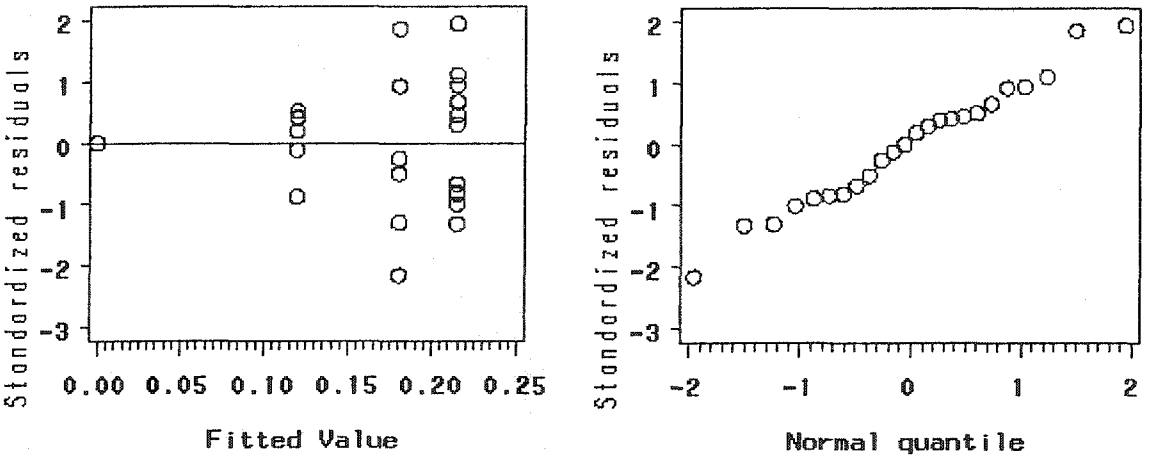
Figure 5.4: Model adequacy checking for Power Law model for Hickam AFB data with weighted least squares

Table 5.6: Lack-of-fit test for GL model for Hickam AFB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	3	387.9000	129.30	149.978	< 0.0001
Lack of Fit	2	0.3435	0.1717	0.0313	0.83
Pure Error	20	18.6233	0.9312		
Total Error	22	18.9700	0.8621		

Table 5.7: Model parameter estimates for GL model for Hickam AFB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.3509	0.1317	0.0779	0.6240
a_0	0.1706	0.0311	0.1060	0.2351
k	0.3110	Infinity	-Infinity	Infinity



Residuals against fitted values plot

Normal probability plot

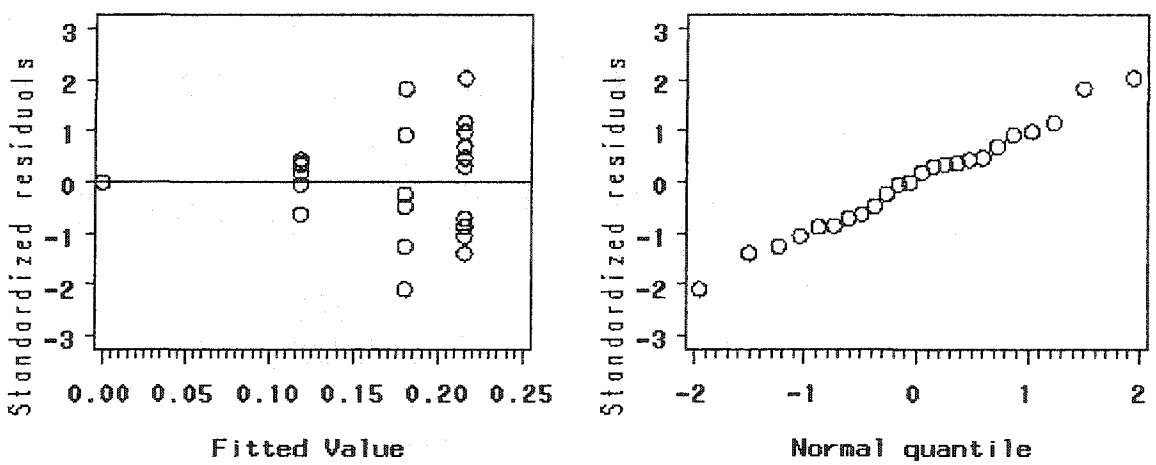
Figure 5.5: Model adequacy checking for GL model for Hickam AFB data with weighted least squares

Table 5.8: Lack-of-fit test for GC model for Hickam AFB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	3	269.9000	89.966	139.87	< 0.0001
Lack of Fit	2	0.0185	0.0092	0.0131	0.98
Pure Error	20	14.1315	0.7066		
Total Error	22	14.1500	0.6432		

Table 5.9: Model parameter estimates for GC model for Hickam AFB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.3595	0.0822	0.1890	0.5300
a_0	0.4634	Infinity	-Infinity	Infinity
k	0.3007	0.3340	-0.3921	0.9935



Residuals against fitted values plot

Normal probability plot

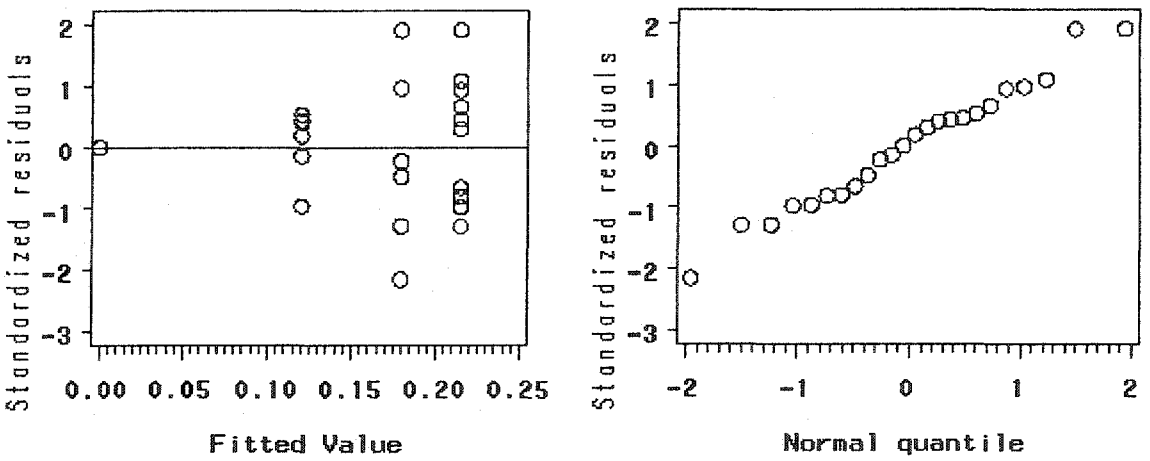
Figure 5.6: Model adequacy checking for GC model for Hickam AFB data with weighted least squares

Table 5.10: Lack-of-fit test for CL model for Hickam AFB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	2	238.5000	119.2500	231.9075	< 0.0001
Lack of Fit	2	0.2718	0.1359	0.2462	0.79
Pure Error	20	11.0409	0.5520		
Total Error	22	11.3127	0.5142		

Table 5.11: Model parameter estimates for CL model for Hickam AFB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.3543	0.0743	0.2003	0.5084
a_0	0.5147	0.2096	0.0800	0.9495



Residuals against fitted values plot

Normal probability plot

Figure 5.7: Model adequacy checking for CL model for Hickam AFB data with weighted least squares

5.2 Discussion

All five growth models were fit to the corrosion growth data sets for four operational air bases except one of the bases (i.e., the GC model of RAF Mildenhall). This is due to the inability of the iterative estimation procedure to converge when performing parameters estimates for the GC model. Table 5.12 summarizes all of the results for the parameter estimation by base, including weighted mean square errors (WMSE) for each corrosion growth model. Note that $\hat{P}$ represents the predicted thickness loss whereas t denotes exposure time. All summaries show that the CL model provides the best fit for all corrosion growth data sets of the four operational air bases and dominates the other models in terms of WMSE. Recall that the CL model is a modified model of the logistic growth model and the confined exponential growth model. Figures 5.8- 5.11 illustrate the five predictive corrosion growth models with upper and lower 95% confidence bounds on the predicted thickness loss for the four operational air bases. The upper and lower 95% confidence bounds of the predicted thickness loss increase as the exposure time increases. This implies that the variability of corrosion growth is higher when the coupons are exposed for longer periods of time. This behavior is supported by examining the corrosion growth measured from the coupons (see Figure 5.1).

Clearly, predictive corrosion growth models of Hickam AFB have higher growth trends compared to those of remaining air bases. According to the results of the corrosion severity ranking in Chapter 4, the PCA ranking for the six air bases in terms of corrosion severity from the most severe site to the least severe site is: Hickam AFB, Kadena AB, Macdill AFB, Seymour Johnson AFB, RAF Mildenhall, and Pease ANGB.

The results from the predictive corrosion growth models thus support the results from the corrosion severity ranking analysis.

Table 5.12: Summary of predictive corrosion growth models for the four bases (* indicating best model by location)

Base	Growth model	Parameterized model	WMSE
Hickam AFB	Confined exponential	$\hat{P} = 0.2507\left(1 - e^{-0.6466\,t}\right)$	0.8632
	Power law equation	$\hat{P} = 0.1212\,t^{0.5304}$	0.8686
	GL model	$\hat{P} = 0.3509 - \frac{0.3509}{\exp\left[1.5633\left(1 - e^{-0.3110\,t}\right)\right]}$	0.8621
	GC model	$\hat{P} = 0.3595 - \frac{0.3595}{\exp\left[1.5411\left(1 - e^{-0.3007\,t}\right)\right]}$	0.6432
	CL* model	$\hat{P} = \frac{0.1824t}{1 + 0.5147t}$	0.5142
Kadena AB	Confined exponential	$\hat{P} = 0.1520\left(1 - e^{-1.0121\,t}\right)$	0.8467
	Power law equation	$\hat{P} = 0.0972\,t^{0.3703}$	0.7956
	GL model	$\hat{P} = 0.1885 - \frac{0.1885}{\exp\left[1.933\left(1 - e^{-0.4693\,t}\right)\right]}$	0.5329
	GC model	$\hat{P} = 0.1937 - \frac{0.1937}{\exp\left[1.8844\left(1 - e^{-0.4499\,t}\right)\right]}$	0.5305
	CL* model	$\hat{P} = \frac{0.1957t}{1 + 1.0146t}$	0.5172
RAF Mildenhall	Confined exponential	$\hat{P} = 0.1940\left(1 - e^{-0.9239\,t}\right)$	0.5328
	Power law equation	$\hat{P} = 0.1174\,t^{0.4216}$	0.5415
	GL model	$\hat{P} = 0.1936 - \frac{0.1936}{\exp\left[-150.9919\left(1 - e^{0.0061\,t}\right)\right]}$	0.5331
	GC model	—	Failed to converge
	CL* model	$\hat{P} = \frac{0.2154t}{1 + 0.8388t}$	0.3189
Seymour Johnson AFB	Confined exponential	$\hat{P} = 0.1258\left(1 - e^{-1.3203\,t}\right)$	0.5497
	Power law equation	$\hat{P} = 0.0926\,t^{0.2832}$	0.5565
	GL model	$\hat{P} = 0.1477 - \frac{0.1477}{\exp\left[2.2296\left(1 - e^{-0.5794\,t}\right)\right]}$	0.5492
	GC model	$\hat{P} = 0.1480 - \frac{0.1480}{\exp\left[2.2237\left(1 - e^{-0.5776\,t}\right)\right]}$	0.4031
	CL* model	$\hat{P} = \frac{0.2365t}{1 + 1.5603t}$	0.3402

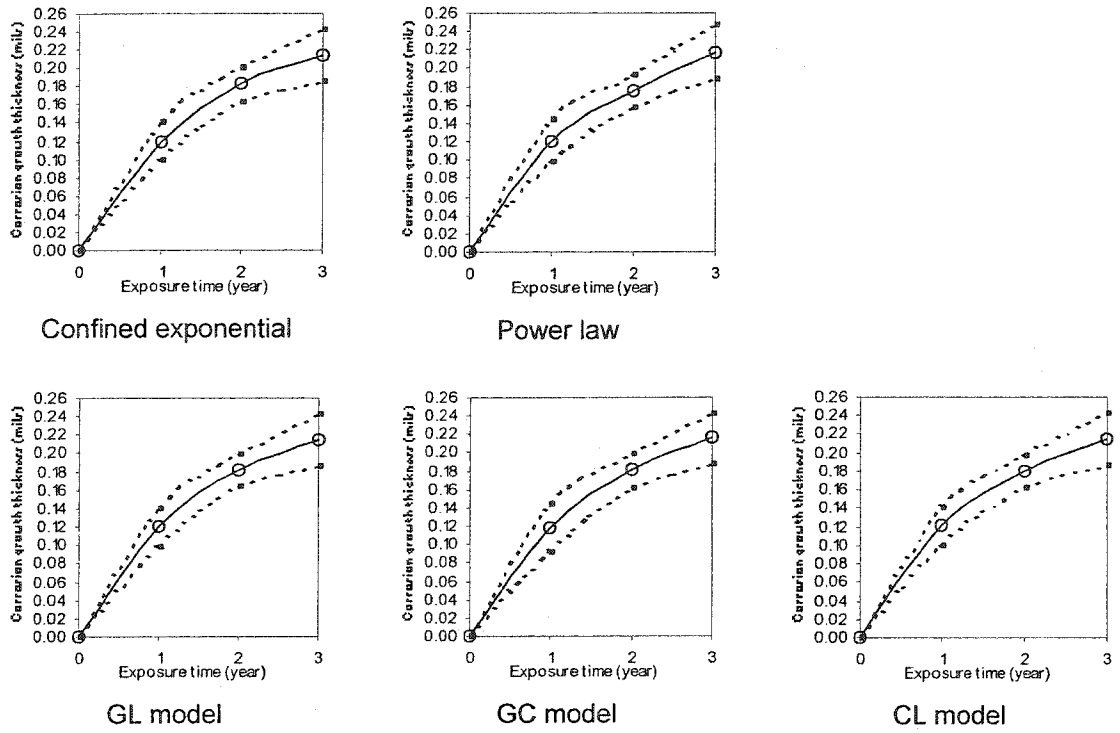


Figure 5.8: Corrosion growth predictive models for Hickam AFB

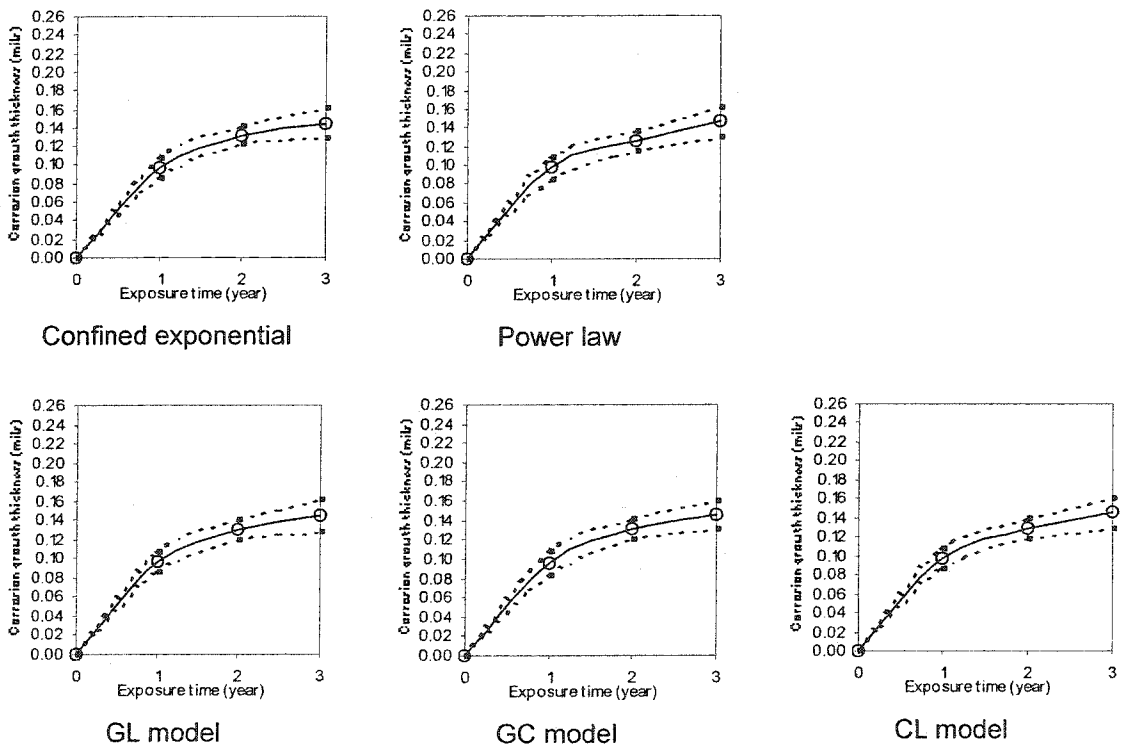


Figure 5.9: Corrosion growth predictive models for Kadena AB

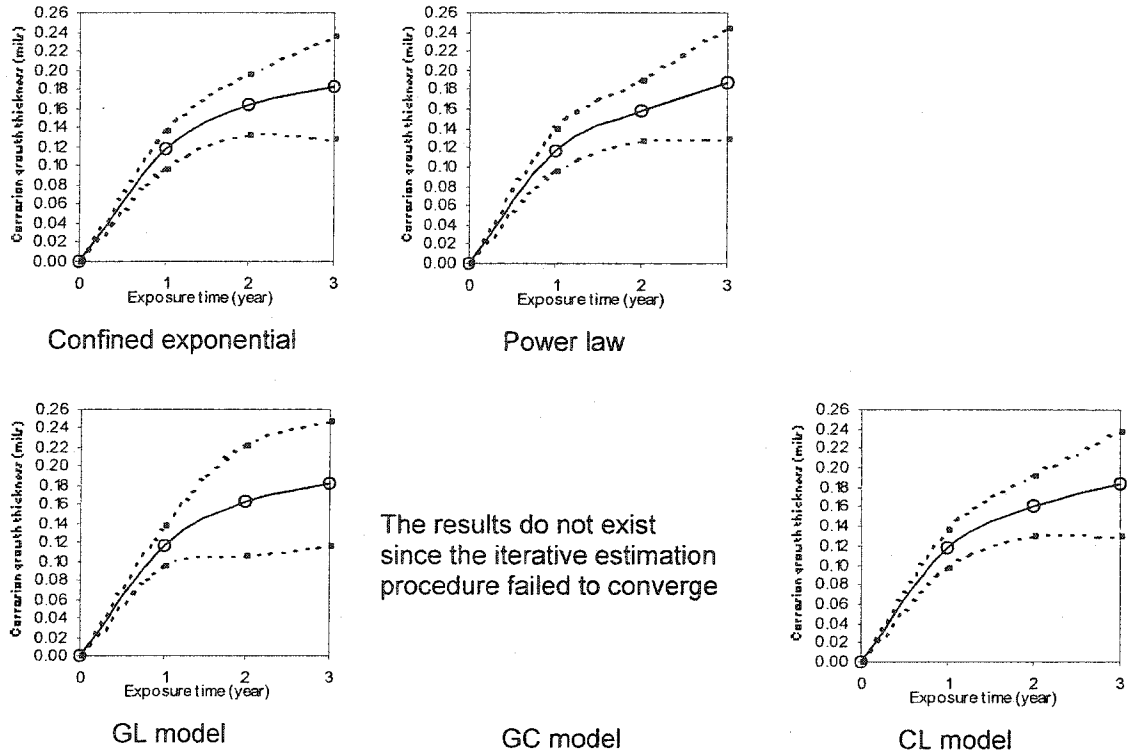


Figure 5.10: Corrosion growth predictive models for RAF Mildenhall

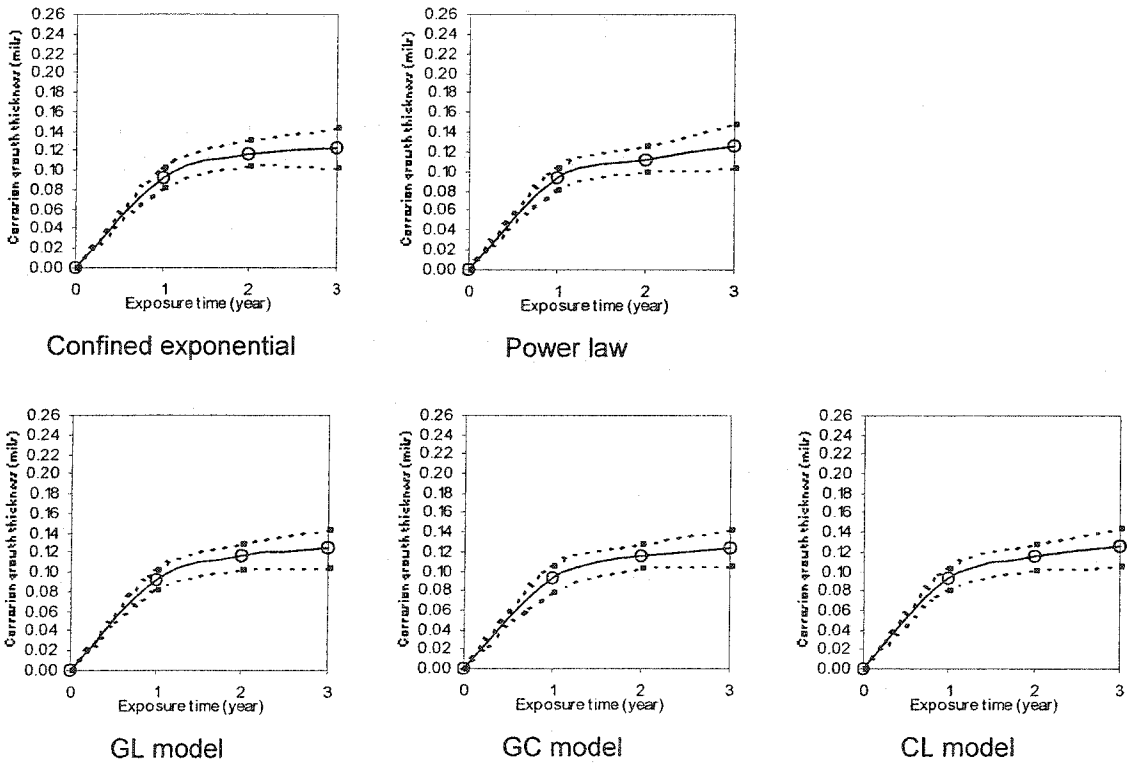


Figure 5.11: Corrosion growth predictive models for Seymour Johnson AFB

CHAPTER 6

CONCLUSIONS AND FUTURE RESEARCH

This chapter provides conclusions and future research. All results from Chapters 4 and 5 are summarized.

6.1 Conclusions

One of the objectives of this research is to identify corrosion severity ranking by location of the six operational air bases. The six operational air force bases included Hickam Air Force Base (AFB) in Hawaii, Kadena Air Base (AB) in Japan, Macdill AFB in Florida, Royal Air Force (RAF) Mildenhall in England, Pease Air National Guard Base (ANGB) in New Hampshire, and Seymour Johnson AFB in North Carolina. These bases were chosen by the USAF and Arinc, Inc. to represent the range of atmospheric conditions C/KC-135 aircraft could be exposed to over their operational life. A corrosion severity ranking scheme for the six operational air bases allows the USAF to concentrate their efforts on proactively inspecting aircraft for corrosion when deployed and operated at highly severe corrosion sites. A corrosion severity ranking scheme for the six operational air bases can also aid future aircraft maintenance programs by prioritizing corrosion inspection and repairs by base.

Atmospheric conditions are clearly important because environment has long been identified as the root-cause of corrosion problems. The atmospheric condition data captured by Arinc, Inc. include air temperature, relative humidity (RH), rain pH, rainfall,

time-of-wetness (TOW), and aircraft's surface temperature. As stated earlier, dew point temperature also plays an important role in atmospheric corrosion and while the dew point temperature was not recorded by Arinc, Inc., it was estimated as a function of air temperature and RH. The methodology for corrosion severity ranking transformed the original data set and the calculated dew point temperature into a compositional data set based on the percentage of 30-minute intervals that an atmospheric variable met a condition conducive to corrosion growth. The percentage of 30-minute intervals is used for corrosion severity ranking analysis of the six air force bases since atmospheric corrosion depends on the length of time that moisture is present on the metal's surface and places all conditions against a common scale for PCA. After obtaining the compositional atmospheric condition data the method of principal component analysis (PCA) was used for the first time to obtain corrosion severity ranking by location of the six operational air bases. The first two principal components were used for corrosion severity ranking analysis since the first two principal components accounted for 85% of the variability of the data and showed that Hickam AFB is the most severe site for corrosion and Pease ANGB is the least severe site for corrosion. The PCA ranking for all six operational air bases ranks from the most severe site to the least severe site as Hickam AFB, Kadena AB, Macdill AFB, Seymour Johnson AFB, RAF Mildenhall, and Pease ANGB.

A predictive model of corrosion growth based on the corrosion thickness loss data gathered from the coupons at each base allows the USAF to look at the rate of corrosion growth by site as a function of time. The foundation for the predictive corrosion growth models developed in this research is based on the known behavior of corrosion growth. That is, the gaseous oxygen interacts with the metal, an oxide is formed on the metal

surface. The oxide forms the surface layer, which protects the metal from further oxidation. If the surface layer reaches a certain thickness, the oxide ceases to grow and the metal passivates. Consequently, while corrosion growth is free initially, it reaches an ultimate limiting value. That is, it cannot corrode more than what amount of metal is initially present or it is stopped by the oxidation process itself. While other models have been developed for metals none have been developed for the alloys used in C/KC-135 operational aircraft or for C/KC-135 alloys that have been "aged" to represent actual operational wear and tear.

The three new predictive corrosion growth models developed by modifying the existing growth models (i.e., the Gompertz growth model and the logistic growth model or GL model, the Gompertz growth model and the confined exponential growth model or the GC model, and the logistic growth model and the confined exponential growth model or the CL model) were compared to the confined exponential growth model and the power law equation. The results showed that of the five models tested, the CL model provides the best fit for all corrosion growth data sets of the four operational air bases and dominates the other models in terms of weighted mean square error (WMSE).

6.2 Future Research

Multiple regression may be an effective approach used to express the relationship between the thickness loss and the eight atmospheric conditions (i.e., air temperature, relative humidity, dew point temperature, rain pH, rainfall, TOW1, TOW2, and aircraft's surface temperature). However, the problem of correlation among independent variables (i.e., multicollinearity) might arise in the independent variables. Multicollinearity will cause the variances to be high. This causes an invalid model while performing multiple

regression. Thus, before performing multiple regression, it is recommended that the problem of multicollinearity should be addressed and alleviated.

Principal component regression (PCR) is an effective approach for handling the problem of multicollinearity (Jackson, 1991). Principal component regression transforms the independent variables into principal components and the dependent variable (i.e., corrosion thickness loss data set) is regressed on the principal components rather than on the original variables (i.e., the numbers of 30-minute intervals exceeding or falling within a threshold for promoting corrosion growth over the thresholds of the atmospheric condition data sets). To overcome the problem of multicollinearity, Khattree and Naik (2000) stated that the last few principal components corresponding to the smallest eigenvalues of the covariance or correlation matrix might be dropped.

However, by using the method of PCR all original independent variables are still present in the multiple regression model. In many applications, it is desirable not only to reduce the number of principal components corresponding to the smallest eigenvalues, but also to reduce the number of the original independent variables. Some original independent variables might not affect the response. It is reasonable to eliminate these ineffectual independent variables. To identify the significant atmospheric conditions influencing corrosion growth, an algorithm for eliminating the ineffectual atmospheric conditions should be developed for supporting the method of PCR. Mansfield et al. (1977) proposed an analytic variable selection algorithm for eliminating independent variables from multiple regression when principal components were removed to adjust for multicollinearity among the independent variables. The procedure of the analytic variable selection algorithm first deleted principal components associated with small eigenvalues and then incorporated a resemblance of the backward elimination procedure

in order to eliminate one or more independent variables. Other proposed algorithms for eliminating ineffectual variables for principal component regression are presented in the articles by McCabe (1984) and Depczynski et al. (2000).

REFERENCES

Afifi, A.A. and Clark, V. (1990), *Computer-Aided Multivariate Analysis*, Second Edition, Chapman & Hall, New York.

Ahamad, B. (1967), "An Analysis of Crimes by the Method of Principal Components", *Applied Statistics*, 16(1), 17-35.

Aitchison, J. (1983), "Principal Component Analysis of Compositional Data", *Biometrika*, 70(1), 57-65.

Anderson, T.W. (1958), *An Introduction to Multivariate Statistical Analysis*, John Wiley & Sons, Inc., New York.

Banks, R.B. (1994), *Growth and Diffusion Phenomena: Mathematical Frameworks and Applications*, Springer-Verlag, New York.

Basu, S. and Reinsel, G.C. (1996), "Relationship Between Missing Data Likehoods and Complete Data Restricted Likelihoods for Regression Time Series Models: an Application to Total Ozone Data", *Applied Statistics*, 45(1), 63-72.

Bates, D.M. and Watts, D.G. (1988), *Nonlinear Regression Analysis and Its Applications*, John Wiley & Sons, Inc., New York.

Bhattacharjee, S., Roy, N., Dey, A.K., and Banerjee, M.K. (1993), "Statistical Appraisal of the Atmospheric Corrosion of Mild Steel", *Corrosion Science,* 34(4), 573-581.

Box, G.E.P. and Jenkins, G.M., and Reinsel, G.C. (1994), *Time Series Analysis: Forecasting and Control*, Third Edition, Prentice Hall Englewood Cliffs, N.J.

Brown, P.W. and Masters, L.W. (1982), "Factors Affecting the Corrosion of Metals in the Atmosphere", *Atmospheric Corrosion*, Ailor, W.H., Ed., John Wiley & Sons, Inc., New York.

Burghes, D.N. and Borrie, M.S. (1981), *Modelling with Differential Equations*, Ellis Horwood Limited, West Sussex, England.

Carroll, R.J. and Ruppert, D. (1988), *Transformation and Weighting in Regression*, Chapman and Hall, New York.

Cattell, E.M. (1966), "The Scree Test for the Number of Factors", *Multivariate Behavior Research*, 1, 245-276.

Chatfield, C. (1984), *The Analysis of Time Series: An Introduction*, Third Edition, Chapman and Hill, London.

Collett, D. and Lewis, T. (1976), " The Subjective Nature of Outlier Rejection Procedures", *Applied Statistics*, 25(3), 228-237.

Cottis, R.A., Qing, L., Owen, G., Gartland, S.J., Helliwell, I.A., and Turega, M. (1999), "Neural Network Methods for Corrosion Data Reduction", *Material and Design*, 20, 169-178.

Dawkins, B. (1989), "Multivariate Analysis of National Track Records", *The American Statistician*, 43(2), 110-115.

Dean, J.A. (1979), *Lange's Handbook of Chemistry*, 12th Edition, McGraw-Hill, New York.

Dean, S.W. (1993), "Classifying Atmospheric Corrosivity--A Challenge for ISO", *Materials Performance*, 30(2), 53-58.

DeGaetano, A.T., Eggleston, K.L., and Knapp, W.W. (1995), "A method to Estimate Missing Daily Maximum and Minimum Temperature Observations", *Journal of Applied Meteorology*, 34(2), 371-380.

Depczynski, U., Frost, V.J., and Molt, K. (2000), "Genetic Algorithms Applied to the Selection of Factors in Principal Component Regression", *Analysis Chimica Acta*, 420, 217-227.

Draper, N.R. and Smith, H. (1998), *Applied Regression Analysis*, Third Edition, John Wiley & Sons, Inc., New York.

Dreyer, T.P. (1993), *Modelling with Ordinary Differential Equations*, CRC Press, Inc., Boca Raton, Florida.

Feinberg, A.A., Gibson, G.J., White, J.V., and Briggs Jr., R.E. (1994), "A Corrosion Simulation Environment for Maintenance of Aging Aircraft", *Proceedings: Institute of Environmental Sciences*, 198-210.

Feliu, S., Morcillo, M., and Feliu, S. Jr. (1993a), "The Prediction of Atmospheric Corrosion from Meteorological and Pollution Parameters—I. Annual Corrosion", *Corrosion Science*, 34(3), 403-414.

Feliu, S., Morcillo, M., and Feliu, S. Jr. (1993b), "The Prediction of Atmospheric Corrosion from Meteorological and Pollution Parameters—II. Long-Term Forecasts", *Corrosion Science*, 34(3), 415-422.

Ferrer, K.S. and Kelly, R.G. (2002), "Development of an Aircraft Lap Joint Simulant Environment", *Corrosion*, 58(5), 452-459.

Fuller, W.A. (1996), *Introduction to Statistical Time Series*, Second Edition, John Wiley & Son, New York.

Greene, W.H. (1993), *Econometric Analysis*, Second Edition, Prentice Hall, Englewood Cliffs, New Jersey.

Groner, D.J. and Nieser, D.E. (1996), "U.S. Air Force Aging Aircraft Corrosion", *Canadian Aeronautics and Space Journal*, 42(2), 63-67.

Hagemaier, D.J., Wendelbo, A.H., and Bar-Cohen, Y. (1985), "Aircraft Corrosion and Detection Methods", *Materials Evaluation*, 43, 426-437.

Haykin, S. (1999), *Neural Networks A Comprehensive Foundation*, Second Edition, Prentice Hall, New Jersey.

Haynie, F.H., Spence, J.W., and Upham, J.B. (1978), "Effects of Air Pollutants on Weathering Steel and Galvanized Steel: A Chamber Study", *Atmospheric Factors Affecting the Corrosion of Engineering Metals, ASTM STP 646*, Coburn, S.K., Ed., American Society for Testing and Materials, 30-47.

Howard, M., Schubert, L.E., and Tietz, H.M. (1999), *First Interim Report: Corrosion Growth Rates Testing*, ARINC, Inc., Oklahoma City, Oklahoma.

Hotelling, H. (1933), "Analysis of a Complex of Statistical Variables into Principal Components", *Journal of Educational Psychology*, 24, 417-441.

Jackson, J.E. (1980), "Principal Components and Factor Analysis: Part I—Principal Components", *Journal of Quality Technology*, 12(4), 201-213.

Jackson, J.E. (1991), *A User's Guide to Principal Components*, John Wiley & Sons, Inc., New York.

Jones , R.H. (1980), "Maximum Likelihood Fitting of ARMA Models to Time Series With Missing Observations", *Technometrics*, 22, 389-395.

Kajiyama, F. and Koyama, Y. (1997), "Statistical Analyses of Field Corrosion Data for Ductile Cast Iron Pipes Buried in Sandy Marine Sediments", *Corrosion*, 53(2), 156-162.

Kemp, W.P. Burnell, D.G., Everson, D.O., and Thomson, A.J. (1983), "Estimating Missing Daily Maximum and Minimum Temperatures", *Journal of Climate and Applied Meteorology*, 22, 1587-1593.

Khattree, R. and Naik, D.N. (1999), *Applied Multivariate Statistics with SAS® Software*, Second Edition, SAS® Institue Inc., Cary, North Carolina.

Khattree, R. and Naik, D.N. (2000), *Multivariate Data Reduction and Discrimination with SAS® Software*, SAS® Institue Inc., Cary, North Carolina.

Klink, K. and Willmont, C.J. (1989), "Principal Components of the Surface Wind Field in the United States: A Comparison of Analyses Based Upon Wind Velocity, Direction, and Speed", *International Journal of Climatology*, 9, 293-308.

Kucera, V. and Mattsson, E. (1987), "Atmospheric Corrosion", *in Corrosion Mechanisms*, Marcel Dekker, New York.

Kutzbach, J.E. (1967), "Empirical Eigenvectors of Sea-Level Pressure, Surface Temperature and Precipitation Complexes Over North America", *Journal of Applied Meteorology*, 6, 791-802.

Lawson, H.H. (1995), "Atmospheric Corrosion Test Methods", Vol. 4, *NACE International*, Houston, Texas, 1-3.

Legault, R.A. and Dalal, J.G. (1982), "The Statistical Analysis of Atmospheric Corrosion Data", *Atmospheric Corrosion*, Ailor, W.H., Ed., John Wiley & Sons, Inc., New York.

Ljung, G.M. (1993), "On Outlier Detection in Time Series", *Journal of the Royal Statistical Society, Series B*, 55(2), 559-567.

Mansfields, E.R., Webster, J.T., and Gunst, R.F. (1977), "An Analytic Variable Selection Technique for Principal Component Regression", *Applied Statistics*, 26(1), 34-40.

Marquardt, D.W. (1963), "An Algorithm for Least-Squares Estimation of Nonlinear Parameter", *Journal of Industrial and Application Mathematics*, 11(2), 431-441.

Matlab® Neural Network Toolbox (2000), the Mathworks Inc., Version 6.

McCabe, G.P. (1984), "Principal Variables", *Technometrics*, 26(2), 137-144.

Mikhailovsky, Y.N. (1982), "Theoretical and Engineering Principles of Atmospheric Corrosion of Metals", *Atmospheric Corrosion*, Ailor, W.H., Ed., John Wiley & Sons, Inc., New York.

Mizaki, R., Draper, N.R., and Johnson, R.A. (1973), "On the Violation of Assumptions in Nonlinear Least Squares by Interchange of Response and Predictor Variables", *Industrial and Engineering Chemistry Fundamentals*, 12(2), 251-254.

Naik, D.N. and Khattree, R. (1996), "Revisiting Olympic Track Records: Some Practical Considerations in the Principal Component Analysis", *The American Statistician*, 50(2), 140-144.

Perry, R.H. and Green, D.W. (1997), *Perry's Chemical Engineers' Handbook*, Seventh Edition, McGraw-Hill, New York.

Pourbaix, M. (1982), "The Linear Bilogarithmic Law for Atmospheric Corrosion", *Atmospheric Corrosion*, Ailor, W.H., Ed., John Wiley & Sons, Inc., New York.

Preisendorfer, R.W. (1988), *Principal Component Analysis in Meteorology and Oceanography*, Mobley, C.D, Ed., Elsevier, Amsterdam.

Ratinger, T. (1996), "Seasonal Time Series with Missing Observations", *Applications of Mathematics*, 41(1), 41-55.

Richards, F.J. (1959), "A Flexible Growth Function for Empirical Use", *Journal of Experimental Botany*, 10(29), 290-300.

Roberge, P.R. (2000), *Handbook of Corrosion Engineering*, McGraw-Hill, New York.

Robinson (1979), "Comment to Klieiner, Martin, and Thomson: Robust Estimation of Power Spectra", *Journal of the Royal Statistical Society Series B*, 41(3), 313-351.

Ryan, T.P. (1997), *Modern Regression Methods*, John Wiley & Son Inc., New York.

SAS/STAT® (1990), *User's Guide*, Version 6, Fourth Edition, Vol.2, Cary, North Carolina.

Schweitzer, P.A. (1991), *Corrosion Resistance Tables: Metals, Nonmetals, Coating, Mortars, Plastics, Elastomers and Linings and Fabrics*, Third Edition, Revised and Expanded, Part B, J-Z, Marcel Dekker, Inc., New York.

Schweitzer, P.A. (1998), *Encyclopedia of Corrosion Technology*, Marcel Dekker, Inc., New York.

Sereda, P.J. (1960), "Atmospheric Factors Affecting the Corrosion of Steel", *Industrial And Engineering Industry*, 52(2), 157-160.

Solus Systems (1994), *User's Manual*, Version 1.5, Portland, Oregon.

Spezzaferro, K.E. (1996), "Applying Logistic Regression to Maintenance Data to Establish Inspection Intervals", IEEE: Proceedings Annual Reliability and Maintainability *Symposium*, 296-300.

Spezzaferro, K.E., Gibson, G.J., Feinberg, A.A., and Peacock, W.J. (1995), "Analysis of Air Force Corrosion Data to Develop Corrosion Maintenance Inspection Strategies", *Structural Integrity in Aging Aircraft ASME*, 47, 213-219.

Srivastava, M.S. (2002), *Methods of Multivariate Statistics*, John Wiley & Son Inc., New York.

Stidd, C.K. (1967), "The Use of Eigenvectors for Climatic Estimates", *Journal of Applied Meteorology*, 6, 255-264.

Tang Z., Almeida, C.D., and Fishwick, P.A. (1991), "Time Series Forecasting Using Neural Networks vs. Box-Jenkins Methodology", *Simulation*, 57(5), 303-310.

Velicer, W.F. (1976), "Determining the Number of Components from the Matrix of Partial Correlations", *Psychometric*, 41(3), 321-327.

Wallace, W., Hoeppner, D.W., and Kandachar, P.V. (1985), "AGARD Corrosion Handbook", Vol. 1, *Aircraft Corrosion: Causes and Case Histories*, Essex, England.

Weihs, C. (1993), "Multivariate Exploratory Data Analysis and Graphics: a Tutorial", *Journal of Chemometrics*, 7, 306-340.

Wilks, D.S. (1995), *Statistical Methods in the Atmospheric Sciences*, Academic Press, San Diego.

Yano, Y., Oguma, T., Nagata, H., and Sasaki, S. (1998), "Application of Logistic Growth Model to Phamacodynamic Analysis of in Vitro Bactericidal Kinetics", *Journal of Pharmaceutical Sciences*, 87(10), 1177-1183.

Young, F.W., Takane, Y. and De Leeuw, J. (1978), "The Principal Components of Mixed Measurement Level Multivariate Data: An Alternating Least Squares Method Optimal Scaling Features", *Psychometrika*, 43(2), 279-281.

APPENDIX A

Predictive Corrosion Growth Models for Kadena AB, RAF Mildenhall, and Seymour Johnson AFB

Kadena AB

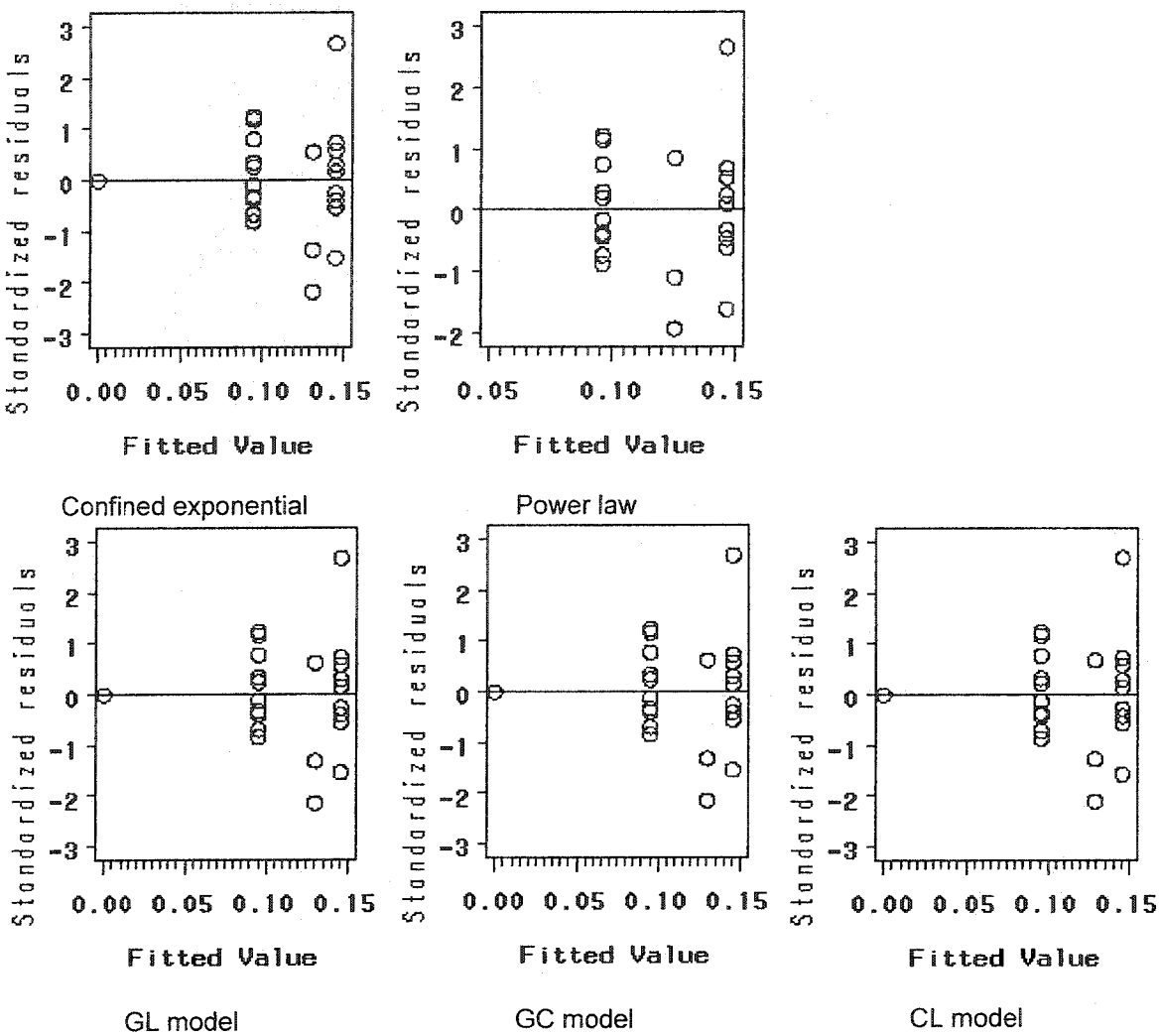


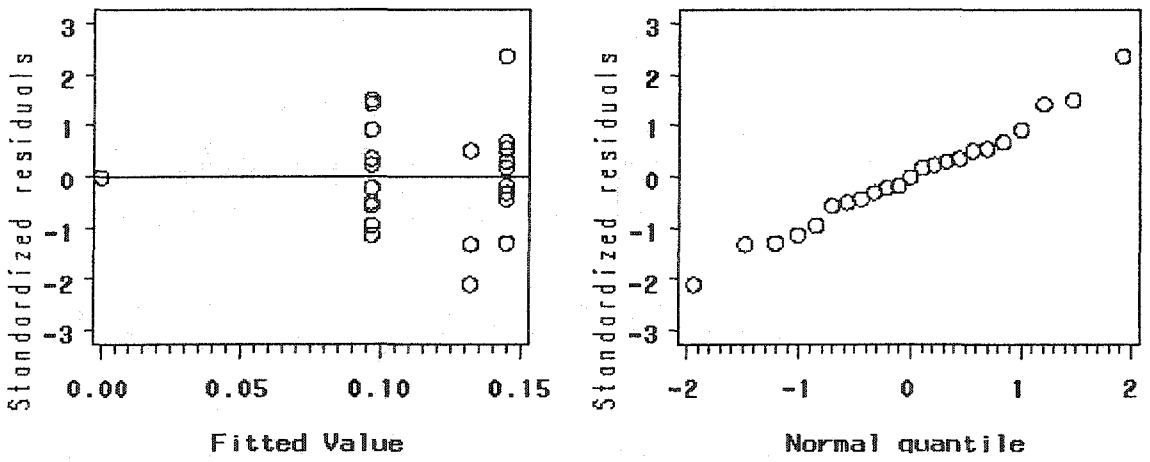
Figure A1: Residual analysis from the results of the five models using OLS for Kadena

Table A1: Lack-of-fit test for Confined Exponential model for Kadena AB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	2	623.2000	311.6000	368.0315	< 0.0001
Lack of Fit	2	2.6274	1.3137	1.6473	0.219
Pure Error	19	15.1526	0.7975		
Total Error	21	17.7841	0.8467		

Table A2: Model parameter estimates for Confined Exponential model for Kadena AB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.1520	0.0117	0.1277	0.1763
a_0	1.0121	0.1881	0.6209	1.4033



Residuals against fitted values plot

Normal probability plot

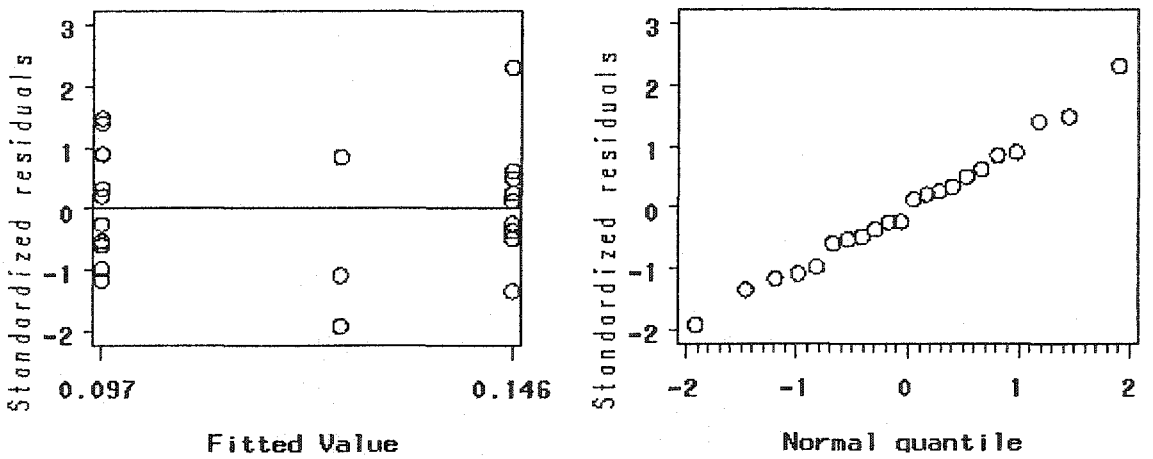
Figure A2: Model adequacy checking for Confined Exponential model for Kadena AB data with weighted least squares

Table A3: Lack-of-fit test for Power Law model for Kadena AB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	2	623.1000	311.5500	391.5725	< 0.0001
Lack of Fit	2	1.4324	0.7162	0.8908	0.43
Pure Error	19	15.2760	0.8040		
Total Error	21	16.7084	0.7956		

Table A4: Model parameter estimates for Power Law model for Kadena AB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.0972	0.00524	0.0863	0.1081
a_0	0.3703	0.0714	0.2214	0.5192



Residuals against fitted values plot

Normal probability plot

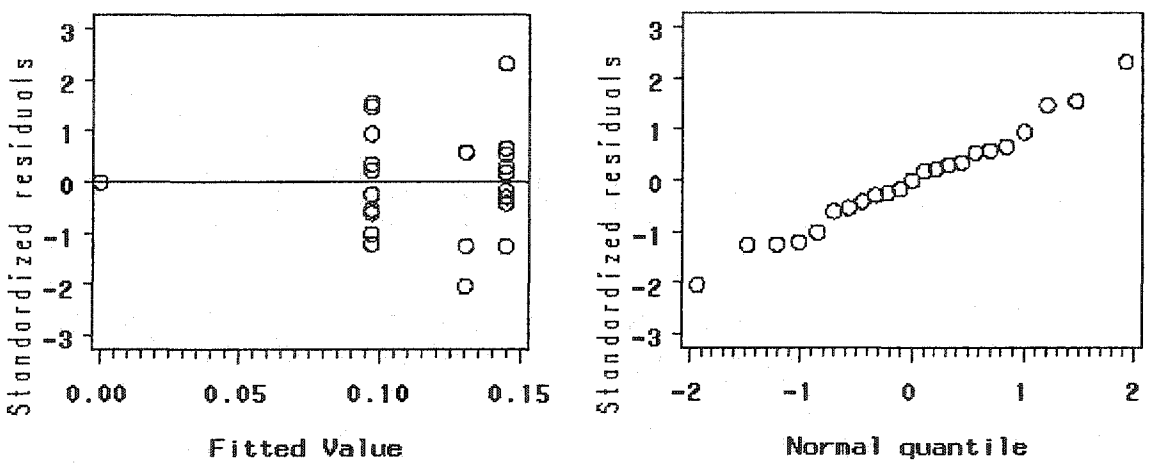
Figure A3: Model adequacy checking for Power Law model for Kadena AB data with weighted least squares

Table A5: Lack-of-fit test for GL model for Kadena AB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	3	402.00	134.00	251.47	< 0.0001
Lack of Fit	2	1.4483	0.7242	1.4124	0.268
Pure Error	19	9.7417	0.5127		
Total Error	21	11.1900	0.5329		

Table A6: Model parameter estimates for GL model for Kadena AB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.1885	0.0353	0.1150	0.2620
a_0	0.1710	0.0248	0.1194	0.225
k	0.4693	Infinity	-Infinity	Infinity



Residuals against fitted values plot

Normal probability plot

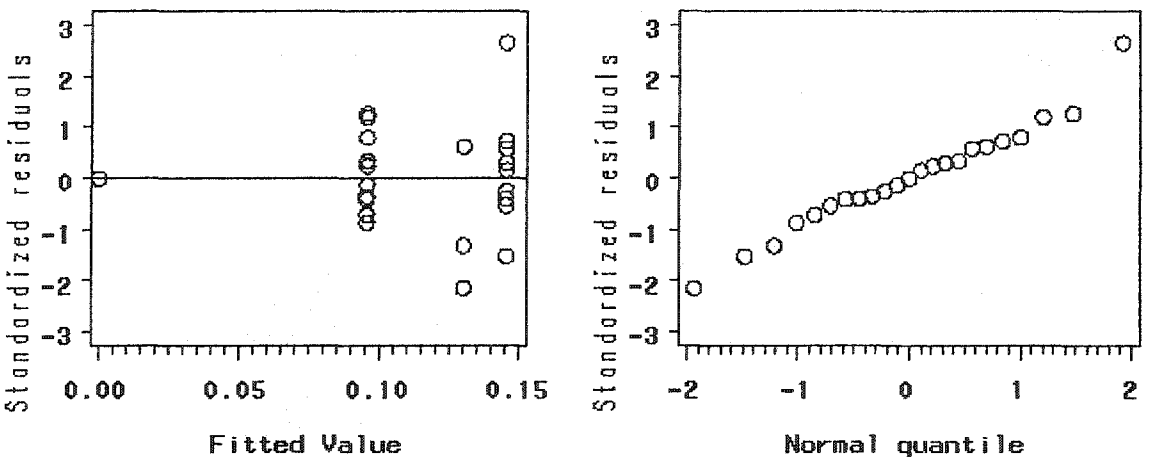
Figure A4: Model adequacy checking for GL model for Kadena AB data with weighted least squares

Table A7: Lack-of-fit test for GC model for Kadena AB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	3	226.10	75.367	142.0736	< 0.0001
Lack of Fit	2	1.5966	0.7983	1.5893	0.23
Pure Error	19	9.5434	0.5023		
Total Error	21	11.1400	0.5305		

Table A8: Model parameter estimates for GC model for Kadena AB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.1937	0.0362	0.1184	0.2689
a_0	0.8478	Infinity	-Infinity	Infinity
k	0.4499	0.3939	-0.3693	1.2691



Residuals against fitted values plot

Normal probability plot

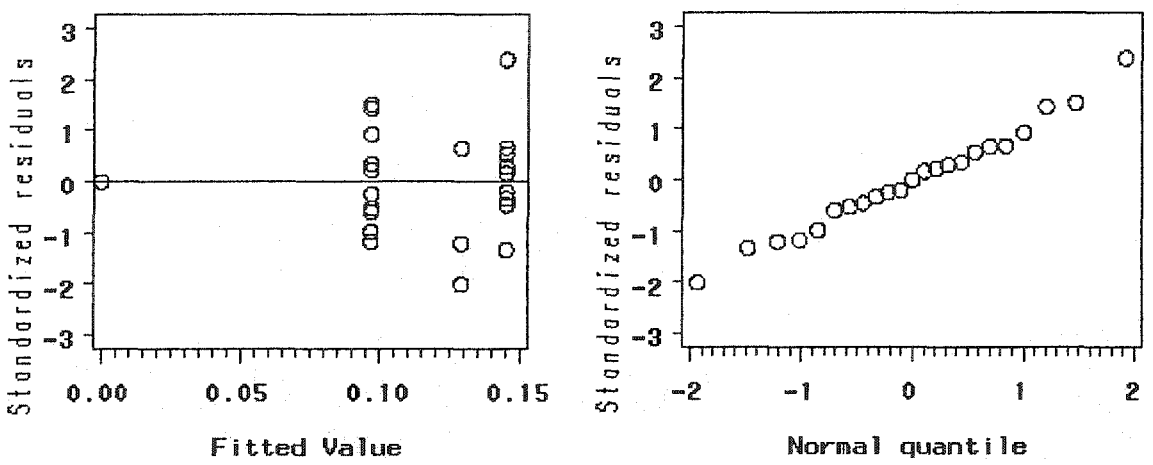
Figure A5: Model adequacy checking for GC model for Kadena AB data with weighted least squares

Table A9: Lack-of-fit test for CL model for Kadena AB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	2	395.1000	197.5500	381.9957	< 0.0001
Lack of Fit	2	1.2672	0.6336	1.2549	0.31
Pure Error	19	9.5930	0.5049		
Total Error	21	10.8602	0.5172		

Table A10: Model parameter estimates for CL model for Kadena AB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.1929	0.0232	0.1447	0.2411
a_0	1.0146	0.3137	0.3622	1.6671



Residuals against fitted values plot

Normal probability plot

Figure A6: Model adequacy checking for CL model for Kadena AB data with weighted least squares

RAF Mildenhall

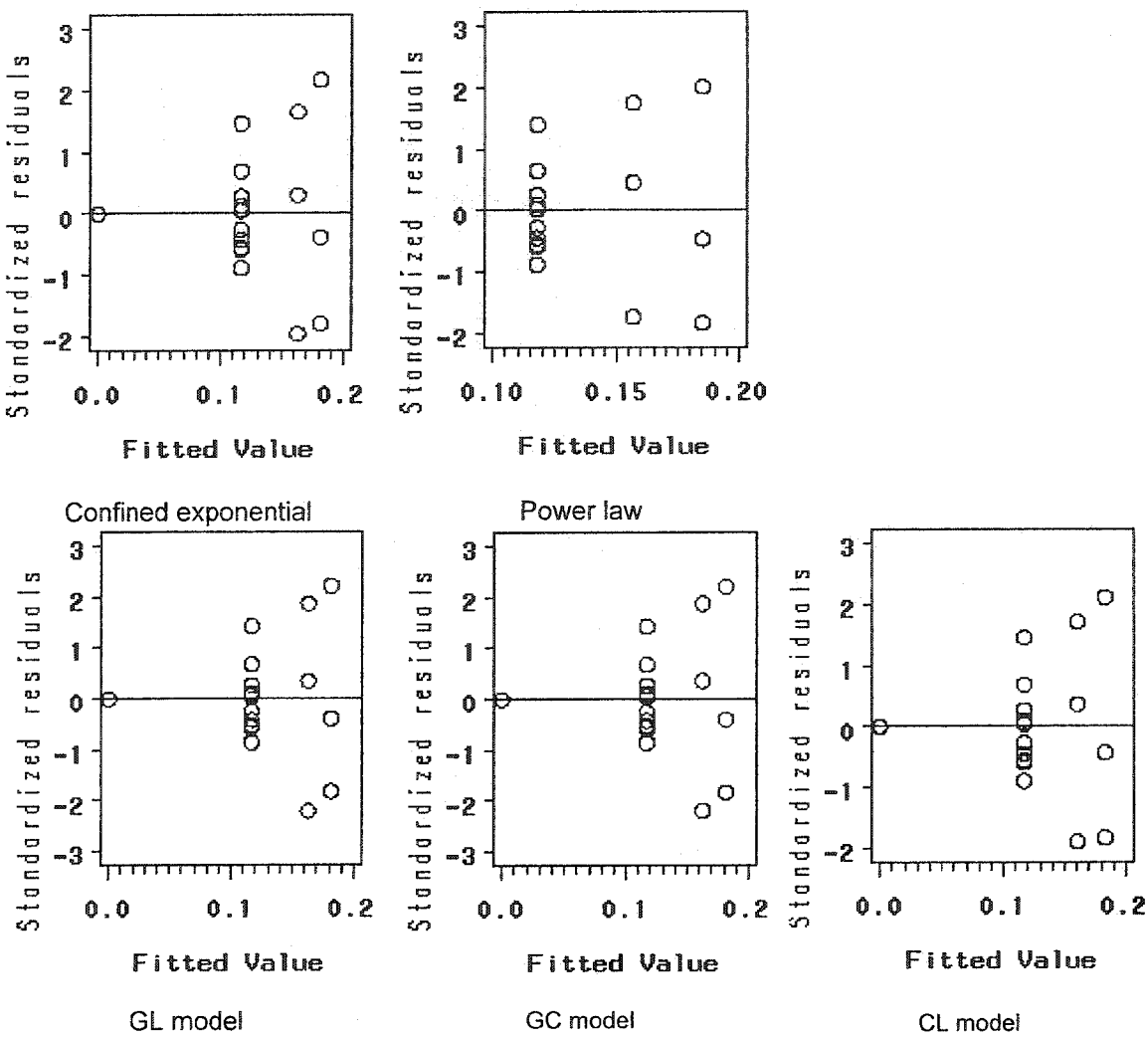


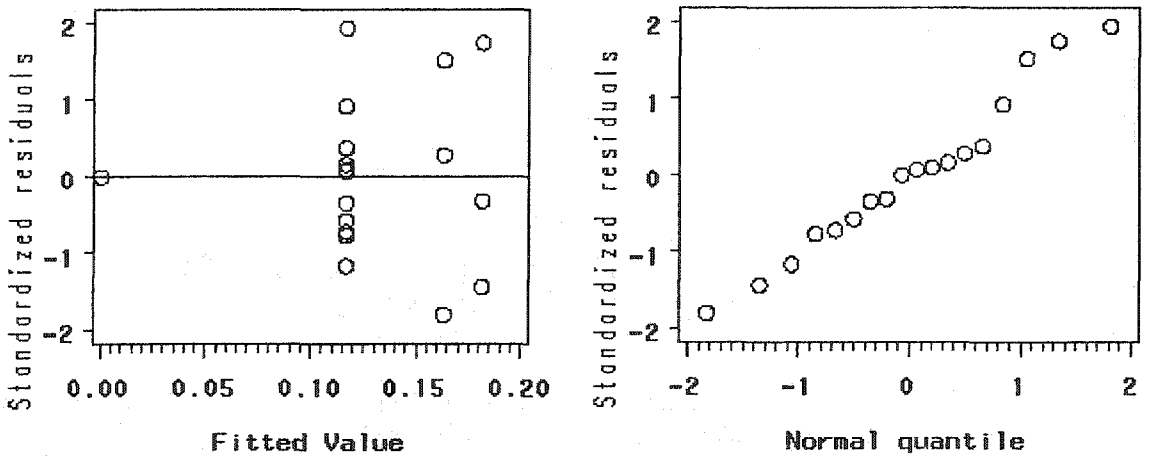
Figure A7: Residual analysis from the results of the five models using OLS for RAF Mildenhall

Table A11: Lack-of-fit test for Confined Exponential model for RAF Mildenhall data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	2	119.3000	59.6500	111.9465	< 0.0001
Lack of Fit	2	0.0024	0.0012	0.0020	0.99
Pure Error	14	8.5231	0.6088		
Total Error	16	8.5255	0.5328		

Table A12: Model parameter estimates for Confined Exponential model for RAF Mildenhall data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.1940	0.0401	0.1089	0.2791
a_0	0.9239	0.3669	0.1461	1.7016



Residuals against fitted values plot

Normal probability plot

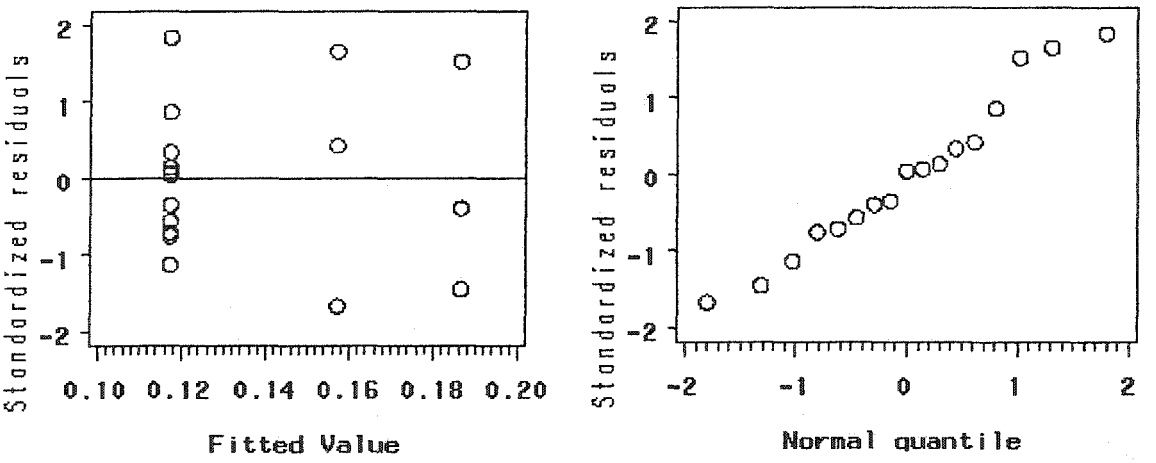
Figure A8: Model adequacy checking for Confined Exponential model for RAF Mildenhall data with weighted least squares

Table A13: Lack-of-fit test for Power Law model for RAF Mildenhall data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	2	119.3000	59.6500	110.1608	< 0.0001
Lack of Fit	2	0.0478	0.0239	0.0388	0.96
Pure Error	14	8.6159	0.6154		
Total Error	16	8.6637	0.5415		

Table A14: Model parameter estimates for Power Law model for RAF Mildenhall data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.1174	0.00991	0.0963	0.1385
a_0	0.4216	0.1587	0.0833	0.7599



Residuals against fitted values plot

Normal probability plot

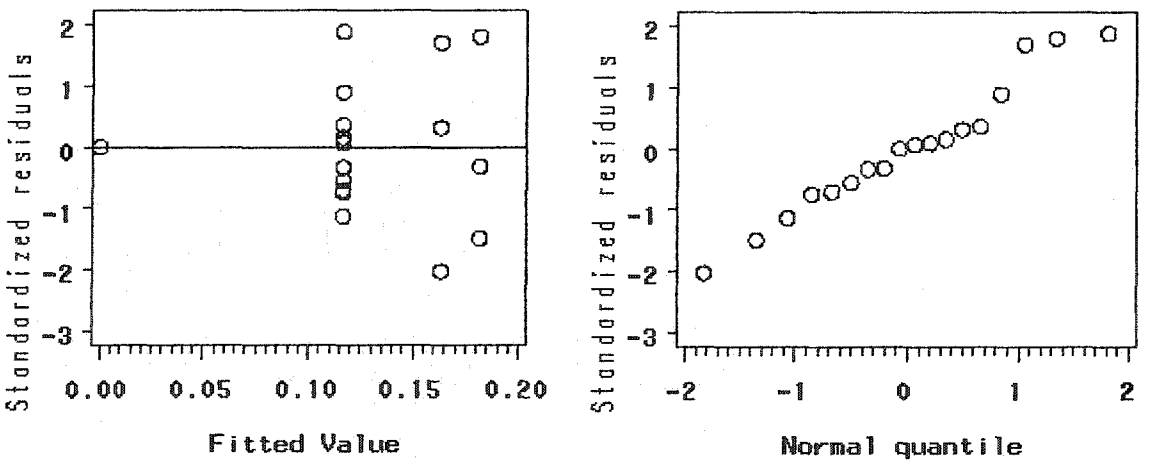
Figure A9: Model adequacy checking for Power Law model for RAF Mildenhall data with weighted least squares

Table A15: Lack-of-fit test for GL model for RAF Mildenhall data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	3	119.300	39.7667	74.5916	< 0.0001
Lack of Fit	2	0.0052	0.0026	0.0043	0.996
Pure Error	14	8.5248	0.6089		
Total Error	16	8.53	0.5331		

Table A16: Model parameter estimates for GL model for RAF Mildenhall data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.1936	0.1989	-0.2304	0.6177
a_0	0.1789	0.1960	-0.2388	0.5966
k	-0.00612	3.3990	-7.2509	7.2387



Residuals against fitted values plot

Normal probability plot

Figure A10: *Model adequacy checking for GL model* for RAF Mildenhall data with weighted least squares

Table A17: Lack-of-fit test for GC model for RAF Mildenhall data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	2	-	-	-	-
Lack of Fit	2	-	-	-	-
Pure Error	14	-	-		
Total Error	16	-	-		

Note: The results do not exist since the iterative estimation procedure failed to converge

Table A18: Model parameter estimates for GC model for RAF Mildenhall data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	-	-	-	-
a_0	-	-	-	-
k	-	-	-	-

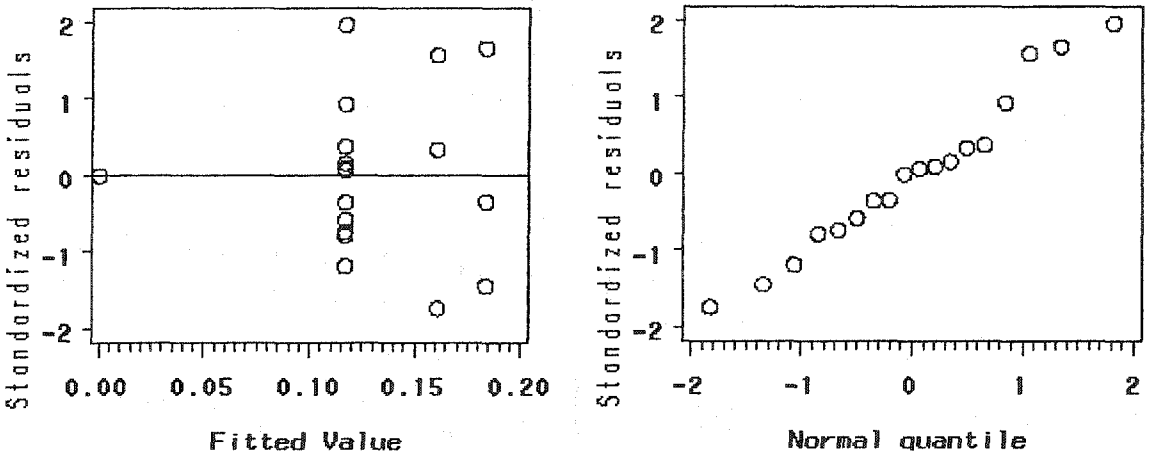
Note: The results do not exist since the iterative estimation procedure failed to converge

Table A19: Lack-of-fit test for CL model for RAF Mildenhall data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	*Pr* > *F*
Regression	2	72.1143	36.0572	113.0761	< 0.0001
Lack of Fit	2	0.0059	0.0030	0.0081	0.99
Pure Error	14	5.0961	0.3640		
Total Error	16	5.1020	0.3189		

Table A20: Model parameter estimates for CL model for RAF Mildenhall data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.2568	0.0800	0.0873	0.4263
a_0	0.8388	0.5379	-0.3014	1.9790



Residuals against fitted values plot

Normal probability plot

Figure A11: Model adequacy checking for CL model for RAF Mildenhall data with weighted least squares

Seymour Johnson AFB

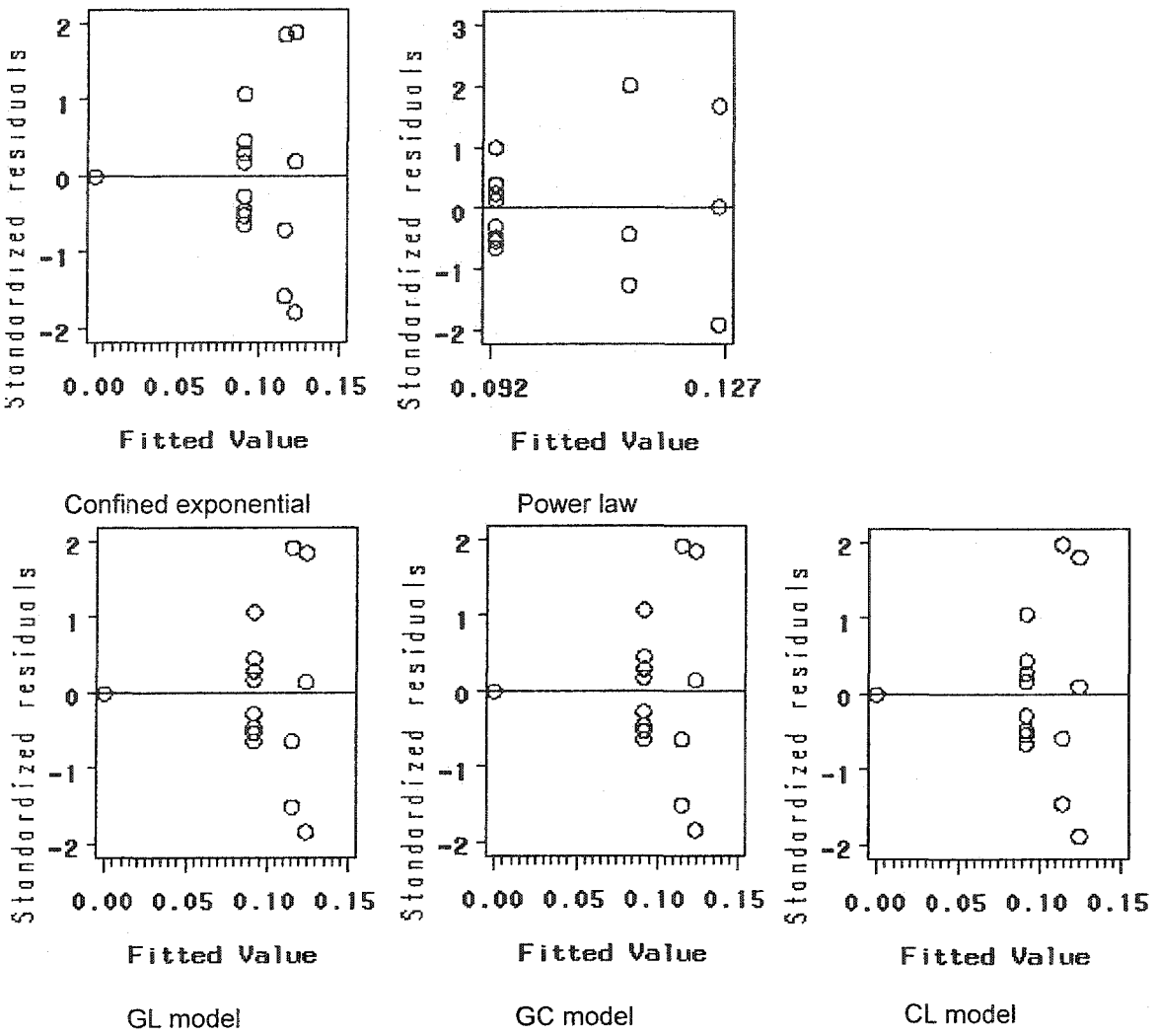


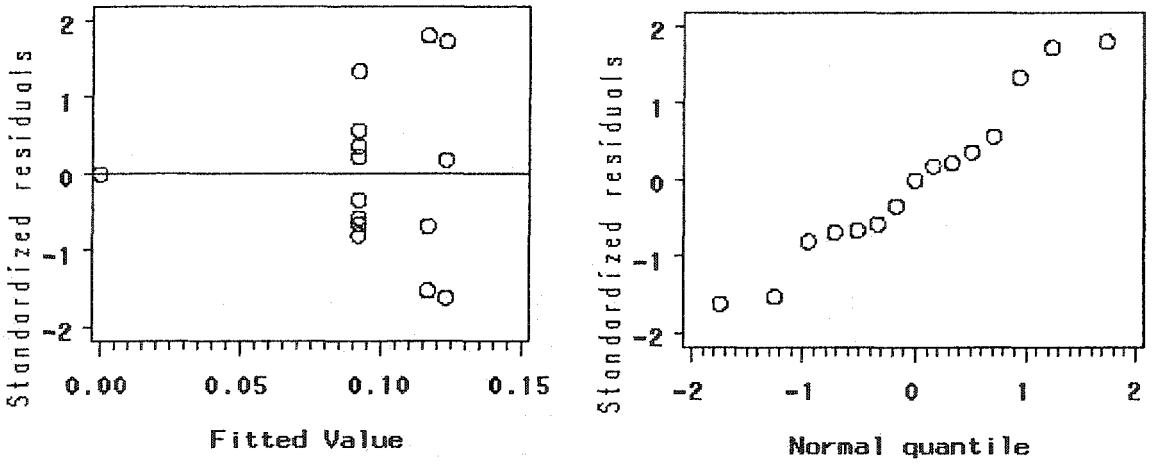
Figure A12: Residual analysis from the results of the five models using OLS for Seymour Johnson AFB

Table A21: Lack-of-fit test for Confined Exponential model for Seymour Johnson AFB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	2	321.8000	160.9000	292.6928	< 0.0001
Lack of Fit	2	0.0497	0.0248	0.0385	0.96
Pure Error	11	7.0967	0.6452		
Total Error	13	7.1464	0.5497		

Table A22: Model parameter estimates for Confined Exponential model for Seymour Johnson AFB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.1258	0.0114	0.1011	0.1505
a_0	1.3203	0.3155	0.6386	2.0019



Residuals against fitted values plot

Normal probability plot

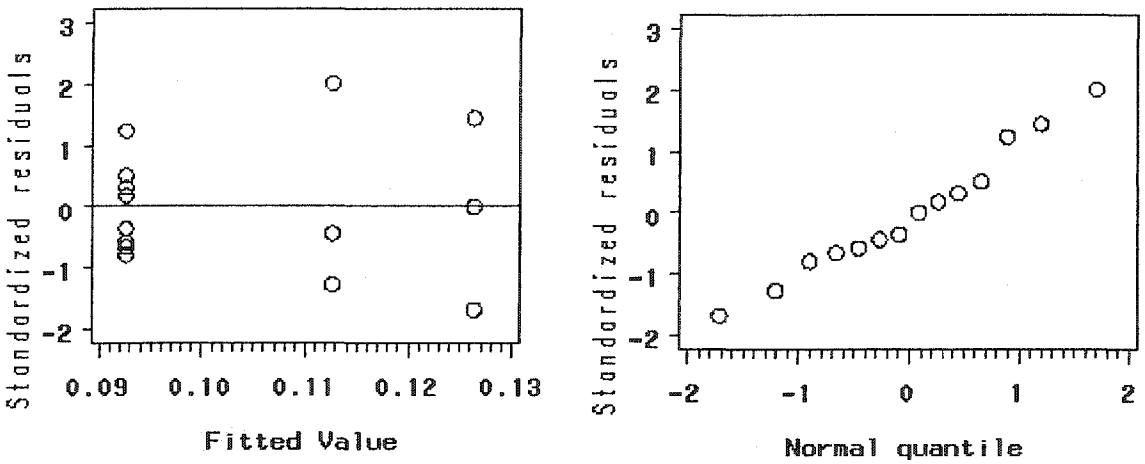
Figure A13: Model adequacy checking for Confined Exponential model for Seymour Johnson AFB data with weighted least squares

Table A23: Lack-of-fit test for Power Law model for Seymour Johnson AFB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	2	321.9000	160.9500	289.2303	<0.0001
Lack of Fit	2	0.0354	0.0177	0.0270	0.97
Pure Error	11	7.1988	0.6544		
Total Error	13	7.2342	0.5565		

Table A24: Model parameter estimates for Power Law model for Seymour Johnson AFB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.0926	0.00516	0.0813	0.1038
a_0	0.2832	0.0941	0.0782	0.4883



Residuals against fitted values plot

Normal probability plot

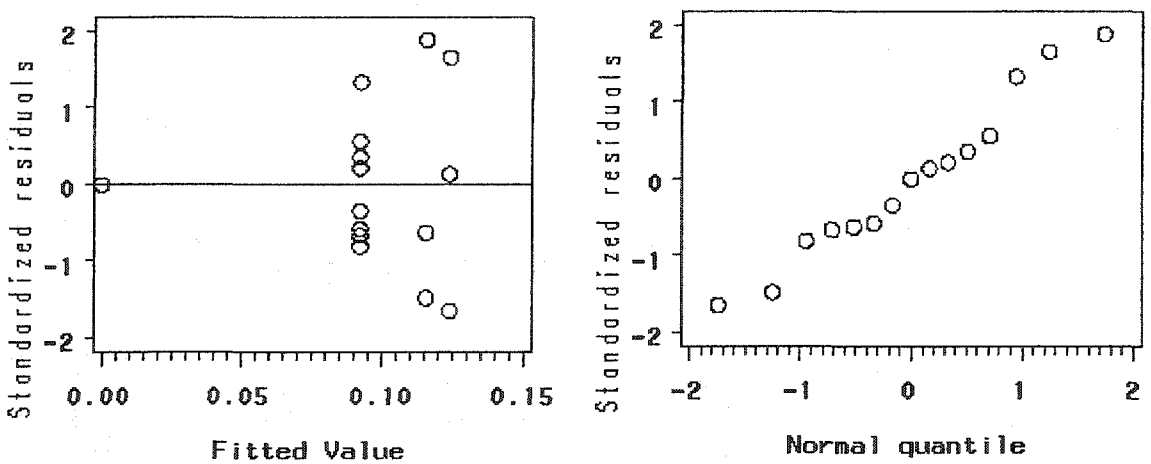
Figure A14: Model adequacy checking for Power Law model for Seymour Johnson AFB data with weighted least squares

Table A25: Lack-of-fit test for GL model for Seymour Johnson AFB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	3	321.900	107.30	195.3641	< 0.0001
Lack of Fit	2	0.0175	0.0088	0.0135	0.99
Pure Error	11	7.1225	0.6475		
Total Error	13	7.1400	0.5492		

Table A26: Model parameter estimates for GL model for Seymour Johnson AFB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.1477	0.0305	0.0818	0.2136
a_0	0.1908	0.0388	0.1070	0.2745
k	0.5794	Infinity	-Infinity	Infinity



Residuals against fitted values plot

Normal probability plot

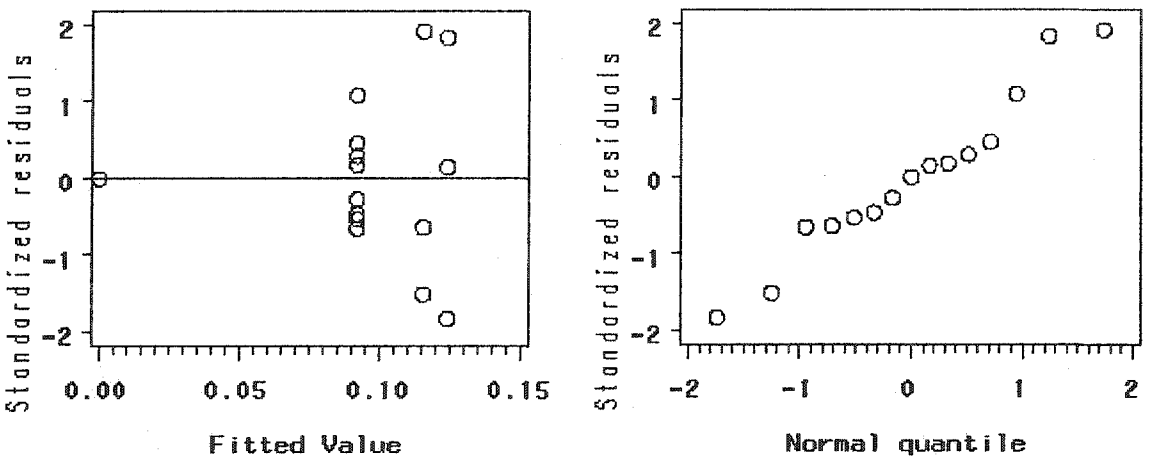
Figure A15: Model adequacy checking for GL model for Seymour Johnson AFB data with weighted least squares

Table A27: Lack-of-fit test for GC model for Seymour Johnson AFB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	3	201.00	67.0000	166.2214	< 0.0001
Lack of Fit	2	0.0159	0.0080	0.0167	0.98
Pure Error	11	5.2241	0.4749		
Total Error	13	5.2400	0.4031		

Table A28: Model parameter estimates for GC model for Seymour Johnson AFB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.1480	0.0305	0.0821	0.2139
a_0	1.2844	0.5427	0.1121	2.4568
k	0.5776	Infinity	-Infinity	Infinity



Residuals against fitted values plot

Normal probability plot

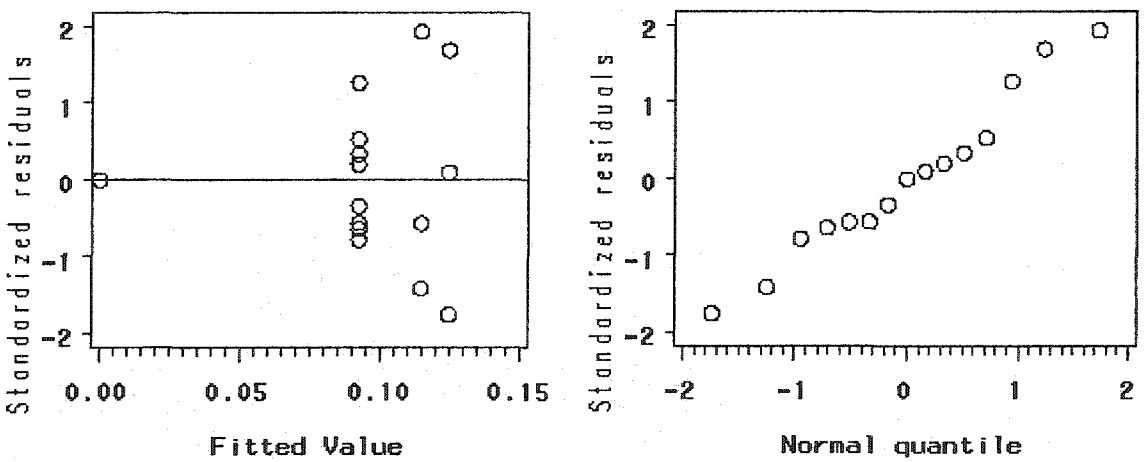
Figure A16: Model adequacy checking for GC model for Seymour Johnson AFB data with weighted least squares

Table A29: Lack-of-fit test for CL model for Seymour Johnson AFB data with weighted least squares

Source	DF	Sum of Squares	Mean Square	*F* Value	Pr > *F*
Regression	2	192.4000	96.2000	282.8004	< 0.0001
Lack of Fit	2	0.0064	0.0032	0.0080	0.99
Pure Error	11	4.4158	0.4014		
Total Error	13	4.4222	0.3402		

Table A30: Model parameter estimates for CL model for Seymour Johnson AFB data with weighted least squares

Parameter	Estimate	Approximate Std Error	Approximate 95% Confidence Limits	
P^*	0.1516	0.0220	0.1041	0.1991
a_0	1.5603	0.7013	0.0452	3.0753



Residuals against fitted values plot

Normal probability plot

Figure A17: Model adequacy checking for CL model for Seymour Johnson AFB data with weighted least squares

APPENDIX B

Weighted Least Squares Calculation

Weighted Least Squares Procedures

Step 1: Plot graph relationship between average and standard deviation of each exposure time

Step 2: Determine expected value (μ_u) at each exposure time from

$$\mu_u = f(x_u, \theta)$$

Step 3: Determine variance function (g), which is defined as the function of expected value

Step 4: Determine the weights defined as

$$w_u = \frac{1}{g^2(\mu_u(\theta), \theta)}$$

Step 5: Fit the growth model with the weights to the corrosion growth data set

Step 6: Update the parameter estimates, weighted sum of squares, and repeat Steps 2-5 until the weighted sum of squares approaches the minimum value. If the model adequacy checking indicates constant variance, obtain the optimum parameter estimates.

Results

Hickam

1. The Confined Exponential Model Fitting to the Corrosion Growth Data Set of Hickam AFB

START

Iteration 0 Running SAS® to obtain the initial parameters without considering the weights.

Iteration 1
Step 1:

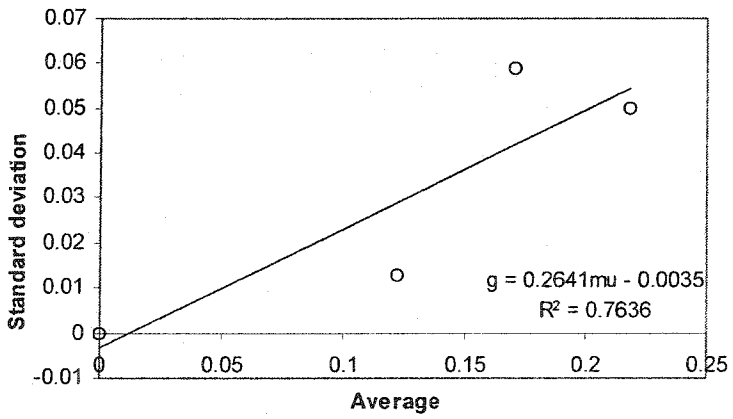


Figure B1: Standard deviation as the function of average for Hickam AFB

Steps 2, 3, and 4:

ConfinedExpo	P=0.2613(1-exp(-0.5866u))			
u	$Mu(u) = P$	$g = 0.2641*Mu-0.0035$	g^2	$w(u)=1/g$^2)
0	0	-0.0035	0.00001225	81632.65
1	0.11596097	0.02712529	0.000735781	1359.10
2	0.18046022	0.04415954	0.001950065	512.80
3	0.21633569	0.05363425	0.002876633	347.63

Step 5:

```
data growth;
  input year growthdepth @@; cards;
0 0 1 0.1323 ... 3 0.2471 3 0.3076
; /*Hickam and Mildenhall combined data*/
proc nlin data = growth method = marquardt;
 parms Pstar=0.1 a0=0.1;
 model growthdepth = Pstar*(1-exp(-a0*year));
       if year = 0 then _weight_ = 81632.65; if year = 1 then _weight_ = 1359.10;
       if year = 2 then _weight_ = 512.80; if year = 3 then _weight_ = 347.63;
run;
```

Step 6:

Iteration	Weighted SS	Pstar	$a0$
0	-	0.2613	0.5866
1	19.07	0.2502	0.6496

Iteration 2

Steps 2, 3, and 4:

ConfinedExpo	P=0.2502(1-exp(-0.6496u))			
u	$Mu(u) = P$	$g = 0.2641*Mu-0.0035$	g^2	$w(u)=1/g$^2)
0	0	-0.0035	0.00001225	81632.65
1	0.11953189	0.02806837	0.000787834	1269.30
2	0.18195797	0.0445551	0.001985157	503.74
3	0.21456029	0.05316537	0.002826557	353.79

Step 5:

```
proc nlin data = growth method = marquardt;
 parms Pstar=0.1 a0=0.1;
 model growthdepth = Pstar*(1-exp(-a0*year));
       if year = 0 then _weight_ = 81632.65; if year = 1 then _weight_ = 1269.30;
       if year = 2 then _weight_ = 503.74; if year = 3 then _weight_ = 353.79;
run;
```

Step 6:

Iteration	Weighted SS	Pstar	$a0$
0	-	0.2613	0.5866
1	19.07	0.2502	0.6496
2	19.01	0.2508	0.6465

Iteration 3

Steps 2, 3, and 4:

ConfinedExpo	$P=0.2508(1-\exp(-0.6465u))$			
u	*Mu(u) = P*	*g = 0.2641*Mu-0.0035*	*g^2*	*w(u)=1/g^2)*
0	0	-0.0035	0.00001225	81632.65
1	0.11941186	0.02803667	0.000786055	1272.18
2	0.18196889	0.04455798	0.001985414	503.67
3	0.21474102	0.0532131	0.002831634	353.15

Step 5:

```
proc nlin data = growth method = marquardt;
 parms Pstar=0.1 a0=0.1;
 model growthdepth = Pstar*(1-exp(-a0*year));
       if year = 0 then _weight_ = 81632.65; if year = 1 then _weight_ = 1272.18;
       if year = 2 then _weight_ = 503.67; if year = 3 then _weight_ = 353.15; run;
```

Step 6:

Iteration	Weighted SS	*Pstar*	*a0*
0	.	0.2613	0.5866
1	19.07	0.2502	0.6496
2	19.01	0.2508	0.6465
3	18.99	0.2507	0.6466

Iteration 4

Steps 2, 3, and 4:

ConfinedExpo	$P=0.2507(1-\exp(-0.6466u))$			
u	*Mu(u) = P*	*g = 0.2641*Mu-0.0035*	*g^2*	*w(u)=1/g^2)*
0	0	-0.0035	0.00001225	81632.65
1	0.11937739	0.02802757	0.000785545	1273.00
2	0.18191009	0.04454246	0.00198403	504.02
3	0.21466621	0.05319335	0.002829532	353.42

Step 5:

```
proc nlin data = growth method = marquardt;
 parms Pstar=0.1 a0=0.1;
 model growthdepth = Pstar*(1-exp(-a0*year));
       if year = 0 then _weight_ = 81632.65; if year = 1 then _weight_ = 1273.00;
       if year = 2 then _weight_ = 504.02; if year = 3 then _weight_ = 353.42; run;
```

Step 6:

Iteration	Weighted SS	*Pstar*	*a0*
0	.	0.2613	0.5866
1	19.07	0.2502	0.6496
2	19.01	0.2508	0.6465
3	18.99	0.2507	0.6466
4	18.99	0.2507	0.6466

STOP

Similarly, the weighted least squares procedures can be used to obtain weighted sum of squares and the optimal model parameters for the power law equation and the "new" model fitting to the corrosion growth data sets of Hickam AFB, Kadena AB, RAF Mildenhall, and Seymour Johnson AFB. Thus, the detailed procedures will not be presented for these models. However, the summary results will be provided.

2. The Power Law Equation Fitting to the Corrosion Growth Data Set of Hickam AFB

PowerLaw	P=0.1212*u^0.5304			
u	*Mu(u) = P*	*g = 0.2641*Mu-0.0035*	*g^2*	*w(u)=1/g^2)*
0	0	-0.0035	0.00001225	81632.65
1	0.1212	0.02850892	0.000812759	1230.38
2	0.17505275	0.04273143	0.001825975	547.65
3	0.21705397	0.05382395	0.002897018	345.18

Iteration	Weighted SS	*k*	*m*
0	-	0.1202	0.5407
1	19.09	0.1212	0.5302
2	19.12	0.1212	0.5304
3	19.11	0.1212	0.5304
4	19.11	0.1212	0.5304

3. The Model GLFitting to the Corrosion Growth Data Set of Hickam AFB

new1				
u	*Mu(u) = P*	*g = 0.2641*Mu-0.0035*	*g^2*	*w(u)=1/g^2)*
0	0	-0.0035	0.00001225	81632.65
1	0.1198441	0.02815083	0.000792469	1261.88
2	0.18078047	0.04424412	0.001957542	510.84
3	0.21496527	0.05327233	0.002837941	352.37

Iteration	Weighted SS	*Pstar*	*k*	*a0*
0	-	0.3713	0.2875	0.1616
1	19.02	0.3501	0.312	0.1709
2	19.00	0.351	0.311	0.1706
3	18.98	0.3509	0.311	0.1706
4	18.97	0.3509	0.311	0.1706

4. The Model GC Fitting to the Corrosion Growth Data Set of Hickam AFB

New2				
u	*Mu(u) = P*	*g = 0.2641*Mu-0.0035*	*g^2*	*w(u)=1/g^2)*
0	0.103336	0.023791	0.000566	1766.74
1	0.1720309	0.0419334	0.0017584	568.70
2	0.2107785	0.0521666	0.0027214	367.46
3	0.2342426	0.0583635	0.0034063	293.57

Replication	Weighted SS	*Pstar*	*k*	*a0*
0	-	0.3713	0.2875	0.4351
1	15.37	0.3568	0.3039	0.4704
2	14.15	0.3602	0.2999	0.4616
3	14.38	0.3595	0.3007	0.4634
4	14.39	0.3597	0.3005	0.463

5. The Model CL Fitting to the Corrosion Growth Data Set of Hickam AFB

new	*P=(0.3543*0.5147*u)/(1+0.5147*u)*			
u	*Mu(u) = P*	*g = 0.2641*Mu-0.0035*	*g^2*	*w(u)=1/g^2)*
0	0	-0.0035	0.00001225	81632.65
1	0.14256107	0.03415038	0.001166248	857.45
2	0.22777547	0.0566555	0.003209846	311.54
3	0.2787115	0.07010771	0.004915091	203.46

Iteratio	Weighted SS	*Pstar*	*a0*
0	-	0.3735	0.4602
1	11.49	0.3535	0.5172
2	11.35	0.3544	0.5146
3	11.31	0.3543	0.5147
4	11.31	0.3543	0.5147

Kadena

6. The Confined Exponential Model Fitting to the Corrosion Growth Data Set of Kadena AB

ConfinedExpo	$P=0.152(1-\exp(-1.0121u))$			
u	$Mu(u) = P$	$g = 0.1785*Mu+0.0011$	g^2	$w(u)=1/g^2)$
0	0	0.0011	0.00000121	826446.28
1	0.09674933	0.01836975	0.000337448	2963.42
2	0.13191686	0.02464716	0.000607482	1646.14
3	0.14469995	0.02692894	0.000725168	1378.99

Iteration	Weighted SS	Pstar	a0
0	-	0.1543	0.9610
1	17.90	0.1519	1.0146
2	17.78	0.152	1.012
3	17.78	0.152	1.0121
4	17.78	0.152	1.0121

7. The Power Law Equation Fitting to the Corrosion Growth Data Set of Kadena AB

PowerLaw	$P=0.0972*u^0.3703$			
u	$Mu(u) = P$	$g = 0.1785*Mu+0.0011$	g^2	$w(u)=1/g^2)$
0	0	0.0011	0.00000121	826446.28
1	0.0972	0.0184502	0.00034041	2937.64
2	0.12564282	0.02352724	0.000553531	1806.58
3	0.14599711	0.02716048	0.000737692	1355.58

Iteration	Weighted SS	k	m
0		0.0963	0.3840
1	16.75	0.0972	0.3698
2	16.72	0.0972	0.3703
3	16.71	0.0972	0.3703
4	16.71	0.0972	0.3703

8. The GL Model Fitting to the Corrosion Growth Data Set of Kadena AB

new1				
u	*Mu(u) = P*	*g = 0.1785*Mu+0.0011*	*g^2*	*w(u)=1/g^2)*
0	0	-0.0035	0.00001225	81632.65
1	0.09714911	0.02215708	0.000490936	2036.92
2	0.13035342	0.03092634	0.000956438	1045.55
3	0.14462902	0.03469652	0.001203849	830.67

Iteration	Weighted SS	*Pstar*	*k*	*a0*
0	-	0.1942	0.4479	0.1635
1	11.27	0.1882	0.4704	0.1713
2	11.19	0.1885	0.4693	0.171
3	11.19	0.1885	0.4693	0.171
4				

9. The GC Model Fitting to the Corrosion Growth Data Set of Kadena AB

New2				
u	*Mu(u) = P*	*g = 0.1785*Mu+0.0011*	*g^2*	*w(u)=1/g^2)*
0	0.1484396	0.0275965	0.0007616	1313.08
1	0.1549766	0.0287633	0.0008273	1208.71
2	0.1586431	0.0294178	0.0008654	1155.53
3	0.1607978	0.0298024	0.0008882	1125.89

Iteration	Weighted SS	*Pstar*	*k*	*a0*
0	-	0.1942	0.4479	0.8418
1	11.17	0.1936	0.4499	0.8479
2	11.15	0.1937	0.4499	0.8478
3	11.14	0.1937	0.4499	0.8478
4				

10. The CL Model Fitting to the Corrosion Growth Data Set of Kadena AB

new	P=(0.1929*0.1.0146*u)/(1+1.0146*u)			
u	Mu(u) = P	g = 0.1785*Mu+0.0011	g^2	w(u)=1/g^2)
0	0	0.0011	0.00000121	826446.28
1	0.12294362	0.02304544	0.000531092	1882.91
2	0.16752988	0.03100408	0.000961253	1040.31
3	0.18369937	0.03389034	0.001148555	870.66

Iteration	Weighted SS	Pstar	a0
0	-	0.1973	0.9472
1	10.90	0.1926	1.0193
2	10.86	0.1929	1.0143
3	10.86	0.1929	1.0146
4	10.86	0.1929	1.0146

RAF Mildenhall

11. The Confined Exponential Model Fitting to the Corrosion Growth Data Set of RAF Mildenhall

ConfinedExpo	$P=0.194(1-\exp(-0.9239u))$			
u	$Mu(u) = P$	$g = 0.4156*Mu-0.005$	g^2	$w(u)=1/g^2)$
0	0	-0.005	0.000025	40000.00
1	0.11698824	0.04362031	0.001902732	525.56
2	0.16342881	0.06292101	0.003959054	252.59
3	0.18186422	0.07058277	0.004981927	200.73

Iteration	Weighted SS	Pstar	$a0$
0	-	0.1940	0.9240
1	8.52	0.194	0.9239
2	8.5255	0.194	0.9239
3	8.5255	0.194	0.9239
4	8.5255	0.194	0.9239

12. The Power Law Equation Fitting to the Corrosion Growth Data Set of RAF Mildenhall

PowerLaw	$P=0.1174*u^0.4216$			
u	$Mu(u) = P$	$g = 0.4156*Mu-0.005$	g^2	$w(u)=1/g^2)$
0	0	-0.005	0.000025	40000.00
1	0.1174	0.04379144	0.00191769	521.46
2	0.15724699	0.06035185	0.003642346	274.55
3	0.18656167	0.07253503	0.005261331	190.07

Iteration	Weighted SS	k	m
0	-	0.1179	0.4121
1	8.67	0.1174	0.4214
2	8.67	0.1174	0.4216
3	8.66	0.1174	0.4216
4	8.66	0.1174	0.4216

13. The GL Model Fitting to the Corrosion Growth Data Set of RAF Mildenhall

new1				
u	*Mu(u) = P*	*g = 0.4156*Mu-0.005*	*g^2*	*w(u)=1/g^2)*
0	0	-0.005	0.000025	40000.00
1	0.11697752	0.04361586	0.001902343	525.67
2	0.16344663	0.06292842	0.003959986	252.53
3	0.18180143	0.07055668	0.004978244	200.87

Iteration	Weighted SS	*Pstar*	*k*	*a0*
0	-	0.1936	-0.0061	0.1789
1	8.53	0.1936	-0.00612	0.1789
2				
3				
4				

14. The GC Model Fitting to the Corrosion Growth Data Set of RAF Mildenhall

The results do not exist since the iterative estimation procedure failed to converge

15. The CL Model Fitting to the Corrosion Growth Data Set of RAF Mildenhall

new	*P=(0.2568*0.8388*u)/(1+0.8388*u)*			
u	*Mu(u) = P*	*g = 0.4156*Mu-0.005*	*g^2*	*w(u)=1/g^2)*
0	0	-0.005	0.000025	40000.00
1	0.14581472	0.0556006	0.003091426	323.48
2	0.20883375	0.08179131	0.006689818	149.48
3	0.23606968	0.09311056	0.008669576	115.35

Iteration	Weighted SS	*Pstar*	*a0*
0	-	0.2553	0.8506
1	5.10	0.2568	0.8389
2	5.10	0.2568	0.8388
3	5.10	0.2568	0.8388
4	5.10	0.2568	0.8388

Seymour Johnson AFB

16. The Confined Exponential Model Fitting to the Corrosion Growth Data Set of Seymour Johnson AFB

ConfinedExpo	$P=0.1258(1-\exp(-0.1.3203u))$			
u	*Mu(u) = P*	*g = 0.2281*Mu-0.002*	*g^2*	*w(u)=1/g^2)*
0	0	-0.002	0.000004	250000.00
1	0.09220446	0.01903184	0.000362211	2760.82
2	0.11682814	0.0246485	0.000607548	1645.96
3	0.12340402	0.02614846	0.000683742	1462.54

Iteration	Weighted SS	*Pstar*	*a0*
0	-	0.1260	1.3108
1	7.15	0.1258	1.3204
2	7.1464	0.1258	1.3203
3	7.1464	0.1258	1.3203
4	7.1464	0.1258	1.3203

17. The Power Law Equation Fitting to the Corrosion Growth Data Set of Seymour Johnson AFB

PowerLaw	$P=0.0926*u\text{^}02832$			
u	*Mu(u) = P*	*g = 0.2281*Mu-0.002*	*g^2*	*w(u)=1/g^2)*
0	0	-0.002	0.000004	250000.00
1	0.0926	0.01912206	0.000365653	2734.83
2	0.11268411	0.02370325	0.000561844	1779.85
3	0.12639553	0.02683082	0.000719893	1389.10

Iteration	Weighted SS	*k*	*m*
0	-	0.0927	0.2803
1	7.25	0.0926	0.2832
2	7.23	0.0926	0.2832
3	7.23	0.0926	0.2832
4	7.23	0.0926	0.2832

18. The GL Model Fitting to the Corrosion Growth Data Set of Seymour Johnson AFB

new1				
u	*Mu(u) = P*	*g = 0.2281*Mu-0.002*	*g^2*	*w(u)=1/g^2)*
0	0	-0.002	0.000004	250000.00
1	0.09229339	0.01905212	0.000362983	2754.95
2	0.11571081	0.02439364	0.00059505	1680.53
3	0.12418448	0.02632648	0.000693084	1442.83

Iteration	Weighted SS	*Pstar*	*k*	*a0*
0	-	0.1480	0.5775	0.1900
1	7.15	0.1477	0.5794	0.1908
2	7.14	0.1477	0.5794	0.1908
3				
4				

19. The GC Model Fitting to the Corrosion Growth Data Set of Seymour Johnson AFB

New2				
u	*Mu(u) = P*	*g = 0.2281*Mu-0.002*	*g^2*	*w(u)=1/g^2)*
0	0.1273077	0.0270389	0.0007311	1367.80
1	0.129508	0.0275408	0.0007585	1318.40
2	0.1306388	0.0277987	0.0007728	1294.05
3	0.1312428	0.0279365	0.0007804	1281.32

Iteration	Weighted SS	*Pstar*	*k*	*a0*
0	-	0.1480	0.5774	0.2840
1	46.41	0.1477	0.5792	1.2908
2	5.25	0.148	0.5776	1.2844
3	5.24	0.148	0.5776	1.2844
4				

20. The CL Model Fitting to the Corrosion Growth Data Set of Seymour Johnson AFB

new	P=(0.1516*0.1.5603*u)/(1+1.5603*u)			
u	Mu(u) = P	g = 0.2281*Mu-0.002	g^2	w(u)=1/g^2)
0	0	-0.002	0.000004	250000.00
1	0.11975293	0.02531564	0.000640882	1560.35
2	0.14490979	0.03105392	0.000964346	1036.97
3	0.15019457	0.03225938	0.001040668	960.92

Iteration	Weighted SS	Pstar	a0
0	-	0.1516	1.5585
1	4.42	0.1516	1.5603
2	4.42	0.1516	1.5603
3	4.42	0.1516	1.5603
4	4.42	0.1516	1.5603