Image Fusion and An Outlier Detection Framework for Hierarchical Modeling with Application to Corrosion Prediction

A Dissertation

Presented to

the faculty of the School of Engineering and Applied Science

University of Virginia

in partial fulfillment

of the requirements for the degree

Doctor of Philosophy

by

Lei Chen

May

2012

APPROVAL SHEET

The dissertation

is submitted in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

AUTHOR

The dissertation has been read and approved by the examining committee:

Advisor

Accepted for the School of Engineering and Applied Science:

James H. Ayl

Dean, School of Engineering and Applied Science

May

2012

ABSTRACT

Semi-continuous data appear frequently in many scientific fields which is a special type of data which include a number of continuous data with reoccurrences of some discrete numbers. For example, in epidemiology studies, data often include both zeros for those areas without certain disease, and positive values indicating the severity degree of the diagnosed epidemic cases in other places. Such data often include outliers and errors from the experiment and the measurement, which are not preventable. Modeling the semi-continuous data in the presence of noise is challenging because of the appearance of outliers and the skewness from the normal distribution within the data. Both of them may mislead the modeling result. It is imperative to model problems with such data, but available techniques are very limited.

This dissertation aims to develop a formal methodology using supervised learning when: (1) relevant information is stored in a series of images; (2) images are usually noisy and distorted from each other; (3) the data set for modeling is a data set with a semi-continuous response variable; and (4) the modeling goal is to understand the causal mechanism between variables and to predict future events accurately. The developed methodology includes three models for this objective: an image fusion algorithm, an outlier detection framework and a two-part generalized hierarchical model for semi-continuous data. They have been applied to a real corrosion problem and modeling results showed that this methodology solved this problem effectively. Corrosion data were efficiently extracted from corrosion images, outliers within the extracted data were detected and treated properly and most importantly, the underlying causal mechanisms between material microstructures and corrosion evolution were revealed by the generalized hierarchical model.

Four major contributions have been made: a supervised learning methodology is constructed for problems with information stored in both semi-continuous data and a series of noisy images; an outlier detection framework for supervised learning is constructed to enhance prediction accuracy; an image fusion algorithm is designed to extract and combine information from multiple noisy images, and the estimation of the generalized hierarchical model helps material scientists to reveal the causal mechanisms between grain boundary characteristics and the intergranular corrosion, as well as to predict future corrosion occurrences. Future works of this dissertation are discussed at last.

ACKNOWLEDGMENTS

My PhD life at University of Virginia is one of the most wonderful experience in my life. I enjoyed every bitter and sweet moment during this journey. All the precious memories will become the real treasure for the rest of my life. I would never have been able to finish this dissertation without the guidance of my committee members, help from my friends, and support from my family, especially my husband.

First, I'd like to express my deepest gratitude to my two advisors, Prof. Donald Brown and Prof. Robert Kelly. I would like to thank Prof. Brown for patiently guiding on my study, for correcting on my writing, and for caring for our family. Thank you for providing me with such an excellent research opportunity. I would like to thank Prof. Kelly for introducing me to the world of corrosion. You were always there whenever I had questions or difficulties with my project. Thank you for your understanding and support during my difficult times. It was my pleasure to work with you and your group. I benefited a great deal from your guidance, and I have learned a lot from you, not only for being a good researcher and experimentalist, but also for being a good person. I would also like to thank my other committee members, Prof. Learmonth, Prof. Patek, Prof. Lambert, for your guidance and insights through my PhD study. Meanwhile, I am truly thankful for talking with Prof. Sean Agnew. Thank you for patiently explaining to me many details of materials microstructures.

I owe sincere and earnest thankfulness to Richard White, who successfully taught me to conduct difficult experiments, such as EBSD, polishing and etching. You are so kind and considerate that I really enjoyed working with you. My special thank goes to Mary Lyn Lim. Thank you for helping me with the complicated experiment

when I had no clue about it, and for cheering me up when I was down. I am also obliged to my dear colleagues Courtney Crane, Joelle Burzinski, and Elissa Bumiller in Dr. Kelly's group who supported me and had valuable discussions with me. Also, financial support for this study from the Office of Naval Research (Dr. Airan Perez), Grant N000140810315, is sincerely appreciated. I would like to thank my friends at UVa who have made my PhD life enjoyable: Xiaohuan, Tiantian, Yonghang, Ruwei, Zhenyu, Hui Hua, Haiyan, Zhang Nan, Yiyi, Jian Kang, Dandan, Mingyi, Wang Lu, Yijing, among others.

Last but foremost, I am sincerely grateful to my family. I would like to thank my parents, for giving me thoughtful care and love that makes my life cheerful, for supporting me to realize my dream, and for taking care of my baby so that I was able to make this dissertation possible. I would also like to thank my parents-in-law for consistently understanding and supporting us, and for helping us to take care of the family when we were too busy to make it. Most importantly, I am truly indebted and thankful to my husband Xiaofeng for the love and support that he has provided me throughout our years together. You have always been there to pick me up, to encourage me and to care for me when my confidence was shaken. I can never thank you enough for all you do for our family. I love you. Finally, thank you for your arrival, Aaron. You are my motivation to finish this and your smile reminds me of the truly meaning of life. Thank you for being the most important part of my life!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

This chapter introduces the background information, motivation, objectives and contributions of this dissertation study.

1.1 Background and Motivation

Semi-continuous data appear frequently in many scientific fields such as economy, ecology, physics, medicine and so on. It is a special type of data which include a number of continuous data with a reoccurrence of some discrete numbers. For example, the medical research about the drinking outcomes among alcohol-dependent individuals in Liu et al. (2008) used individual alcohol consumption data, which include positive and continuous values indicating the percentage of daily alcohol consumption and a large amount of zeros indicating no alcohol consumption. In epidemiology studies, data often include both zeros for those areas without certain disease, and positive values indicating the degree of severity of diagnosed epidemic cases in other places. Such data often include outliers and errors from the experiment and the measurement, which are not preventable. Modeling the semi-continuous data in the presence of noise is challenging because of the appearance of outliers and the skewness from the normal distribution within the data. Both of them may mislead the modeling result.

Semi-continuous data in the presence of noise are also common in material science studies. For instance, the variable *Percent β Coverage* records the percentage of the amount of corrosion occurred along each grain boundary from several aluminum alloy samples. These observations are collected from experimental images. This variable takes values from 0 to 1, with a large proportion of reoccurrences of 0's and 1's, corresponding to the situations of no corrosion occurred and totally corroded. This special distribution is due to the grain boundary characteristics of different grain boundaries and this variable is considered as a semi-continuous variable. Modeling the noisy semi-continuous corrosion data is a challenging and imperative task, because we are eager to understand the underlying causal mechanisms behind the data. However, the appearance of noise makes the statistical models inaccurate and unstable. We need a robust model which can resist to those noises. Although this is a similar problem to robust regression, classic robust regression models were not designed for semi-continuous data. This dissertation aims to construct a formal methodology to satisfy the need for the robust modeling of semi-continuous data in the presence of noise. The methodology is applied to an important example, which is the prediction of intergranular corrosion in AA5XXX-series alloys.

According to a national study of the corrosion cost in the United States funded by the Federal Highway Administration (Koch et al., 2002), the annual estimated direct cost of corrosion in 2001 was \$276 billion, equal to 3.1% of the national Gross Domestic Product (GDP). This amounted to \$981 per U.S. resident per year, based on the population of 2001. Indirect annual cost of corrosion was up to \$552 billion, which doubled the direct cost. These facts show the imperative need of intensive corrosion control studies. Koch et al. (2002) also pointed out in the report that one of the effective strategies for corrosion control was corrosion performance assessments and corrosion prediction, by improving the determination of severity of corrosion damage, rates of corrosion and methods of evaluating corrosion growth.

Among many metallic and nonmetallic materials, aluminum alloys are most widely used in marine, aerospace, automobile and manufacture industries because they have many advantages such as low density, high strength-to-weight ratio and tunable strength (Mondolfo, 1976). However, corrosion of aluminum alloys also limits their usage in these areas, especially for marine structures due to their corrosive seawater environment. Increasing the corrosion resistance of such alloys can be a challenging yet important problem and numerous studies and strategies have been carried out to improve the corrosion resistance of such alloys (Kim et al., 2001; Lo et al., 2009; Unwin et al., 1969; Yuan, 2006) since their usage became popular. According to those previous studies, there are many dominant factors that keep this problem so challenging. This dissertation deals with some of these significant difficulties which were either not paid enough attention to or were not solved efficiently by previous researchers in the material science and other similar engineering areas. These difficulties are discussed as follow.

First, collecting data for modeling can be difficult and time-consuming when data are stored in different forms of images. For instance, corrosion occurrences on alloy materials can be captured by optical microscopes in the form of gray scale images. Usually these images only have a limited number of features such as lines and dots. They are different from regular pictures because of the lack of meaningful complex objects like faces, trees or buildings. Figure 1.1 shows an example of such image. To characterize features of materials, another type of image, the electron backscatter diffraction (EBSD) image, is obtained by the HKL Channel EBSD acquisition system (Day and Trimby, 2004), as shown in Figure 1.2. Both Figure 1.1 and Figure 1.2 are containing different information about the corrosion degrees and physical properties of the same aluminum alloy. In order to build prediction models for such alloy, relevant corrosion data such as the degree of corrosion, the length of each grain boundary (black lines in Figure 1.1), need to be extracted from these two types of images, which are the only sources of such data. This essential task is called image fusion, which means combining information from different types of images to construct a complete data set for modeling use. The quality of the extracted data directly affects the quality of the prediction models to be built, so it is significantly important to have images fused as accurately as possible. Due to its broad application areas from medical science to material science, image fusion is one of the difficulties that is worth being analyzed and resolved.

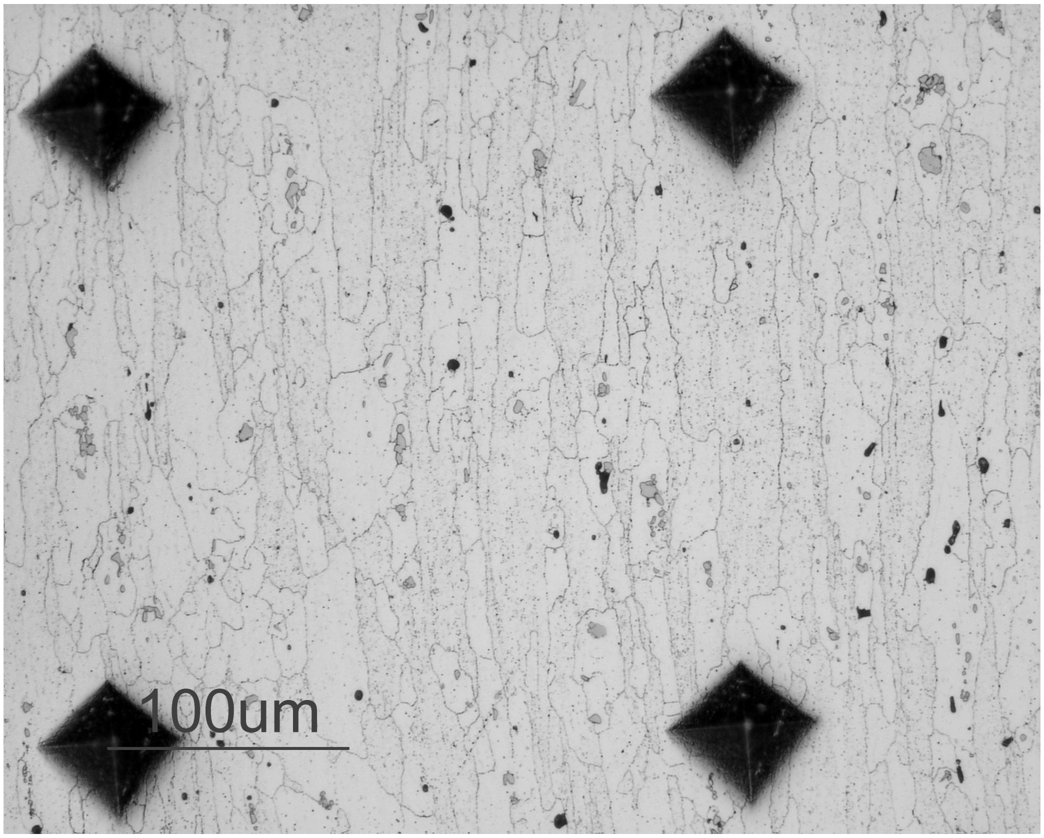


Figure 1.1: A corrosion image of Alloy AA5083-H131 from the optical microscope at magnification=200X. Degree of sensitization is 57 mg/cm^2. Sample was sensitized at 100°C for 45 days. Sample was etched in the solution of 20g ammonium persulfate and 100ml water at room temperature for 1 hour.

Second, as the major resource of data, images from experiments or scientific measurements are usually noisy, sometimes even distorted, due to the limitation of equipments being used or the properties of objects being recorded. Such noisiness can be observed in Figure 1.1, for example. Black dots within grains are all unwanted noises from the corrosive environment, and they are irrelevant to the desired data from this

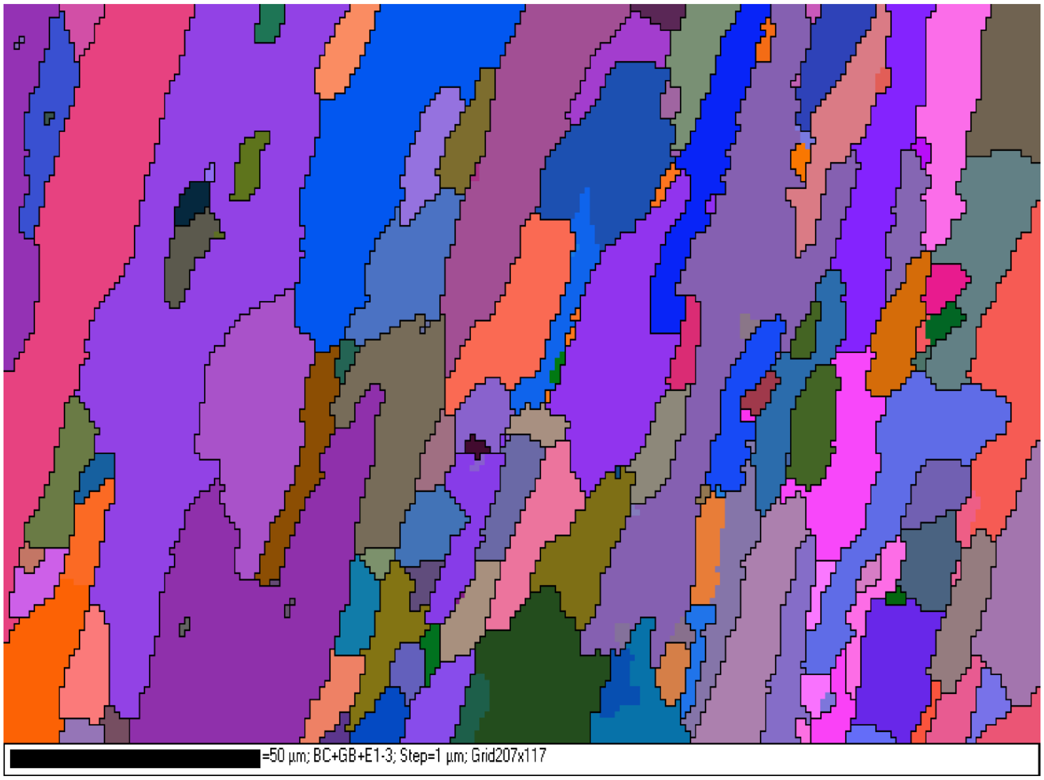


Figure 1.2: A microstructure image from Electron Backscatter Diffraction system at magnification=200X. Degree of sensitization is 24 mg/cm^2. Sample was sensitized at 100°C for 7 days.

image. Unfortunately, data extracted from such images via image fusion often include unwanted noisiness unavoidably. The noisiness shows up in the collected data set as outliers, which would indirectly affect the performance of the prediction models constructed based on them. Detecting and removing these outliers is a practical method to enhance the prediction accuracy of modeling, as well as to ensure the real causality mechanisms are explained. This task is as important as the image fusion mentioned above. It is worth extra deep analysis to make data clean for modeling.

Third, the special types of data extracted from the images often limits the usage of classic efficient and popular statistical learning methods. For instance, the data extracted from the above two images are often semi-continuous data, which are different from the assumptions of traditional statistical models. The data can be characterized as a mixture of non-zero continuously distributed values and a certain proportion of repeated single values, such as 0's (Olsen and Schafer, 2001). On one hand, such data do not have a normal distribution, so typical regression models would not be suitable for them; on the other hand, semi-continuous data usually come from surveys or experiments, so those single values such as zeros have important real meanings, and researchers are especially interested in those values. For example, zeros mean no corrosion occurred in corrosion science, and researchers are interested in why those areas are free of corrosion. Because those data are from the scientific experiment, statistical learning models constructed for such data are desired to provide causality mechanism explanations. However, in practice, only a limited number of such models are available there for scientists and engineers who are dealing with this special type. Therefore, it is imperative to explore more deeply in this area. It will be beneficial to design models especially for this specific type of data.

Not limited to the corrosion problems, there are lots of similar phenomena, in which valuable information is usually in the form of both special experimental data and a series of noisy images with few features. For example, biological scientists investigate tumor growth using test results and multiple MRI images. Meteorologists forecast future weather through local meteorological data and several satellite cloud pictures. The addressed difficulties with their frequent appearances in many scientific areas give the motivation of this dissertation. A method for robust modeling of semi-continuous data is described and analyzed in this dissertation.

1.2 Objective and Contributions

The objective of this dissertation is to develop a formal methodology using supervised learning for problems with the following characteristics:

(1) Relevant information is stored in a series of images;

(2) Images are noisy and distorted;

(3) Data for modeling include semi-continuous variables; and

(4) The modeling goal is to understand the causal mechanism and to predict future events accurately.

The developed formal methodology is applied to an intergranular corrosion problem in this dissertation, but it is also applicable to similar phenomena, such as weather forecasting of different areas based on satellite cloud pictures, tumor growth prediction based on magnetic resonance images (MRI) and epidemic disease spreading modeling. This dissertation provides four major contributions:

(1) A supervised learning methodology is constructed to model processes with information stored in both semi-continuous data and a series of noisy images. Corrosion prediction problems can be solved by this methodology and we provide a case study.

(2) An outlier detection framework for supervised learning is constructed to enhance prediction accuracy. It is applied to a generalized hierarchical model for the corrosion prediction problem. It is applicable to other supervised learning models, such as linear regression, random forest and generalized additive model.

(3) An image fusion algorithm is presented that can extract and combine information from multiple noisy images. Corrosion images taken by different equipment are fused by this algorithm to collect data for modeling. It might also be applicable to a wider range of areas such as magnetic resonance imaging (MRI) for tumor growth monitoring and geographic information integrating.

(4) The estimation of the generalized hierarchical model can help material scientists to reveal the causal mechanisms between grain boundary characteristics and the intergranular corrosion, as well as to predict future corrosion occurrences.

1.3 Scope of the Rest of Dissertation

The remainder of this dissertation addresses the developed formal methodology for problems with specific characteristics described above. This formal methodology includes various models and algorithms for data collecting, data cleaning and prediction model construction, which are introduced and analyzed in the following chapters respectively. Also, modeling results and conclusions from the application to an intergranular corrosion prediction problem are presented as the model evaluation.

Chapter 2 introduces related research on supervised learning, outlier detection, image registration and a specific topic about intergranular corrosion; chapter 3 focuses on the framework for the methodology; Chapter 4, 5 and 6 describes each component of the methodology development in detail; Chapter 7 gives an application example of the methodology, which is to predict intergranular corrosion given corrosion image data. It also shows the modeling results of the applied methodology; Chapter 8 addresses dissertation conclusions and recommendations for future work.

CHAPTER 2

LITERATURE REVIEW

In this chapter, five research areas relevant to this dissertation are reviewed: supervised learning, semi-continous data, outlier detection, image registration and background information about corrosion. Supervised learning is the main topic of this dissertation and several classical regression and classification models are introduced under this section. The relevant information about intergranular corrosion and grain boundary characteristics is also presented, which motivates this research and will be used as an application of the developed methodology in Chapter 7.

2.1 Supervised Learning

With the development of tools for data collecting, vast amount of data are generated from many fields. Thus, it is important to learn from these data and explain patterns as well as trends behind the data. According to Friedman et al. (2001), if we predict the values of outputs based on a number of input observations by some learning algorithms, then this learning process is called supervised learning. Suppose there is an observation set $T = \{(x_i, y_i)|i = 1, 2, \ldots, N\}$, and it is assumed that the model $Y = f(x) + \epsilon$ represents the relationship between inputs and outputs. Supervised learning aims to learn the function f through some learning algorithms, and produces outputs $\hat{f}(x_i)$ corresponding to inputs x_i. The learning algorithm aims to minimize the difference between the original and the generated outputs $y_i - \hat{f}(x_i)$, which is known as *learning by example.*

Supervised learning is a broad topic in the statistical learning field. The two major components of supervised learning are linear models and nonlinear models for both regression and classification problems. Linear regression models assume that the relationship between inputs and outputs is linear, or similarly the function f is assumed to be linear. In many cases, linear regression models are simple and effective to explain how inputs affect outputs. They can also be generalized into many complex models. If the predictor takes values from a discrete set, then it is possible to partition the input space into a collection of labeled regions. Linear models for classification aim to find linear boundaries to separate these regions. There are various methods to find linear decision boundaries, such as the indicator matrix method, discriminant analysis, logistic regression and separating hyperplanes. More detailed reviews about linear regression and classification models are well presented in Seber (2004), Weisberg (2005), Duda et al. (2001) and Friedman (1994). For most of the real data sets, it is not always the truth that $f(X) = E(Y|X)$ is linear in response to X. Thus, we transform the input vector X in a nonlinear way, and utilize linear models with this newly derived input space. This nonlinear method is considered as an extension of the linear methods, and is known as the basis-function method. Important basis-function based methods for nonlinear regression and classification problems include the piecewise polynomials, smoothing splines (de Boor, 1978; Green and Silverman, 1994), nonparametric logistic regression and wavelet smoothing (Daubechies, 1992). The kernel-based technique is another class of regression techniques. It estimates the regression function $f(X)$ by fitting a simple but different model separately at each query point x_0 and only uses observations close to that point (Loader, 1999).

There exist many other important techniques in the supervised learning area. The generalized additive model is an effective and flexible statistical technique to characterize the nonlinear relationship between inputs and outputs. It can be used for both regression and classification problems. They are fitted by a scatterplot smoother and all p functions can be estimated simultaneously, as shown in Hastie and Tibshirani (1990). Tree-based methods play an important role in supervised learning area. They partition the input space into a set of rectangles and fit separate simple model for each one. Other learning techniques include the support vector machine (SVM), neural network and random forest. Boosting is one of the most powerful learning method developed in the last decade. It is a procedure of combining the results of many weak learners and reweighting their importances by majority vote (Schapire and Freund, 1999). More explanations about boosting methods can be found in Schapire and Singer (1999) and Friedman (2002). Details of some widely used supervised learning methods are discussed in the following sections.

2.1.1 Linear Model

Linear models for regression are the most popular and well developed supervised learning methods in statistical learning area. Suppose the input vector is $X = \{X_1, X_2, \ldots, X_N\}$, and the real-value output is Y. The linear regression model which represents the relationship between the input and output is in the form of

$$f(X) = \beta_0 + \sum_{i=1}^{N} X_i \beta_i \tag{2.1}$$

where $\beta_i's$ are unknown parameters or coefficients that need to be estimated, and $X_i's$ are quantitative values. This model is linear in the parameters, so it is called a linear regression model. To estimate unknown parameters, the most popular method is least squares estimation (Sorenson, 1970), which is to minimize the residual sum of squares (RSS)

$$\min\ RSS(\beta) = \sum_{i=1}^{N} (y_i - f(x_i))^2 \tag{2.2}$$

where $(x_i, y_i), i = 1, \ldots, N$ are from a set of training data (Friedman et al., 2001).

Linear models for classification basically have the same modeling strategy as linear models for regression. The difference between them is that classification problems have numeric levels of qualitative inputs. For instance, if a variable T is a three-level factor input, then we can define $X_j, j = 1, 2, 3$, such that $X_j = I(T = j)$. Therefore, the effect of T is represented by a group of X_j along with a set of level-dependent constants (Friedman et al., 2001). Popular linear models for classification include logistic regression, linear discriminant analysis (Mardia et al., 1980), and linear separating hyperplanes addressed in Vapnik (2000).

As a flexible generalization of linear models, generalized linear models focus on the problems where the response variable is related to the predictors through a link function (Nelder and Wedderburn, 1972). The link function can take various forms such as identity, inverse, log and logit. Nelder and Wedderburn (1972) developed an iteratively re-weighted least squares method for maximum likelihood estimation, in order to estimate the parameters of generalized linear models.

2.1.2 Generalized Additive Model

The generalized additive model is a supervised learning model developed by Hastie and Tibshirani (1990). It is a flexible combination of the properties of generalized linear models and additive models. It is able to capture and characterize the nonlinear regression effects between the response variable and predictors, which is valuable and essential when linear models fail for some problems. A generalized additive model has a form of

$$Y = \alpha + \sum_{i=1}^{N} f_i(X_i) + \epsilon \tag{2.3}$$

where f_j's are nonparametric smooth functions, and ϵ is the error term with mean zero. These smooth functions can be estimated by many different algorithms, such as cubic smoothing spline (Green and Silverman, 1994) and kernel smoother (Wand and Jones, 1995). Hastie and Tibshirani (1987) developed a local scoring algorithm to estimate the smooth functions f_j nonparametrically, with a scatterplot smoother as a building block. Wood (2008) discussed further about estimating the smooth functions of GAMs within a penalized likelihood framework. Wood (2008) pointed out that existing methods for penalized likelihood fitting may not converge under some circumstances, and summarized the three basic approaches for smoothness estimation: generalized cross-validation (GCV) (Golub et al., 1979), restricted maximum likelihood or penalized quasi-likelihood (Breslow and Clayton, 1993), and direct smoothness estimation based on Akaike information criterion (Akaike, 1973) or GCV. A more stable and direct estimation method was developed in this work to avoid drawbacks of the pre-existing methods.

2.1.3 Hierarchical Model

As an extension of generalized linear regression models, hierarchical models (also known as multilevel models) are widely used to deal with data having hierarchical structures in biology, pharmacology, psychology, education and so on. Hierarchical models enable us to explore the variation at different levels within the hierarchy, which can bring good interpretability of variances from different variables. Hierarchical models were systematically studied by Goldstein (1995) and they can be classified into linear hierarchical models and nonlinear hierarchical models (Pinheiro, 1994).

Linear hierarchical models were developed as variance components models in the first place. Henderson (1953) introduced the ANOVA method to estimate variance components for unbalanced data and it became the standard estimation method for linear hierarchical models before fast computation tools were invented. Crump (1946) developed a maximum likelihood estimation (MLE) method and it was extended by Hartley and Rao (1967) for a large class of variance components models. After MLE, restricted maximum likelihood estimation (RMLE) was introduced and developed by Thompson Jr (1962) and Patterson and Thompson (1971). Then, generalized linear hierarchical models were able to be estimated using EM algorithms to obtain (R)MLEs of the variance components. More general estimation approaches were proposed after MLE and RMLE, such as Bayesian inference (Harville, 1974), a combination of empirical Bayes and MLE (Laird and Ware, 1982), and MLE with a scoring method (Chi and Reinsel, 1989). Bayesian analysis using the Gibbs sampler, introduced by Geman et al. (1984), is another important topic in this area (Gelfand et al., 1990; Wakefield et al., 1994). This method can relax the restricted assumption of Gaussian distributions for the random effects and error terms. However, it still has drawbacks including intensive computation efforts and the need of prior distributions of all individual level parameters. Liang and Zeger (1986) and Zeger et al. (1988) discussed the generalized linear hierarchical models with a link function, which can also be considered as a type of nonlinear hierarchical models. Detailed and comprehensive reviews of linear hierarchical models and estimation methods are shown in books of Searle et al. (1992), Lindsey and Aickin (1994), Longford (1993) and Gelman et al. (2007).

Nonlinear hierarchical models were firstly introduced in the pharmacokinetics area by Beal and Sheiner (1980). Many nonlinear hierarchical models were discussed by Gallant and Corporation (1987), Pinheiro and Bates (2009), Grossman and Koops (1988). MLE methods have been developed based on Taylor expansions to estimate this type of model. Many efforts have been made to estimate nonlinear hierarchical models. Mallet et al. (1988) created a nonparametric MLE method for nonlinear hierarchical models without assumptions of the distribution of random effects. Davidian and Gallant (1992) proposed a smooth nonparametric MLE method for nonlinear hierarchical models. Bennett et al. (1996) applied a Bayesian method to estimate the nonlinear hierarchical models. The assumption was that the distributions of random effects and error terms were known and prior distributions of individual level parameters were also known.

Nonlinear hierarchical models include those models in which some levels are linear and other levels are nonlinear as well. This type of model was introduced by Vonesh and Carter (1992). They are also called mixed effects models. Model parameters can be estimated by a generalized least squares procedure. As addressed by Pinheiro and Bates (2009), nonlinear hierarchical models are useful when we need simplicity, good interpretability and more importantly, validity from the modeling beyond the observed range of data.

2.1.4 Other Supervised Learning Methods

Many supervised learning algorithms, other than classical ones described above have been recently developed since the using of high efficiency computation tools. Support Vector Machine (SVM) is one of the most successful and effective methods, which was originally introduced by Vapnik in 1963 (Vapnik, 2000). It can be used for both classification and regression problems. The basic idea of SVM is that it works as a classifier, and generates a hyperplane or a set of hyperplanes in a high dimensional space, in order to separate the inputs into different classes. The optimal hyperplane has the greatest distance to the nearest data point of any class. SVM was firstly developed as a linear classifier, and later nonlinear classifier algorithms were introduced in Boser et al. (1992) by using kernel functions. Frequently used kernel functions include polynomial functions and the Gaussian radial-basis function.

Random Forest (Breiman, 2001) is another popular supervised learning algorithm developed recently, and is also one of the most accurate learning methods. Performing as a classifier, Random Forest combines many decision trees together, and gives a class as an output which is the mode of the classes by those trees. It is characterized by the “bagging” algorithm (Breiman, 1996) with the random selection of features (Amit and Geman, 1997; Ho, 1998). Random Forest is able to handle large size data sets with many input variables efficiently, and generate unbiased estimations of the generalization error. However, it also has some disadvantages as other methods have, such as overfitting some data sets with noisy data points. The choice of the number of decision trees to be used in the model also affects its performance.

Multivariate Adaptive Regression Splines (MARS) is a nonparametric adaptive regression procedure introduced by Friedman (1991), and it works well for high dimensional data. It can be considered as a generalization of the stepwise linear regression which automatically models the nonlinearity. MARS takes the form

$$f(X) = \beta_0 + \sum_{j=1}^{N} \beta_j B_j(X) \tag{2.4}$$

where $B_j(X)$ is a basis function or a linear spline, c_j is a coefficient. MARS can also be applied to classification problems, given special developed modeling strategies such as PolyMARS (Kooperberg et al., 1997). MARS can deal with both numeric and categorical data, and is well suited for large data sets. MARS is a nonparametric regression, so model validation can only be done indirectly with techniques such as cross-validation.

Boosting is known as one of the most important and powerful statistical learning algorithms developed in the last decade. As the most popular boosting method, "Adaboost" was introduced by Schapire and Freund (1999). The idea of boosting is to combine the outputs of a group of "weak" learners or classifiers, in order to generate a strong learner. A learner is considered as a "weak" learner if its error rate is only slightly better than random guessing. The final prediction is made based on a reweighted majority vote of many "weak" learners. After repeated modifications, such a voting procedure is able to give a powerful prediction:

$$C(x) = sign(\sum_{i=1}^{N} \alpha_i C_i(x)) \tag{2.5}$$

where $C(x)$ is the updated classifier, C_i are "weak" learners, $\alpha_1, \alpha_2, \ldots, \alpha_N$ are computed by the boosting algorithm. Different boosting algorithms may give different weighting strategies. Suppose there are M observations. "Adaboost.M.1" updates α_i as

$$\alpha_i = log(\frac{1 - err_i}{err_i}) \tag{2.6}$$

where $err_i = \frac{\sum_{j=1}^{M} w_j I(y_j \neq C_i(x_j))}{\sum_{j=1}^{M} w_j}$, and w_j is the weight of each observation, with an initial value of $1/M$.

2.2 Semi-continuous Data

Semi-continuous data are often characterized as a mixture of non-zero continuously distributed values and a certain proportion of repeated single values, such as 0's (Olsen and Schafer, 2001). Such data appear frequently in economics, ecology, physics and medicine. For example, the research about the drinking outcomes among alcohol-dependent individuals in Liu et al. (2008) has individual alcohol consumption data, which include positive and continuous values indicating the percentage of daily alcohol consumption and a large amount of zeros indicating no alcohol consumption. In epidemiology studies, data often include both zeros for those areas without certain disease, and positive values indicating the severity degree of diagnosed epidemic cases at other places. Semi-continuous data are also common in material science studies. Here in the intergranular corrosion problem we are solving, the response variable *Percent β Coverage* records the percentage of the amount of corrosion occurred along each grain boundary from several aluminum alloy samples. This variable takes values from 0 to 1, with a large proportion of reoccurrences of 0's and 1's, corresponding to situations of no corrosion occurred and totally corroded. This special distribution is due to the grain boundary characteristic differences among grain boundaries and we consider such a variable as a semi-continuous variable. Analysis of the semi-continuous data is challenging because of the presence of skewness from the normal distribution within the data.

2.3 Outlier Detection

With the development of information technology, more and more data have been collected and analyzed for the purpose of knowledge discovery. Sometimes, collected data include human or machine errors. These errors are considered as outliers for the modeling purpose. For example, a measurement is wrongly recorded due to limitations of designed algorithms in a science experiment. Another type of outliers is the unusual incident such as credit card frauds Kou et al. (2005) and computer virus attacks Ertoz et al. (2003). It is desirable to remove or correct errors from the dataset for a better understanding of patterns underneath, and it is also important to report unusual incidents for safety consideration. Both of these needs are called outlier detection.

Outlier detection has been one of the most important topics in statistics since the 19^{th} century. With the rapid growth of data dimensions and types of data sources, it is worth making more effort to develop outlier detection methods for these data. In practice, various outlier detection techniques have been developed. Chandola et al. (2009) provided a comprehensive survey of outlier detection techniques. These techniques are based on different statistical and data mining methods such as classification models, neural network models, Bayesian networks, SVM, rule-based models, the nearest-neighbor models (using distance to the k^{th} nearest neighbor and using relative density), clustering algorithms, and other information theoretic models.

Besides the above techniques, there are many new approaches developed recently for outlier detection. Abe et al. (2006) presented an approach to reduce outlier detection problem to a classification problem, and then applied a selective sampling mechanism based on active learning to the reduced classification problem. Aggarwal and Yu (2008) examined and applied a density based approach to the problem of how to remove the uncertainty from data. Latecki et al. (2007) developed an unsupervised algorithm for outlier detection. A nonparametric density estimation was modified to obtain a robust local density estimation and then outliers were detected by comparing the local density of each point to the local density of its neighbors . Kriegel et al. (2008) discussed outlier detection techniques for high-dimensional data. Instead of using distance-based approaches to high-dimensional data, the authors proposed a novel approach named angle-based outlier detection and also compared it to other distance-based methods. Filzmoser et al. (2008) also discussed outlier detection in the high dimension. The algorithm adopted properties of principal components to identify outliers in the transformed space, so that the computational time was less than other existing methods. Another outlier detection schema for high dimensional data was proposed by Kriegel et al. (2009), utilizing axis-parallel subspaces. Their model determined how much the object deviates from the neighbors in this subspace. She and Owen (2010) developed an outlier detection method based on penalized regression, utilizing a thresholding based iterative procedure to detect outliers. Cerioli (2010) developed multivariate outlier tests based on the high-breakdown Minimum Covariance Determinant estimator. The rules have a good performance under the null hypothesis of no outliers in the data. Riani et al. (2009) applied the forward search method to obtain robust Mahalanobis distances for outlier detection in a multivariate normal data set. Several new robust distances were also introduced, and comparison results showed the power of this approach. Nguyen and Welsch (2009) presented a robust linear regression approach utilizing the “maximum trimmed squares” (MTS), which maximized the sum of the q smallest squared residuals, instead of the least trimmed squares (LTS), to capture the set of outliers. Plus, MTS can be solved efficiently in polynomial time.

Generally, there are two types outlier detection models, based on different objectives of outlier detection problems. For problems aiming to just find out unusual incidents or outliers, unsupervised learning methods are typically utilized, such as distance-based models in Knorr and Ng (1998). Observation points far from the majority points are considered as outliers. These methods are applicable to problems such as credit card fraud detection and computer virus attack detection. To evaluate this type of methods, a predefined data set including known outliers is used as a test set (Aggarwal and Yu, 2001). The more accurately known outliers are filtered out, the better a method performs. The other type of problems belongs to the realm of supervised learning. Outliers in such problems usually mislead the result of regression or classification, which leads to bad prediction performance. Outlier detection for such problems aims to remove noisy data or correct wrong data and to enhance the performance of regression models. Robust regression is a type of methods for noisy data when discarding noisy data is not ideal for the purpose of modeling (Wilcox, 2012). It performs better than the ordinary least squares in the presence of heavy tailed distributions. Most common methods of robust regression include M-estimation introduced by Huber (1964), least median squares (LMS) and least trimmed squares (LTS) (Rousseeuw, 1984). The general M-estimator minimizes the objective function

$$\sum_{i=1}^{n} \rho(e_i) = \sum_{i=1}^{n} \rho(y_i - X^T \beta) \tag{2.7}$$

where the symmetric function ρ gives the weight of each residual to the objective function. LMS orders all the residuals and replaces the sum of residual squares in the ordinary least squares with the median residual squares. LTS minimizes the sum of squared residuals over a subset of all points, and the residuals are ordered as well (Rousseeuw, 1984). Improved prediction accuracy is considered as the criterion to evaluate an outlier detection method for problems of regression (Brossart et al., 2011; Naseem et al., 2011; Rousseeuw and Leroy, 1987; Xu et al., 2008; Zhang et al., 2010). Naseem et al. (2011) used the Receiver Operating Characteristic curve (ROC curve) to indicate the predicted classification accuracy of their model, in order to evaluate the performance of their outlier detection model for face recognition. Xu et al. (2008)

conducted a linear robust regression using a recursive outlier-removal algorithm and evaluated the model by giving an increased R^2 of the regression.

Numerous outlier detection methods have been developed, from unsupervised learning to supervised learning, but it is still challenging for practitioners to select an appropriate one among these techniques and treat outliers correctly. In addition, because of the complexity of the outlier types and different objectives of studies, ad hoc outlier detection models should be developed individually.

2.4 Image Registration

Image fusion is a process to combine information from multiple images into a single image or data set, and image registration is a critical step for image fusion. Image registration aligns two or more images of the same scene, but taken at different times, from different angles, and by different sensors (Zitova and Flusser, 2003). In order to gain adequate and accurate information from the combination of images, studies have been done on image registration since the last decades (Brown, 1992). Image registration is being frequently utilized in many areas, such as medical imaging (Lester and Arridge, 1999; Maintz and Viergever, 1998; Van den Elsen et al., 2002), remote sensing (Fonseca and Manjunath, 1996; Le Moigne et al., 2002), microscopy images (Ozdemir and Casasent, 1999), and so on. Detailed surveys of these methods are presented by Brown (1992) and Zitova and Flusser (2003).

Based on the type of image acquisition methods, image registration problems are classified into three categories: different viewpoints, different times and different sensors to model registration (Zitova and Flusser, 2003). Since image sources and types of distortions are various, it is difficult to develop a universal image registration methodology applicable to all problems. Generally, image registration methods firstly detect features such as areas, lines and points from images. Then, they define a similarity measurement to match features. At last, they estimate transform models, which can transform an image into its target image (Russ, 2007). Many efforts have been made for feature detection methods (Goshtasby, 2005; Goshtasby and Stockman, 1985; Kapur and Casasent, 2000; Mahadevan and Casasent, 2003). The detected features from unregistered images need to be matched with matching methods, or similarity measure functions. The objective of matching is to find a feature correspondence, estimate parameters and then align two images together with the minimum misalignment errors. Many feature matching and similarity measurement techniques have been developed, including the correlation-like methods (Diniz, 2010), sequential similarity detection algorithms (Barnea and Silverman, 1972), projection-based registrations (Cain et al., 2002) and the Fourier methods (Bracewell, 2000; De Castro and Morandi, 2009). Detailed descriptions of these methods could be found in surveys of Lester and Arridge (1999) and Brown (1992). To estimate the transform model, approaches were developed for various specific problems. Among these approaches, the local mapping (Goshtasby, 1986; Wiemker et al., 1996), radial basis functions (Ehlers and Fogel, 1994; West et al., 1997) and elastic registration methods (Bajcsy and Kovacic, 1989; Wollny and Kruggel, 2002) are frequently adopted. Control point registration is a method which allows users to manually select common features in each image for mapping to the same location. It is best used for images with distinct features, but also with noises and distortions. It has been implemented in the image processing toolbox in Matlab. Since its performance relies on the manual selection of distinct features, this method is not computationally efficient and lack of stability. Few studies have been carried out on image registration for pairs of experimental images such as EBSD images and microscope images. Although many studies have been conducted to improve the performance of image registration, it is still one of the most difficult and critical tasks in image processing.

2.5 Background of Corrosion

2.5.1 Intergranular Corrosion of Aluminum Alloys

Aluminum alloys are widely used in marine, aerospace, automobile and manufacture industries because they have many advantages such as light weight, high strength-to-weight ratio and tunable strength. However, corrosion of aluminum alloys also limits their usage in these areas, especially for marine structures. Numerous studies have been carried out to improve the corrosion resistance of such alloys (Kim et al., 2001; Lo et al., 2009; Pan et al., 1996; Unwin et al., 1969; Yuan, 2006).

5XXX series alloys are used in this dissertation, which contain magnesium (Mg) more than 3 wt%. It has been shown that alloys with Mg more than 3 wt% can become saturated with Mg, leading to the formation of an intermetallic compound β (Al_2Mg_3). β usually precipitates along grain boundaries when exposed to temperatures higher than 70°C (Dix et al., 1958). Therefore, such alloy is susceptible to intergranular corrosion (IGC) due to the sensitization caused by active β which only precipitates along grain boundaries. Evaluation of intergranular attack can be difficult. For 5XXX series alloys, the susceptibility to IGC is quantitatively measured by ASTM-G67 standard test, known as the Nitric Acid Mass Loss Test (NAMLT), according to the Annual book of ASTM standards for Testing and Materials (2002).

2.5.2 Grain Boundary Characteristics

When two dissimilarly orientated crystals or grains meet, the space between them constitute a grain boundary on the surface (Yuan, 2006). Grain boundaries play an important role in corrosion science, because many studies have shown that a grain boundary provides a preferential site for precipitation. And grain boundary characteristics (GBCs), such as grain boundary misorientation, boundary plane and grain orientations, are critical factors in determining the precipitation rate and morphology (Butler and Swann, 1976; Garg and Howe, 1992; Vaughan, 1968; Yuan, 2006). Grain boundary misorientation (θ) is defined the angle required to rotate the set of crystal axes of one grain into coincidence with that of the other one in its neighbor. Mathematically, θ is calculated as

$$\theta = \min|cos^{-1}\{\frac{tr[O_{24} \cdot (g_1^{-1} g_2)] - 1}{2}\}| \tag{2.8}$$

where O_{24} is a set of 24 symmetry operators for cubic crystals, and g_1, g_2 are rotation matrices of grain 1 and grain 2. Intergranular corrosion (IGC) is a type of corrosion selectively attacking grain boundaries or closely adjacent regions without attacking grain bodies. It occurs between adjacent grain bodies, and different alloys show various electrochemically active paths for IGC.

Pan et al. (1996) and Unwin and Nicholson (1969) well studied the property and behavior of grain boundaries and concluded that they are strongly affected by local chemistry and atomic structures. The only GBC parameter considered correlating to IGC is misorientation (denoted as θ). It is sometimes expressed in terms of Coincidence Site Lattice (CSL). Chan et al. (2008) found that low-angle boundaries and grain boundaries with $\Sigma 3$ ($60^\circ\langle 111\rangle$) and $\Sigma 7$ ($38.21^\circ\langle 111\rangle$) are more likely to have higher corrosion resistance than other random boundaries. Grain boundaries with CSL equal to $\Sigma 13$ ($27.79^\circ\langle 111\rangle$) are also found to have higher corrosion resistance, but this is uncertain due to limited observations(only 18 boundaries). Unwin and Nicholson (1969) found that there were a large amount of β-phase on grain boundaries with θ in the range of 2° to 15°. Jakupi et al. (2010) also studied the corrosion resistance of $\Sigma 3$ grain boundaries and they stated that $\Sigma 3$ has high corrosion resistance in alloy 22 and the corrosion resistance capability decreases in the grain orientation order, which is $\langle 111\rangle > \langle 110\rangle > \langle 100\rangle$. Yuan (2006) has shown that in addition to misorientation, grain boundary plane also has an influence on IGC growth and β phase preferentially grew on $\langle 111 \rangle$ with $\langle 112 \rangle$ and boundaries with θ less than 20° are resistant to IGC. Kim et al. (2001) found that grain boundaries with θ below 15° can be hardly attacked by IGC in pure aluminum.

Although many efforts and studies have been made to explore the correlation between GBCs and IGC, physical origins underlying relationships between IGC and GBCs are still unknown and the utilized analysis approaches are very limited (Lo et al., 2009). In order to explain the underlying causal relationships and predict future IGC based on GBCs, a rigorous statistical analysis is needed.

CHAPTER 3

METHODOLOGY

As discussed in Chapter 1, methods are needed when: (1) relevant information is stored in a series of images; (2) images are noisy and distorted; (3) the data set for modeling is a structured data set with semi-continuous variables; and (4) the modeling goal is to understand the causal mechanism between variables and to predict future events accurately. In order to construct this formal methodology and solve this problem, three procedures have been considered and three models have been carried out respectively.

In this chapter, the statement of the problem to be modeled and the mathematical definition of the problem are firstly given in Section 3.1 and Section 3.2. These two sections address the objective of the developed methodology as well. Then, Section 3.3 describes the three components of the methodology.

Chapter 4, Chapter 5 and Chapter 6 describe the three components of this methodology in more details. Each chapter includes a problem definition where the constructed model is applicable. Chapter 4 gives data collection procedure, which gives an image fusion algorithm to extract information from multiple noisy images; Chapter 5 describes a data cleaning procedure, which provides an outlier detection framework to prepare clean data for supervised learning models; and Chapter 6 shows the final data modeling procedure, which models the data set and predicts future events based on a specially designed generalized hierarchical model for a special data type known as the semi-continuous data.

3.1 Problem Statement

As discussed in Chapter 1, semi-continuous data appear frequently in many scientific fields such as economy, ecology, physics, medicine and so on. It is a special type of data which includes a number of continuous data with a reoccurrence of some discrete numbers. Such data often include outliers and errors from the experiment and the measurement, which are not preventable. Modeling the semi-continuous data in the presence of noise is challenging because of the appearance of outliers and the skewness from the normal distribution within the data. Both of them may mislead the modeling result. Semi-continuous data in the presence of noise is common in material science studies. For instance, the variable *Percent β Coverage* records the percentage of the amount of corrosion occurred along each grain boundary from several aluminum alloy samples. These observations are collected from experimental images. This variable takes values from 0 to 1, with a large proportion of reoccurrences of 0's and 1's, corresponding to situations of no corrosion occurred and totally corroded. This special distribution is due to the grain boundary characteristics of different grain boundaries and this variable is considered as a semi-continuous variable.

Generally speaking, there are some scientific processes of which information is mainly stored in a series of noisy images with only few features. Data extracted from images are in the form of semi-continuity with outliers. Modeling the noisy semi-continuous data is a challenging and imperative task, because we want to understand the underlying causal mechanisms behind the data, but few available models are suitable for such data. This dissertation aims to construct a formal methodology to satisfy the need for the robust modeling of semi-continuous data in the presence of noise. The methodology is applied to an important example, which is the prediction of intergranular corrosion in AA5XXX-series alloys.

In order to model these processes, it is essential to fuse multiple noisy images together to extract relevant data stored in these images. This data collection step is known as image fusion. For instance, corrosion occurrences on materials are captured by optical microscopes in the form of gray scale images in corrosion science. Usually these images only have a limited number of features such as lines and dots.They are different from regular pictures because of the lack of meaningful composite features such as faces, special objects or buildings. To characterize features of materials, electron backscatter diffraction (EBSD) images are obtained by the HKL^{TM} Channel EBSD acquisition system (Day and Trimby, 2004). Fusing the optical microscope image with the EBSD image together could present valuable characteristics of the material which is being studied, and the corrosion evolution process on that material is able to be modeled with the extracted information. Many problems in other fields have similar characteristics, such as medicine and meteorology. More specifically, it is imperative to align multiple magnetic resonance images (MRI) taken from the same object at different times to monitor the tumor growth. It is also important to register several satellite cloud pictures to track meteorological changes and make weather forecasting based on the path changes of meteorological features.

Therefore, images are the main source of essential information to model event processes in many scientific fields. However, there are only a limited number of automated methods available to utilize these images. Take corrosion images as an example. Corrosion occurrences are usually detected by human eyes and then are measured manually regarding length or width of the corrosion events. This procedure not only limits the processing efficiency, but also extracts data inconsistently because of the objective views of different operators. In this way, these would lead to a data set with a number of biased data and a very limited size.

Even with a highly efficient image fusion method, wrong data can be unavoidably extracted from such noisy and sometimes distorted images, due to both the limitation of the image fusion method and measurement errors. For example, on the corrosion images, noises within grains due to the excessive exposure to the corrosive solution are easily considered as intergranular corrosion. Including such data in the intergranular corrosion data set can mislead the regression models built based on them. An outlier detection method is needed to filter out and remove these wrong data, in order to improve the quality of data for modeling. This step is considered as data cleaning.

Utilizing the clean data, a supervised learning method is constructed to make predictions of future events, as well as to provide the causal mechanism operating in the process. The data being used in this dissertation are from corrosion experiments with a special data type called semi-continuous data. The constructed supervised learning method is able to deal with semi-continuous data very well, by providing a high prediction accuracy and good model interpretability.

In summary, the objective of this dissertation is to develop a formal methodology using supervised learning methods for problems with characteristics:(1) relevant information is stored in a series of images; (2) images are usually noisy and distorted; (3) the data set for modeling is a structured data set with semi-continuous variables; and (4) the modeling goal is to understand the causal mechanisms between variables and to predict future events accurately.

3.2 Problem Definition

Suppose there is a process $\Omega = \{Y, X_1, X_2, \cdots, X_P\}$, where Y is the response variable of interest and X_i are features or predictors, $i = 1, 2, \cdots, P$. In the real world, it is almost impossible to observe the process Ω directly. But instead, what is available for observation is the process $\Omega^o = \{X_1^o, \cdots, X_m^o, Img_1, \cdots, Img_n\}$, where $\{X_1^o, \cdots, X_m^o\}$ are observations of features $\{X_1, \cdots, X_m\}$, and $\{Img_1, \cdots, Img_n\}$ are images. The difference between the two processes Ω and Ω^o is that the response variable Y and features $\{X_{m+1}, \cdots, X_P\}$ in Ω are only accessible from images $\{Img_1, \cdots, Img_n\}$ in Ω^o.

Therefore, to model the process Ω, the first step is extracting relevant information from images $\{Img_1, \cdots, Img_n\}$ in Ω^o. Mathematically, there is a need to find a function h

$$\{Y^o, X^o_{m+1}, \cdots, X^o_P\} = h(Img_1, \cdots, Img_n) + \epsilon_{img} \tag{3.1}$$

such that

$$dist(\{Y^o, X^o_{m+1}, \cdots, X^o_P\}, \{Y, X_{m+1}, \cdots, X_P\}) < \delta_{img} \tag{3.2}$$

where

(1) Y^o is the extracted response variable from the process Ω^o,

(2) $\{X^o_{m+1}, \cdots, X^o_P\}$ are extracted features from images $\{Img_1, \cdots, Img_n\}$ in Ω^o,

(3) h is a mapping from images to underlying information,

(4) ϵ_{img} is a random error,

(5) $dist()$ is a distance function measuring the difference between two sets of variables (such as Euclidean distances between two sets), and

(6) δ_{img} is a threshold.

After extracting valuable information from images successfully, the final objective is finding a function f characterizing complex relationships between the response variable Y and observed features $X^o_1, \ldots, X^o_m, X^o_{m+1}, \ldots, X^o_P$:

$$Y = f(X^o_1, \cdots, X^o_m, X^o_{m+1}, \cdots, X^o_P) + \epsilon \tag{3.3}$$

such that

$$||Y^o - \hat{Y}||^2 < \delta^* \tag{3.4}$$

where

(1) $\{X^o_{m+1}, \cdots, X^o_P\} \subset h(Img_1, \cdots, Img_n)$,

(2) Y^o is the extracted response variable in the future,

(3) $\hat{Y}$ is the prediction of model $f(\cdot)$,

(4) ϵ is a random noise, and

(5) δ^* is a threshold.

3.3 Methodology Development Strategy

In this dissertation, the developed formal methodology is considered as a system, which includes three components. The framework to implement the three components gives the solution to problems with the four characteristics described above. Each component indicates one individual component of the system, which is also able to solve a specific problem independently. The flowchart of the framework of this formal methodology is shown in Figure 3.1. As shown in Figure 3.1, the whole system can be divided into three components or subsystems: (1) the image fusion system with image data; (2) outlier detection system with noisy data; and (3) the modeling system with clean data. The final model for the problem of interest is the output of this whole system. Within each subsystem, a test procedure is included as a validation step. Only if the test result is reliable, it is possible to continue with next subsystem. The following three chapters describe the three subsystems in detail, and they constitute the content of methodology development in this dissertation.

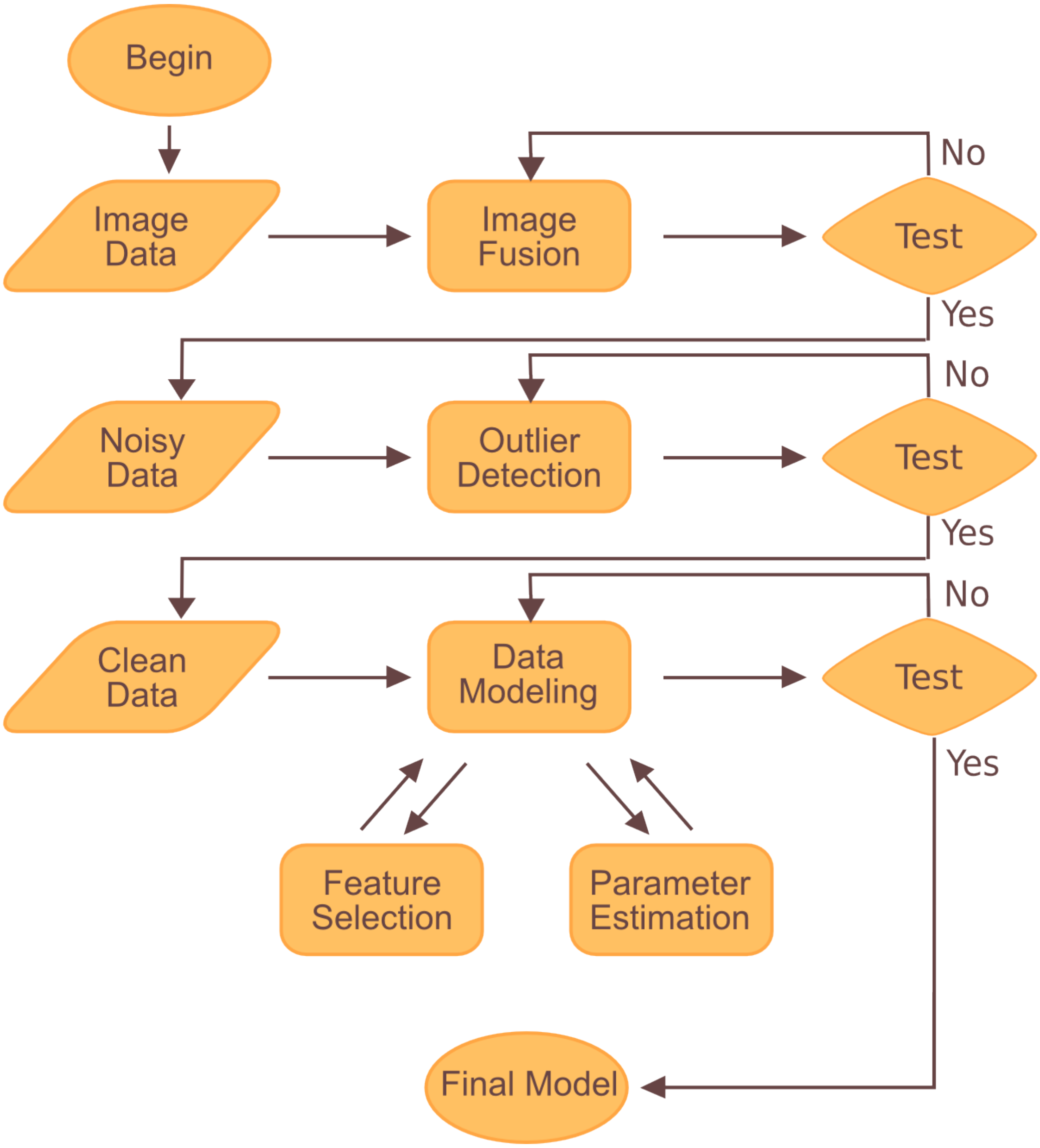


Figure 3.1: The flowchart of Methodology Development

CHAPTER 4

IMAGE FUSION

This chapter introduces the first subsystem in the developed formal methodology, the image fusion model. The problem is firstly defined formally and then developed algorithms are described in details.

4.1 Problem Definition

Image fusion is a general process of combining relevant information from two or more images into a single image or data set. The content of relevant information depends on the application under consideration. Images to be fused should be registered or aligned first, and therefore, image registration is a critical step for obtaining a highly accurate fusion result. This dissertation mainly focuses on the development of an effective image registration method for noisy and distorted experimental images, as the first model of the developed formal methodology.

Image registration is a process of setting up a point-by-point correspondence projection between two or more images, which are acquired either at different times, by different equipment, or from different points of view. This process is able to align two or more images together so that contents of relevant information stored within images can be extracted, combined and compared (Chu et al., 2010). Only when images are aligned successfully, the process of image fusion can be finished by extracting and combining information from aligned images.

According to studies of Sabuncu (2004), Zitova and Flusser (2003), Chu et al.

(2010) and Caner et al. (2006), the image registration problem can be theoretically defined as follows:

Let $\mathcal{G}_1 = \{G_{1i} \mid G_{1i} \subset \mathbb{R}^d, i \in \mathbb{N}\}$, and $\mathcal{G}_2 = \{G_{2i} \mid G_{2i} \subset \mathbb{R}^d, i \in \mathbb{N}\}$ be two sets of real valued images defined on $\mathbb{R}^d$, where $d \in \mathbb{Z}^+$ is the dimension of the images (usually $d \in \{2, 3\}$). For example, suppose that G_{1i} is an image of intergranular corrosion occurred on an area of aluminum alloy. G_{1i} is taken by an optical microscope at time t, then all images taken by this optical microscope at any time from 1 to $T, T \in N$ constitute $\mathcal{G}_1$ $(t \leq T)$. Similarly, if G_{2i} is a magnetic resonance image (MRI) of a tumor taken at some time point p, then $\mathcal{G}_2$ may include the series of MRI images taken at different times, such as $p+1, p+2, \ldots, p+N$. Equation 4.1 describes the mapping between G_{1i} and G_{2i} as

$$G_{1i} = H[\Psi(G_{2i})] + \Upsilon_i \tag{4.1}$$

where

(1) $G_{1i} \in \mathcal{G}_1$, $G_{2i} \in \mathcal{G}_2$;

(2) $\Psi : \mathbb{R}^d \mapsto \mathbb{R}^d$ is a geometric transformation that models the alignment;

(3) $H : \mathcal{G}_2 \mapsto \mathcal{G}_1$ captures variations across image sets;

(4) Υ_i is a random noise;

(5) $\parallel G_{1i} \parallel = \parallel \Psi(G_{2i}) \parallel$.

The objective of image registration techniques is to estimate the geometric transformation function Ψ by maximizing the alignment measure function $p : \mathcal{G}_1 \times \mathcal{G}_2 \mapsto \mathbb{R}$. The alignment measure function p measures the degree of alignment between two images to be registered. For 2-dimension images ($d = 2$), p can be defined as

$$p(G_{1i}, \Psi(G_{2i})) = \sum_{x=1}^{m} \sum_{y=1}^{n} I\{[G_{1i}]_{xy}, [H \circ \Psi(G_{2i})]_{xy}\} \tag{4.2}$$

and

$$I = \begin{cases} 1 & \text{if } [G_{1i}]_{xy} = [H \circ \Psi(G_{2i})]_{xy}, \\ 0 & \text{if } [G_{1i}]_{xy} \neq [H \circ \Psi(G_{2i})]_{xy}. \end{cases} \tag{4.3}$$

where

(1) $[G_{1i}]$ is an $m \times n$ image;

(2) $[G_{1i}]_{xy}$ is the value of pixel (x, y) on image G_{1i};

(3) $[H \circ \Psi(G_{2i})]_{xy}$ is the value of pixel (x, y) on the registered image of G_{2i}.

Thus, the registration problem can be phrased as an optimization problem:

$$\Psi^* = \arg\max p(G_{1i}, \Psi(G_{2i})|H) \tag{4.4}$$

Equation 4.4 defines the problem of interest, and addresses the objective of the image registration problem which is to find the optimal transformation function Ψ^* such that the degree of alignment between two images is maximized. In the next section, an image fusion algorithm is designed and implemented in this dissertation to achieve this objective, serving as the first subsystem of the whole system for the formal methodology, as shown in Figure 3.1.

4.2 Data Collection: Image Fusion

In this dissertation, the developed formal methodology is applied to an intergranular corrosion prediction problem as a methodology application. The provided image data are from corrosion experiment with noisiness and distortion. The desired image fusion technique for corrosion images should be able to firstly register a series of images with different kinds, taken by different equipments, but from the same surface area. And then it is able to extract valuable information from the registered image and calculate the value of relevant features as well.

As observed in Figure 4.1, corrosion images used in this dissertation have several special characteristics, which would prevent us from utilizing classical and popular image fusion algorithms directly. First of all, both types of images do not have as many features as pictures from our daily life. For example, there is no colorful or any meaningful objects shown on the optical microscope images of corrosion (the bottom one). Only grains with boundaries and noisy particles from the experiment. Similarly, the electron backscatter diffraction (EBSD) image (the top one) only has grains with different colors which indicate different characteristics of grains. Second, the optical microscope images of corrosion are noisy (black particles), especially for those which are highly corroded. When being implemented with widely used image fusion algorithms, such noisiness is usually not taken into consideration, so that the extracted information is very likely to include unwanted or misleading data. For example, it is possible that the black particles on the optical microscope image are considered as corrosion, which is not the true case. Third, EBSD images are actually not images taken by cameras, but projected images from numerical computations processed by the HKL software through some algorithms. Therefore, when it is not possible to detect some boundaries on EBSD images, they are generated by numerical interpolations. Fourth, EBSD images are distorted from a flat surface due to the 70° tilted sample placed in the chamber of EBSD acquisition system. Such distortion makes it more difficult to align the EBSD image with the optical microscope image which is taken from a flat surface. Fifth, these two types of images measure different properties of materials independently. Based on these special features and the addressed objective, an ad hoc image registration technique is designed and implemented with these images. The following sections introduce this technique which includes three algorithms working consequently to achieve the objective.

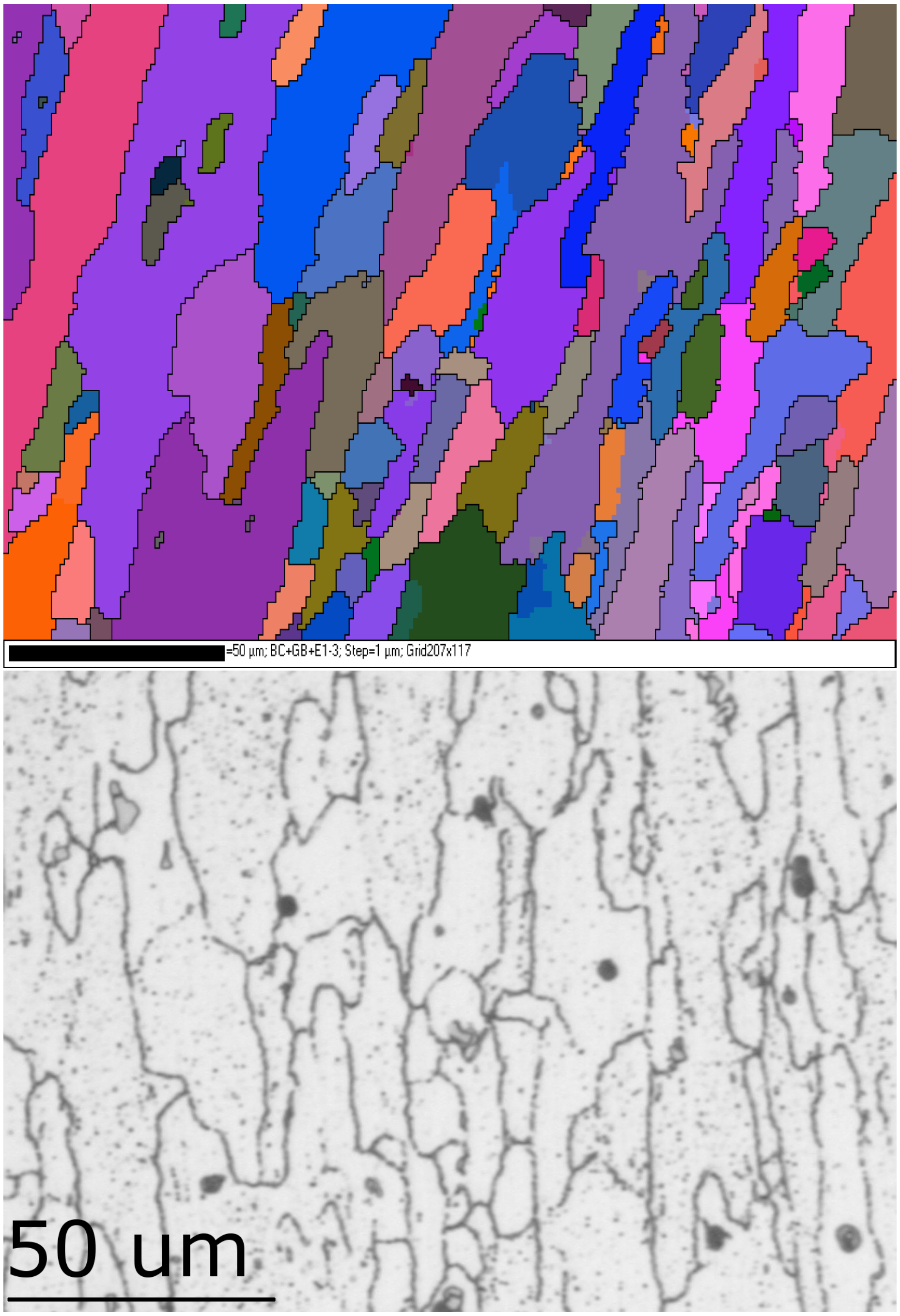


Figure 4.1: Example of a pair of images to be registered. Top: an EBSD image at Magnification=200X; Bottom: an optical microscope image of intergranular corrosion at Magnification=200X

Algorithm 1 computeIGC(): Image Registration with Percentage β Coverage Calculation

```
Require: $G_E$, $G_C$, $\alpha^*$, $\sigma^*$
$G_E^{bw} = bw(G_E)$,$G_C^{bw} = bw(G_C)$
$\{G_E^{adj}, G_C^{adj}\} = roughAdj(G_E^{bw}, G_C^{bw})$
$\{G_{E,b}\} = identifyBound(G_E^{adj})$,$B = size(\{G_{E,b}\})$
create a zero vector $CorrPect$, and $size(CorrPect) == B$
for $b = 1$ to $B$ do
  $\alpha \leftarrow 0$
  $CorrPect_b = cover(G_{E,b}, G_C^{adj}, \alpha)$
  while $\alpha \leq \alpha^*$ AND $CorrPect_b < \sigma^*$ do
    $\alpha \leftarrow \alpha + 1$
    $CorrPect_b = cover(G_{E,b}, G_C^{adj}, \alpha)$
  end while
end for
return $CorrPect$
```

Before explaining the proposed algorithms in detail, definitions of several single functions used in these algorithms are given in Table 4.1. All of these functions are available in Matlab and are coded already for use.

Function *computeIGC*() in Algorithm 1 is the main function in the method of image registration and the calculation of *Percentage β Coverage*. Inputs of the function *computeIGC*() include two different types of images, such as an EBSD image and an optical microscope image in this application, and two hyperparameters α^* and σ^*. α^* and σ^* can be estimated from a collection of test images. The procedures of

Algorithm 2 roughAdj(): Rough alignment of two images with resizing

```
Require: $G_1^{bw}$, $G_2^{bw}$, $\delta_{rough}$, $N_{max}$
$n \leftarrow 0$, $match \leftarrow 0$, $bestMatch \leftarrow 0$, $\Psi_{best} = \{rotate(0), enlarge(0), move(0)\}$
while $n < N_{max}$ AND $match < \delta_{rough}$ do
    $\Psi_n = \{rotate(\varphi_{r,n}), enlarge(\varphi_{e,n}), move(\varphi_{m,n})\}$
    $match = p(G_1^{bw}, \Psi_n(G_2^{bw}))/size(G_1^{bw})$
    if $bestMatch < match$ then
        $bestMatch \leftarrow match$
        $\Psi_{best} \leftarrow \Psi_n$
    end if
    $n \leftarrow n + 1$
end while
$G_{2r}^{bw} = resize(\Psi_{best}(G_2^{bw}))$, such that $size(G_1^{bw}) == size(G_{2r}^{bw})$
$G_{1f} = skeleton(G_1^{bw})$, $G_{2f} = skeleton(G_{1r}^{bw})$
return $\{G_{1f}, G_{2f}\}$
```

Algorithm 3 cover(): Computation of Percentage β Coverage on each boundary

Require: G_b, G_C, α

```
if $size(G_b! = size(G_C))$ then
    return error message
end if
$[M, N] = size(G_b)$
create a zero matrix $G_b^{r1}$ and $G_b^{r2}$, such that $size(G_b^{r1}) == size(G_b^{r2}) == [M, N]$
for $m = 1$ to $M$ do
    for $n = 1$ to $N$ do
        if $[G_b]_{[}m - \alpha : m + \alpha, n - \alpha : n + \alpha]! =$a square with only white color then
            if $[G_C]_{mn} ==$ black color then
                $[G_b^{r2}]_{mn} \leftarrow 1$
            end if
        end if
        if $[G_b]_{mn}! =$ white color then
            $[G_b^{r1}]_{mn} \leftarrow 1$
        end if
    end for
end for
$corrPect_b = \max\{1, \frac{\sum_m \sum_n [G_b^{r2}]_{mn}}{\sum_m \sum_n [G_b^{r1}]_{mn}}\}$
return $corrPect_b$
```

Function Name	Function Explanation
$bw()$	a function to transform color images into black and white images.
$size()$	a function to compute the size of a vector or matrix.
$rotate()$	a function to rotate images by a certain degree
$enlarge()$	a function to enlarge images by a certain percentage
$move()$	a function to move images by a number of pixels along a certain direction
$resize()$	a function to zoom images to a given size
$skeleton()$	a function to reduce widths of lines to the unit width using algorithms shown in Lam et al. (2002).

Table 4.1: Functions used in the developed image fusion algorithms

Algorithm 1 are introduced as follow.

First, both types of images are transformed into binary images using function $bw()$. Then they are roughly aligned together using Algorithm 2, which is described particularly in the following part. After a rough alignment, both images have an equal size. Objects, such as black lines on both images, are processed to have the same unit width. Next, based on the adjusted EBSD image G_E^{adj}, function $identifyBound()$ is utilized to generate a series of B images $G_{E,b}$, $b = 1, \ldots, B, B \in \mathbb{N}$. Each $G_{E,b}$ has the same image size as the adjusted image G_E^{adj}, but only contains one unique object, such as grain boundary in the corrosion example, from G_E^{adj}. Thus, B is equal to the total number of unique objects on image G_E^{adj}. Function $identifyBound()$ is able to be realized with the original feature data by the EBSD acquisition system HKL. Line 5 to Line 12 in Algorithm 1 are designed to compute the percentage of β coverage on each grain boundary. The general idea behind the calculation is to thicken each grain boundary on the EBSD image in order to make it fully cover the corresponding grain boundary on the optical microscope image as completely as possible, with certain constraints enforced by the hyperparameters α^* and σ^*. The output of Algorithm 1 is a vector of the percentage of β coverage for all grain boundaries detected by the EBSD image.

Function $roughAdj()$ in Algorithm 2 is a function to roughly align two images together. It assists the function performance of Algorithm 1. The inputs of function $roughAdj()$ are two black and white images G_1^{bw}, G_2^{bw}, along with two hyperparameters δ_{rough}, N_{max}. Here in the corrosion example, image G_1^{bw} can be the optical microscope image and G_2^{bw} can be the EBSD image. This function applies different kinds of transformations to image G_2^{bw} in order to maximize criteria defined in Equation 4.4. Different transformations include rotating, enlarging and moving images, as explained in Table 4.1. In Algorithm 2, function $p()$ is defined in Equation 4.3. After the optimal transformation combination is found, two images are resized to an equal size. Objects such as black lines on both of them are skeletonized to the unit width. The output of function $roughAdj()$ is a pair of roughly aligned images with the same image size.

Function $cover()$ in Algorithm 3 computes the percentage of β coverage on each boundary detected by the EBSD image. The inputs of Algorithm 3 are two images G_b, G_C and a parameter named as the thickening factor α. This algorithm can be considered as a two-stage process. In the first stage, the width of each object (black lines, for instance) on G_b is thickened from the unit width to the width of $2\alpha+1$. The position of each object does not change. In the second stage, the algorithm covers the thickened line on G_C and then counts the total number of black pixels within the covered area. The percentage of β coverage on each grain boundary is calculated as a normalized term of the total number of covered pixels.

4.3 Evaluation

As discussed in Chapter 2 Section 2.4, numerous image registration algorithms have been developed in image processing, but few existing techniques of image registration are applicable to images from experiments with few features such as the EBSD experiment and optical microscopes. Noisy pixels are shown on such images due to the limitation of the experiment, and the geometric distortion between the EBSD image and the microscope image also prevents the application of many available methods. Our image registration algorithms are especially developed for such images, in order to align both images and extract valuable data and relevant properties of the objects, where images are taken from.

To evaluate the performance of our image registration algorithms, they were tested on an EBSD image and a microscope image of the intergranular corrosion from the aluminum alloy sample of DoS (degree of sensitization) = 49 mg/cm^2. In order to measure the performance of the image fusion algorithms, intergranular corrosion was manually outlined on all grain boundaries (252 grain boundaries), and the average percentage of β coverage (E_{igc}) was calculated as $E_{igc} = 0.7693$. It is assumed that a well-performed algorithm should give an average percentage of β coverage as close to E_{igc} as possible.

The result of our image registration algorithms were compared with that of an existing widely used technique named as Control Point Registration (Ingle and Proakis, 1999), which is implemented in Matlab. Details of this technique are introduced in Chapter 2 Section 2.4. E_p is denoted as the result of our image fusion algorithms, and E_m is denoted as the result of the Control Point Registration method. The Control Point Registration method requires users to identify several pairs of points on both images at first, and then uses those points as references to complete the registration operation. In this evaluation, ten trials have been tried with different groups of paired points. The average value of mean percent β coverage of 252 grain boundaries this method output was $E_m = 0.1056$ (sd $= 0.0543$). The fusion algorithm has been applied to the same image, with ten different pairs of hyperparameters. Our fusion algorithm gave an average result of mean percent β coverage on 252 grain boundaries as $E_p = 0.7273$ (sd $= 0.0774$). Because the mean percent β coverage on the manually outlined image is $E_{igc} = 0.7693$, the fusion algorithm performs much better than the Control Point Registration method.

Also, the distributions of the percent β coverage calculated from the manually outlined image and the fused image were compared. Distribution plots are shown in Figure 4.2. Similar patterns are displayed on both of the distributions. Furthermore, the two-sample Kolmogorov-Smirnov test (K-S test) (Lilliefors, 1967) has been utilized to test whether the two datasets are significantly different from each other. The null hypothesis of the K-S test is that the two samples are drawn from the same distribution. The p-value of this test was 0.1108, which means we failed to reject the null hypothesis at the 5% significance level.

The evaluations show that our image fusion model performs better than the Control Point Registration, in terms of the accuracy of the data extraction, and our image registration algorithms are more suitable for image registration problems with images from the EBSD experiment and the optical microscope.

In summary, this chapter introduces three algorithms which have been designed and implemented with images from corrosion experiments. These three algorithms constitute the formal image fusion algorithm desired in this dissertation, which is served as the first subsystem in the whole system shown in Figure 3.1.

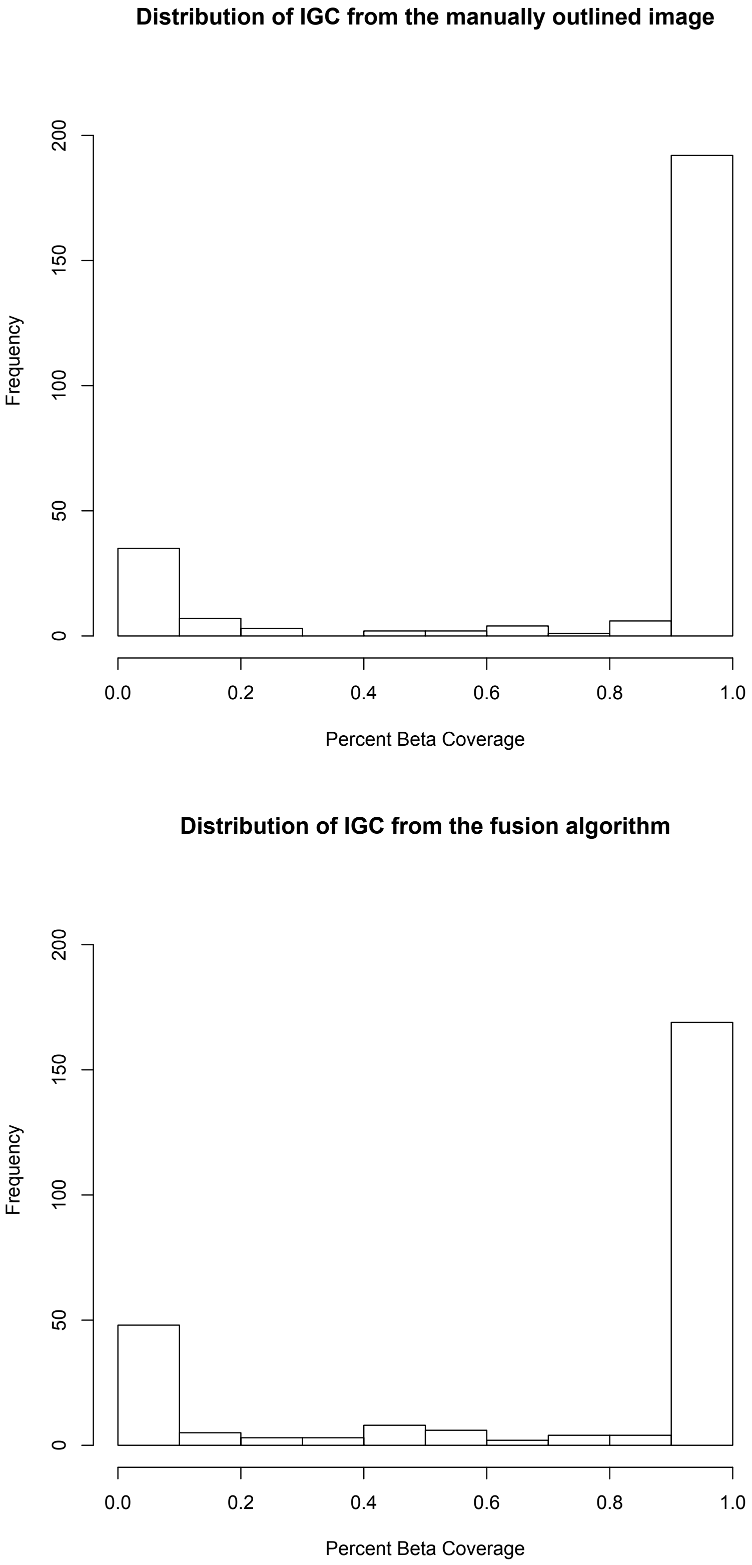


Figure 4.2: Distributions of percent β coverage calculated from manually outlined image and the fused image

CHAPTER 5

OUTLIER DETECTION

This chapter introduces the developed outlier detection framework. The formal problem definition and the detailed algorithm are described here.

5.1 Problem Definition

Outliers are usually considered as errors or noise in the dataset, and they are defined as "observations appearing to be inconsistent with the remainder of that data set" (Barnett and Lewis, 1994; Hawkins, 1980; Johnson and Wichern, 2002). Sometimes outliers within a dataset are carrying important information, but they may mislead the statistical modeling and affect the prediction accuracy as well. Therefore, it is desirable to identify these significant outliers and treat them properly before constructing models using these data. The outlier detection problem is a popular topic and has been studied for a long time in statistical learning. It can be defined in many ways and from different perspectives, such as the distance-based method, density-based method and residual-based method (Chatterjee and Hadi, 1986; Hoaglin and Welsch, 1978; Rousseeuw and Leroy, 1987; Velleman and Welsch, 1981). In this dissertation, the outlier detection problem is defined as follows in terms of the objective of this study.

Suppose the extracted data from the image fusion step are $D = [X, Y]$, where X are explanatory variables and Y is the response variable. Because the images where the data are extracted from contain noises and the method of image fusion is not perfect, D contains noises and errors as well. These noises and errors are considered as outliers. Our final objective is to build a statistical model $E[Y] = f_R(X)$ which is able to reveal the underlying relationship between X and Y, where $E[Y]$ is the expectation of Y and $f_R(X)$ is the model of X. To achieve this objective, an outlier detection model $f_O(D)$ is constructed which can provide us with a better data set than D. Formally, the outlier detection problem can be formulated as follows:

An outlier detection method $f_O(D)$ is said to be *useful* if it satisfies:

$$MSE[f_R(f_O(D))] \leq MSE[f_R(X)] \tag{5.1}$$

where $MSE[\cdot]$ means the mean square errors of the statistical model f_R; the smaller MSE is, the better the model is; here $MSE[\cdot]$ can also be exchanged with other loss functions such as the mean absolute errors (MAE); $f_O(D) = \{D^M, D^O, I_{action}\}$ is the outlier detection method, which has three types of outputs: the majority data set D^M including most of the data in D, the outlier data set $D^O = D - D^M$ including the rest data in D; and the indicator function I_{action} indicating how to treat the detected outliers.

Our objective is to find an useful outlier detection model $f_O(D)$ for the data set D extracted from images. The next section introduces the outlier detection framework developed in this dissertation which is compatible with different supervised learning methods. In the corrosion problem where this method is applied, it is built on a two-part generalized hierarchical model to filter out outliers and improve its prediction performance. The outlier detection framework is considered as a data cleaning subsystem in the formal methodology, shown in Figure 3.1.

5.2 Data Cleaning: Outlier Detection

In order to minimize the influence of outliers on the prediction performance of a supervised learning model, an outlier detection framework based on boosting is developed as follows. Boosting is a supervised learning method, which is capable of enhancing the accuracy of a statistical learning method. It combines a group of learners to enhance their performance. This idea is one of the most successful learning methods in the last decade (Friedman et al., 2001). Here in this dissertation, boosting is applied with the outlier detection framework, so that the accuracy of this method is greatly enhanced, compared with other available outlier detection methods. This framework is able to detect outliers for general supervised learning models and determine outliers automatically. The details of this framework are addressed below.

- Inputs

  A training data set $T = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_n, y_n)\}$.

- Procedures

  1. For $k = 1$ to K, do

     (a) Take a random sample S from T with size N, $N < n$

     (b) Build a general regression model on sample S: $y_i = f(\boldsymbol{x}_i) + \epsilon_i$.

     (c) Label the i^{th} observation as 0 or 1 in terms of its fitted square residual $\gamma_i = ||y_i - f(\boldsymbol{x}_i)||^2, i = 1, \ldots, N$. The label function $L_i^{(m)}$ is defined as:

$$L_i^{(m)} = \begin{cases} 0 & \text{if } \gamma_i \leq \delta \\ 1 & \text{if } \gamma_i > \delta. \end{cases} \tag{5.2}$$

     where $\delta = \frac{1}{N}\sum_{i=1}^{N} \gamma_i + sd(\boldsymbol{\gamma}), \boldsymbol{\gamma} = (\gamma_1, ..., \gamma_N)$. δ is the outlier threshold.

2. Compute $V_i = \sum_{k=1}^{K} L_i^{(k)}$.

   V_i is the vote as an outlier for the i^{th} observation over the K samples. Let y' be a binary response variable defined as:

$$y'_i = \begin{cases} -1 & \text{if } V_i \geq \alpha^* \\ +1 & \text{if } V_i < \alpha^* \end{cases} \tag{5.3}$$

   where α^* is the threshold for V_i.

3. Partition T into two subsets: M (the majority set) and O (the outlier set).

   $M = \{(x_i, y_i) \mid y'_i = -1, i = 1, 2, ..., n\}$, $O = \{(x_i, y_i) \mid y'_i = +1, i = 1, 2, ..., n\}$.

4. Do classification on training set T' with AdaBoosting (Freund and Schapire, 1995).

   $T' = \{(x_1, y'_1), (x_2, y'_2), ..., (x_n, y'_n)\}$ and the classification error is calculated by $\eta = (\text{False Positive} + \text{False Negative})/n$.

5. Modeling strategy for two subsets M and O is established.

   - If η is larger than η^*, then build model f_T with all of observations and no outliers are detected.
   - If η is less than η^* and the size of the outlier set O is less than a threshold s^*, then drop set O and only model the majority set M with f_M.
   - If η is less than η^* and the size of O is large than s^*, then fit set O and M with models f_O and f_M, respectively.

   This modeling strategy is summarized in Table 5.1.

- Outputs

Classification Error	Size of the Outlier Set	Outlier Treatment	Model
$\eta \geq \eta^*$	$s < s^*$ or $s \geq s^*$	No outlier	f_T
$\eta < \eta^*$	$s < s^*$	Drop outliers	f_M
$\eta < \eta^*$	$s \geq s^*$	Fit outliers	f_O and f_M

Table 5.1: Strategy of Outlier Detection Framework

The integrated outlier detection framework model can be written as:

$$f(x) = \begin{cases} f_T(x) & \text{if } \eta \geq \eta^* \\ f_M(x) & \text{if } \eta < \eta^*, s < s^* \\ f_M(x) + f_O(x) & \text{if } \eta < \eta^*, s \geq s^* \end{cases} \tag{5.4}$$

where

(1) s is the size of the outlier set O;

(2) η is the classification error; and

(3) s^* and η^* are threshold values of s and η respectively.

- Parameters

  s^* and α^* are estimated by the cross-validation algorithm (Golub et al., 1979) to get the minimum mean square error. K and η^* are pre-defined parameters. Here it is tested that $K = 10$ and $\eta^* = 0.3$ performs the best on the corrosion data.

5.3 Evaluation

This outlier detection framework (ODF) is especially developed for supervised learning methods. It uses Adaboosting as a classification tool to vote for outliers based on residuals. Basis functions can be supervised learning methods such as linear regression, support vector machine (SVM), random forest and so on. When outliers are detected by this framework, they are either removed from the data or are used for a different model. To apply it to real problems, the objective of applicable problems should be enhancing the prediction accuracy of the models. As discussed in Chapter 2 Section 2.3, current major outlier detection methods usually include distance-based methods, density-based methods and robust regression models such as M-estimation, least median of squares and least trimmed of squares. For the purpose of regression, robust regression plays an important role in the presence of outliers. To evaluate the performance of our outlier detection framework, three groups of different comparisons have been conducted.

Because the purpose of the ODF is to enhance the prediction accuracy of a regression model, the first comparison is between the ODF based on a linear regression model and three common robust regression models, robust linear model with M-estimation (RLM), least median of squares (LMS) and least trimmed of squares (LTS). The predicted mean absolute error (PMAE) is used as the criterion to compare their prediction performances. The intergranular corrosion data extracted from noisy images are used here for the comparison. Using a test set evaluation method with 50 replicates, the ODF with a linear regression model has a mean PMAE of 0.249 with a 95% confidence interval of (0.247, 0.252). The mean PMAE is 0.265 (0.263, 0.266) for RLM, 0.306 (0.293, 0.320) for LMS, 0.349 (0.334, 0.364) for LTS. Because the confidence interval of ODF has no overlap with other confidence intervals, the differences are statistically significant at the level of 0.05. These results show that the ODF with an ordinary least square regression outperforms the three major robust regression models in terms of PMAE, and it is effective to enhance the prediction accuracy of a regression model with the presence of outliers. Detailed comparison results are shown in Section 7.5.2. The breakdown point of this outlier detection technique is $1/n$ because the mean residual is used as the estimator.

The second comparison is to test the effectiveness of the ODF by comparing the PMAE of each of four different supervised learning methods with and without the ODF. In this evaluation, four widely used supervised learning methods are applied. They are support vector machine with a radial basis (SVMR), support vector machine with a linear basis (SVML), random forest (RF) and the linear hierarchical model (MLM). The test is still taken on the intergranular corrosion data from noisy images. The regression results show that with ODF, the prediction performances of all four models are improved significantly because of a lower value of PMAE. Detailed comparison results are shown in Section 7.5.2.

The last evaluation is to test if there is a significant difference between the outlier set and the majority set defined in this outlier detection framework. The Multiple Response Permutation Procedure (MRPP) in Mielke et al. (1976), Good and Wang (2005) and Park et al. (2009) are used to do this evaluation. In MRPP, a new data set is generated by permuting each predictor in the original data set while keeping each observation's outlier label unchanged. Here, the outlier label is obtained from the ODF. This randomized data set is fitted by four different regression models SVML, SVMR, RF and MLM. Adaboosting is used to classify if an observation is an outlier or from the majority set on the training data. Test result shows that the classification error obtained from the randomized data increases, and the PMAE of each of the regression models increases as well. This suggests that there exist clusters in the original data set, which are detected by the ODF successfully. But the permutation after MRPP disrupts the clustering patterns, which leads to a poor prediction performance for the randomized data. Therefore, it is shown that the ODF is able to detect clusters within the data based on different regression models and it is effective to detect the unusual patterns hidden in the data set. Detailed comparison results are shown in Section 7.5.2.

CHAPTER 6

GENERALIZED HIERARCHICAL MODELING

This chapter gives the mathematical definition of the generalized hierarchical modeling approach, and discusses its applicability. Then the characteristics of semi-continuous data are introduced. A two-part generalized hierarchical model is built for semi-continuous data, which is the third model in the formal methodology in this dissertation.

6.1 Problem Definition

Suppose a statistical model

$$Y = f(X) + \epsilon \tag{6.1}$$

is a reasonable assumption for a dataset $D_{n \times m}$, where

(1) Y is an $n \times 1$ response vector;

(2) X is the covariate matrix;

(3) ϵ is the random error vector with $E(\epsilon) = 0$;

(4) ϵ is independent of X.

Supervised learning attempts to learn the function $f(\cdot)$ by examples through a learning algorithm. A training dataset of observations $T = (x_i, y_i), i = 1, \ldots, N, N \leq n$ is assembled, and the learning algorithm produces outputs $\hat{f}(x_i)$ in response to the inputs in T. The learning algorithm is able to modify $\hat{f}(\cdot)$ by adjusting a group of its parameters Θ corresponding to the difference $y_i - \hat{f}(x_i)$. Once the learning process is completed, the objective of the supervised learning model is to minimize the difference between the predicted outputs and the actual outputs for all sets of inputs by estimating the parameters Θ in $\hat{f}(\cdot)$ (Friedman et al., 2001). In summary, in terms of the squared error, this objective can be represented as

$$\min\ \mathcal{L}(\Theta) = \sum_{i=1}^{N}(y_i - \hat{f}_\Theta(x_i))^2 \tag{6.2}$$

where

(1) y_i is the real output for the i^{th} observation;

(2) $\hat{f}_\Theta(x_i)$ is the artificial output of function $\hat{f}(\cdot)$;

(3) Θ is the parameter vector for function $\hat{f}(\cdot)$; and

(4) $\mathcal{L}(\Theta)$ is a loss function of Θ.

The generalized hierarchical model is one of many supervised learning models within the supervised learning framework above. This dissertation focuses on the development of a specific type of generalized hierarchical model developed for semi-continuous data type. The next section addresses the need for such a model, and its construction procedures. The developed generalized hierarchical model for semi-continuous data is applied to the corrosion prediction problem in Chapter 7, in order to provide users with a high prediction accuracy and good model interpretability.

6.2 Data Modeling: A Generalized Hierarchical Model for Semi-continuous Data

6.2.1 Semi-continuous Data

Semi-continuous data are often characterized as a mixture of non-zero continuously distributed values and a certain proportion of repeated single values, such as 0's (Olsen and Schafer, 2001). Such data appear frequently in economics, ecology, physics and medicine. For example, the research about the drinking outcomes among alcohol-dependent individuals in Liu et al. (2008) has individual alcohol consumption data, which include positive and continuous values and a large amount of zeros. In epidemiology, data often include both zeros for those areas without certain disease, and positive values indicating the degree of severity of diagnosed epidemic cases in other places. Semi-continuous data are also common in material science studies. Here in the intergranular corrosion problem we are solving, the response variable *Percent β Coverage* records the percentage of the amount of corrosion occurred along each grain boundary from several aluminum alloy samples. This variable takes values from 0 to 1, with a large proportion of reoccurrences of 0's and 1's, corresponding to situations of no corrosion occurred and totally corroded. This special distribution is due to the grain boundary characteristic differences among grain boundaries and we consider such a variable as a semi-continuous variable.

Analysis of the semi-continuous data is challenging. For instance, the ordinary least squares method performs not quite well because of the presence of skewness from the normal distribution within the data, which means that the assumption about the normal distribution is not satisfied. Such skewness can not always be removed by transformations on the data (Stanghellini and Gottard, 2011). In this dissertation, we present a two-part generalized hierarchical model with a basis of the generalized additive model for such data. Figure 6.1 shows an example of the distribution of semi-continuous data. As shown in Figure 6.1, the semi-continuous data can be characterized as a combination of two types of sample data: 1) samples from a continuous probability distribution, and 2) samples from a Bernoulli distribution. Unlike sampling from a continuous distribution (e.g. Gaussian distribution or the uniform distribution), the semi-continuous data do not have zero probability at the maximum and the minimum values.

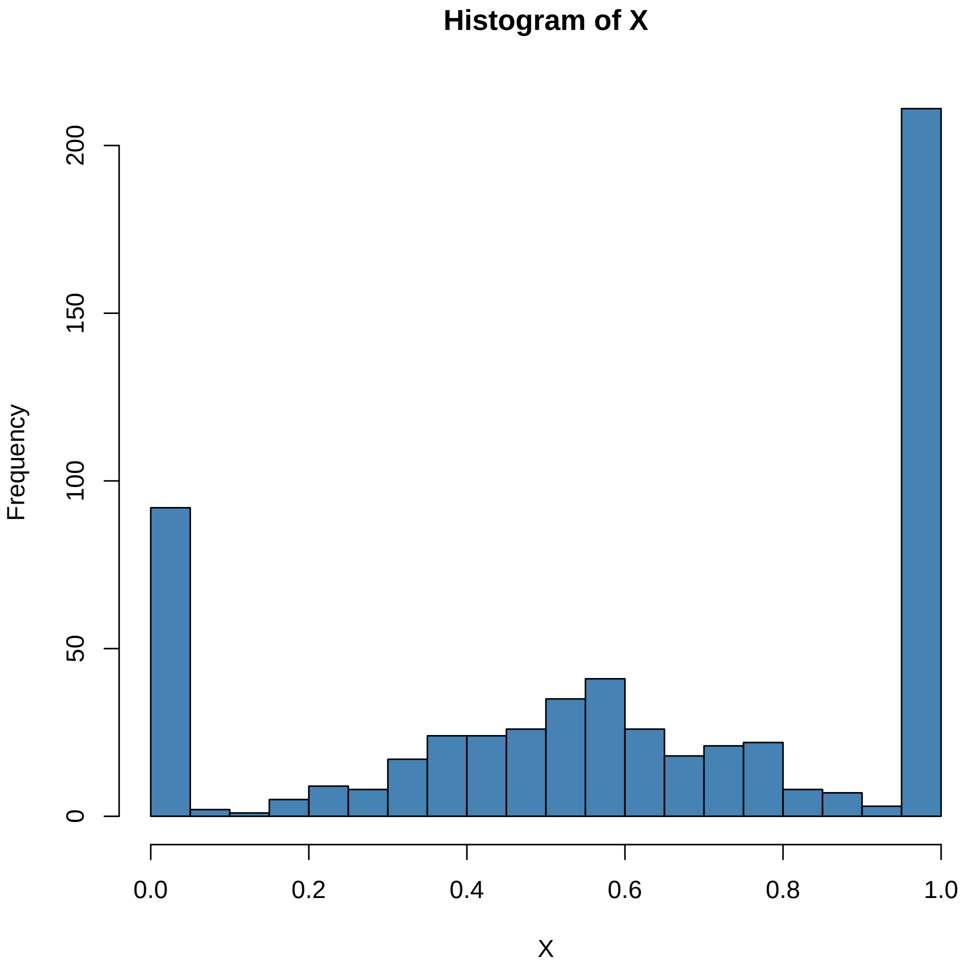


Figure 6.1: An example of the distribution of semi-continuous data

Before defining the semi-continuous data, the definition of the semi-continuous distribution is given below.

Definition 1 *A random variable X is said to have a* ***semi-continuous distribution*** *if X and its density $f(x)$ satisfy the following conditions:*

$$C_1 \leq X \leq C_2 \tag{6.3}$$

$$f(x) = w_1 p_1(x) + w_2 p_2(x) \tag{6.4}$$

$$w_1 + w_2 = 1, w_1 > 0, w_2 > 0 \tag{6.5}$$

$$p_1(X = k) = \begin{cases} q & \mathit{if}\ k = C_1 \\ 1 - q & \mathit{if}\ k = C_2 \end{cases} \tag{6.6}$$

In Definition 1, p_1 is the density function of a Bernoulli distribution, and p_2 is the density function of a continuous variable. C_1 and C_2 are two constants, and $C_1 < C_2$. If $w_1 \to 1$, then X is close to a discrete random variable. If $w_2 \to 1$, then X is close to a discrete variable. Definition 1 shows that X is different from a discrete random variable, because X is able to take an infinite number of values ranging from C_1 to C_2. On the other hand, X is different from a continuous random variable, because X has non zero probability at the maximum and the minimum values. Therefore, the semi-continuous data is defined as follows.

Definition 2 *The data is said to be the* ***semi-continuous data*** *if it is a sample generated from a semi-continuous distribution.*

This dissertation focuses on the problem of modeling the semi-continuous data as a response variable in the data set, due to the property of the corrosion data which the generalized hierarchical model is applied with. After defining the semi-continuous data, the problem where a special supervised learning method is needed is well defined as well in the next section.

6.2.2 Mathematical Definition of the Problem

Suppose there is a sample $S_y = \{y_1, \cdots, y_n\}$ from a semi-continuous distribution. S_y represents the semi-continuous measurement outcomes. For each $y_i \in S_y$, there is a vector of features associated with it: $\mathbf{x}_i = (x_{1i}, x_{2i}, \cdots, x_{ni})$.

Each problem with the semi-continous data which needs a supervised learning model usually has two parts. The first part is a classification problem. As it is shown in Definition 1, there are two different parts of S_y, discrete values $\{C_1, C_2\}$ and continuous values $\{c|C_1 < c < C_2\}$. Take the intergranular corrosion prediction problem as an example. It is known that $C_1 = 0$ means grain boundaries without corrosion, $C_2 = 1$ means completely corroded grain boundaries, and $C_1 < c < C_2$ denotes partially corroded grain boundaries. It is necessary to classify the two extremes, in order to study the patterns separately.

In other words, a classification function $f_{classify}(\mathbf{x}) \in \{Label_{C_1}, Label_{C_2}, Label_c\}$ is desired, where $Label_{C_1}$ and $Label_{C_2}$ indicate that y_i is likely to have the minimum and maximum value respectively; and $Label_c$ means that y_i is not likely to have the extreme values. Ideally, the classification function $f_{classify}(\mathbf{x_i})$ can be represented as:

$$f_{classify}(\mathbf{x_i}) = \begin{cases} Label_{C_1} & \text{if } y_i = C_1 \\ Label_{C_2} & \text{if } y_i = C_2 \\ Label_c & \text{if } C_1 < y_i < C_2 \end{cases} \tag{6.7}$$

The overall objective is to find a function $f(\mathbf{x}_i)$ to fit and predict the value of y_i. Mathematically, a function $f(\mathbf{x}_i)$ is needed to satisfy $\sum_i |y_i - f(\mathbf{x}_i)| < \epsilon$ for a small ϵ. The smaller ϵ is, the better model $f(\mathbf{x}_i)$ is. In next section, a two-part generalized hierarchical model is built for the semi-continuous data, in order to solve the problem stated above.

6.3 A Two-Part Generalized Hierarchical Model for Semi-Continuous Data

This section describes the development of a two-part generalized hierarchical model for solving the problem addressed in Section 6.2.2. This model is considered as a framework consisting of two generalized additive models (GAMs). Section 6.3.1 gives the model definition and Section 6.3.2 discusses an algorithm to estimate the model parameters.

6.3.1 Model Definition

As defined in Section 6.2.2, two models are need to be built: $f_{classify}(\mathbf{x})$ and $f(\mathbf{x})$. A two-part generalized hierarchical model consisting both of the two models has been constructed to solve the problem. First, a classification model $f_{classify}(\mathbf{x})$ with all the data is built. Then, given $f_{classify}(\mathbf{x})$, the overall model $f(\mathbf{x})$ is defined as follows.

$$f(\mathbf{x}) = \begin{cases} C_1 & \text{if } f_{classify}(\mathbf{x}) = Label_{C_1} \\ C_2 & \text{if } f_{classify}(\mathbf{x}) = Label_{C_2} \\ f_{regression}(\mathbf{x}) & \text{if } f_{classify}(\mathbf{x}) = Label_c \end{cases} \tag{6.8}$$

Equation 6.8 is the overall definition of the two-part generalized hierarchical model. Model part I is a classification model $f_{classify}$, and model part II is a hierarchical regression model $f_{regression}$. Generally, any classification model, such as tree, support vector machine, or neural network, are applicable. Similarly, any regression model, such as a linear regression model, is suitable. Based on the discussion in Section 6.1, a supervised learning model with good prediction performance as well as good interpretability is desirable for problems such as the corrosion prediction. Furthermore, finding out the individual impacts of grain boundary characteristics on the growth of intergranular corrosion is another objective for the application problem. Therefore, the Generalized Additive Model (GAM) for both $f_{classify}$ and $f_{regression}$ is chosen as the base model. GAM assumes that there is additivity between individual features, and it also allows nonlinear impacts from each feature on the response variable. Because of the additivity, individual impacts are able to be separated. Because of the nonlinearity, each feature can be modeled more flexibly than with regular linear regression models.

Next, the definition of the first part of the model, $f_{classify}$ is given below.

Model Part I: The Classification Model

$$f_{classify}(\mathbf{x_i}) = \begin{cases} Label_{C_1} & \text{if } Pr(\mathbf{x_i}) < \delta_l \\ Label_{C_2} & \text{if } Pr(\mathbf{x_i}) > \delta_u \\ Label_c & \text{if } \delta_l \leq Pr(\mathbf{x_i}) \leq \delta_u \end{cases} \tag{6.9}$$

and

$$\log(\frac{Pr(\mathbf{x_i})}{1 - Pr(\mathbf{x_i})}) = \sum_{j=\{1,\cdots,J\}} f_j^c(x_j) \tag{6.10}$$

where

(1) $Pr(\mathbf{x_i})$ is the probability of $y_i = C_2$;

(2) δ_l and δ_h are two threshold values satisfying $0 < \delta_l < \delta_h < 1$;

(3) J is the total number of features; and

(4) $f_j^c(x_j)$ is the smooth function of feature x_j.

Equation 6.10 is a generalized additive model which predicts the probability of $y_i = C_2$. Equation 6.9 is a decision function mapping from the probability of $y_i = C_2$ to three different labels: when the probability is small ($Pr(\mathbf{x_i}) < \delta_l$), y_i is assumed to have the value of C_1; when the probability is large ($Pr(\mathbf{x_i}) > \delta_u$), y_i is assumed to take the value of C_2; if the probability is between the two thresholds ($\delta_l \leq Pr(\mathbf{x_i}) \leq \delta_u$), y_i is assumed to be between C_1 and C_2.

Given model part I, the second part of the model is defined as follows.

Model Part II: The Hierarchical Regression Model

$$E[y_i|\mathbf{x}_i] = \begin{cases} C_1 & \text{if } f_{classify}(\mathbf{x_i}) = Label_{C_1} \\ C_2 & \text{if } f_{classify}(\mathbf{x_i}) = Label_{C_2} \\ f_{regression}(\mathbf{x}_i) & \text{if } f_{classify}(\mathbf{x_i}) = Label_c \end{cases} \tag{6.11}$$

and

$$f_{regression}(\mathbf{x}_i) = E[y_i|\mathbf{x}_i] = g^{-1}(\sum_{j=\{1,\cdots,J\}} f_j^r(x_j)) \tag{6.12}$$

where

(1) $g(\cdot)$ is a link function and it is used to constrain the response variable to (C_1, C_2). A popular choice of link function can be the logit function or the probit function.

(2) J is the total number of features;

(3) $f_j^r(x_j)$ is the smooth function of feature x_j.

(4) Equation 6.12 is equivalent to a generalized additive model $g(E[y_i|\mathbf{x_i}]) = \sum_{j=\{1,\cdots,J\}} f_j^r(x_j)$

Equation 6.12 is a generalized additive model constrained by a link function $g(\cdot)$. It is different from the one in Equation 6.10. First, Equation 6.10 predicts the probability of the event that $y_i = C_2$, while Equation 6.12 predicts numerical values between C_1 and C_2. Second, these two models are estimated with different sets of data.

To estimate the model in Equation 6.10, the response variable should be a binary categorical variable. To estimate the model in Equation 6.12, the response variable should be a continuous numerical variable. Therefore, smooth functions $f_j^c(x_j)$ and $f_j^r(x_j)$ in the two models are different.

From the above model definition, it is clear that the basic idea of the two-part generalized hierarchical model is to build a classification model on the entire data set first, in order to classify the discrete extreme values (e.g. 0 and 1, in the corrosion example) and the continuous data between extreme values. Then a hierarchical regression model is constructed for the continuous data only.

In the developed two-part generalized hierarchical model, two types of parameters are need to be estimated. One type is the threshold parameter δ_l and δ_u. The other type is the smooth functions $f_j^r(x_j)$ and $f_j^r(x_j)$. Section 6.3.2 discusses details of how to estimate both types of parameters. Given the estimated model parameters, the prediction with new data $\{\mathbf{x}_{new}\}$ is straightforward. Model part I (Equation 6.9 and 6.10) is used to predict labels, and model part II (Equation 6.11 and 6.12) is used to predict outcomes $\{y_{new}\}$.

6.3.2 Model Estimation

As discussed in Section 6.3.1, the two generalized additive models $f_{classify}$ and $f_{regression}$, as well as the two threshold parameters, δ_l and δ_u, need to be estimated. Generalized additive models have been studied extensively in statistics. There are several well developed algorithms to estimate smooth functions for generalized additive models. For example, Hastie and Tibshirani (1990) developed a back-fitting algorithm to estimate GAM and Wood (2006) used spline functions and the iteratively re-weighted least squares algorithm. Both algorithms are available in R, which is a widely-used open statistical software. However, the major problem in our estimation is about how to generate a training data set for each GAM. This section first briefly reviews the method in Wood (2006) to estimate GAM. Then, a detailed algorithm is given on how to generate the training data set for each model in the two-part generalized hierarchical model and how to incorporate it with the well developed GAM estimation algorithms. Lastly, the estimation of δ_l and δ_u using the cross-validation method is discussed.

According to Wood (2006), each smooth function in GAM can be represented by a sum of basis functions b_i. It can be described as:

$$f_j(x) = \sum_{i=1}^{B} \beta_i \cdot b_i(x) \tag{6.13}$$

A popular choice of basis functions is the cubic regression spline, which includes:

$$b_1(x) = 1 \tag{6.14}$$

$$b_2(x) = x \tag{6.15}$$

$$b_{i+2}(x) = R(x, x_i^*) \tag{6.16}$$

where

(1) $\{x_i^* | i \in \{1, \cdots, B-2\}\}$ are break points of the spline;

(2) $R(x, z)$ is defined as:

$$R(x,z) = \frac{[(z-\frac{1}{2})^2 - \frac{1}{12}][(x-\frac{1}{2})^2 - \frac{1}{12}]}{4} - \frac{(|x-z| - \frac{1}{2})^4 - \frac{1}{2}(|x-z| - \frac{1}{2})^2 + \frac{7}{240}}{24} \tag{6.17}$$

With the above transformation, a GAM is converted to a generalized linear regression model (GLM). GLM can be estimated efficiently by maximum likelihood estimation using the iteratively re-weighted least squares method (IRLS). Penalizing the likelihood with the complexity of the model, which is called the penalized iteratively re-weighted least squares method (P-IRLS) is usually adopted to overcome the overfitting problem. This algorithm is implemented with the package "mgcv" in R by Wood (2007).

To use the above algorithm to estimate GAM, training data are needed. Given the original data set $D = \{(y_i, \mathbf{x}_i)\}$, training data and estimate smooth functions for both GAMs are estimated with Algorithm 4 below. In this algorithm, it is assumed δ_l and δ_u are given.

Algorithm 4 estimates both GAMs with the original data set D. The R code of this algorithm is supplied with this dissertation as supplementary materials. It is assumed that δ_l and δ_u are known in the above algorithm. Practically, different combinations of values of δ_l and δ_u have been tried, and the cross-validation method is chosen to estimate the mean absolute errors because it is an effective and widely used estimation method available (Golub et al., 1979). Then, the parameters with the lowest mean absolute error are chosen as the optimal δ_l and δ_u, denoted as δ_l^* and δ_u^*.

6.4 Evaluation

A two-part generalized hierarchical model based on the generalized additive model is presented for semi-continuous data in this dissertation. It is especially developed for problems with a semi-continous response variable skew from the normal distribution, and the relationship between the response variable and predictors are nonlinear and complex. The purpose of modeling such data is to capture the causal mechanisms between variables and to predict future event. Our two-part generalized hierarchical model separate the modeling process into two parts, one of which is a classification procedure and the other one is a regression procedure. As discussed in Min and Agresti (2002), the two-part model is preferable to models with transformation because it addresses the data in their original form. Such form of modeling is simple

Algorithm 4 Algorithm 1: Estimation of Two-Part Generalized Hierarchical Model

```
Select subset $D_{classify} = \{(y_i, \mathbf{x}_i) | y_i \in \{C_1, C_2\}, (y_i, \mathbf{x}_i) \in D\}$;
Estimate $f_{classify}$ with $D_{classify}$ using the GAM estimation algorithm with $y_i$ as the binary response variable;
Apply $f_{classify}$ to predict label probability for $D$; the output is the probability set $Pr(D) = \{Pr_i\}$
Label each observation in $D$ by the following rules:
if $Pr_i < \delta_l$ then
  $Label_i = Label_{C_1}$
else
  if $Pr_i > \delta_l$ then
    $Label_i = Label_{C_2}$
  else
    $Label_i = Label_c$
  end if
end if
Select subset $D_{regression} = \{(y_i, \mathbf{x}_i) | Label_i = Label_c, (y_i, \mathbf{x}_i) \in D\}$;
Estimate $f_{regression}$ with $D_{regression}$ using GAM estimation algorithm with $y_i$ as the continuous response variable;
```

to fit and to interpret. In this dissertation, the generalized additive model is utilized as the base model for each part, taking the advantage of the interpretability of GAM because the relationships between variables are complex.

This section focuses on the performance evaluation of the developed two-part generalized hierarchical model for cleaned semi-continous data. The model is evaluated by being compared its predicted mean absolute error (PMAE) with that of several other classic models on the same data set. These classic models include linear regression model, generalized additive model, support vector machine, random forest, multivariate additive regression splines and boosted generalized linear and additive models. These models are estimated in R and details of used estimation packages are listed in Table 7.4 in Section 7.5.3. The two-part generalized hierarchical model $f_{classify}$ and $f_{regression}$ are fitted using GAM with the "mgcv" package in R. For $f_{regression}$, we used a logit link function and a probit link function to constrain the response between 0 and 1, and compared those with the identity link function. In order to evaluate the prediction performance of this developed model, the mean of predicted mean absolute error (PMAE) has been utilized as a comparison criteria with 100 trials of the test set method. From the results, it can be concluded that the two-part generalized hierarchical models with the three link functions outperform all other tested models in terms of the mean PMAE. For the different link functions, the logit and the probit link function are able to constrain the response to (C_1, C_2), so we prefer these two functions to the identity link function, and the 2P-GHM with the logit link function gives a slightly better prediction result than the one with the probit link function. In addition, since the generalized additive model is chosen as the base model in both parts, the two-part generalized hierarchical model has a good interpretability performance as well. Nonlinear impact patterns from each feature have been captured and displayed to the users clearly. Detailed results of the two-part generalized hierarchical model with three types of link functions, along with 8 classic models on the cleaned intergranular corrosion data are shown in Table 7.5 and Figure 7.22 in Section 7.5.3.

In summary, we present a two-part generalized hierarchical model based on the generalized additive model for semi-continuous data, which is skewed from the normal distribution. The two-part structure does not require a data transformation, so that the model is constructed on the original data form. We use a link function to constrain the response variable between two extreme values, and compare the performance of three different link functions. We implement the generalized additive model to the presented model, by taking advantage of its good interpretability as well as its ability to capture complex patterns within the data. Evaluation is provided using the intergranular corrosion data and the result shows that the presented model outperforms eight other classic regression models, in terms of the mean PMAE.

CHAPTER 7

METHODOLOGY APPLICATION

This chapter describes the application procedures of the developed formal methodology by individual models on the intergranular corrosion prediction problem. The background of the intergranular corrosion problem is introduced first, following with a formal problem definition. Then, the experiment procedures are stated in detail from sample preparation to image collecting. After describing the modeling data carefully, the three models are estimated and evaluated with the corrosion data in consequence, and the final results are given in terms of two criteria. This formal methodology is also evaluated by a comparison with several other classical supervised learning methods. The comparison results show that the developed formal methodology outperforms others significantly.

7.1 Background of the Intergranular Corrosion Problem

This application is part of a program to develop quantitative models that predict corrosion damage evolution from Intergranular Corrosion (IGC) to Intergranular Stress Corrosion Cracking (IGSCC) in 5XXX-series alloys (aluminum alloys with magnesium as the main alloying element). IGC and IGSCC are well-known failure modes experienced by 5XXX-series alloys. Aluminum (Al) alloys are widely used in marine, aerospace, automobile and manufacture industries because they have many advantages such as low density, high strength-to-weight ratio and tunable strength (Mondolfo, 1976). However, corrosion of aluminum alloys also limits their usage in these areas, especially for marine structures due to the corrosive environment. During service, these materials can be sensitized due to the precipitation of β (Al_3Mg_2) phase along grain boundaries. Various etching solutions show that some grain boundaries are preferentially attacked relative to others, indicating preferential precipitation of β. It has been speculated by Chan et al. (2008) that variations in the nature of grain boundaries would lead to different precipitation preferences. Numerous studies have been carried out to improve the corrosion resistance of such alloys (Kim et al., 2001; Lo et al., 2009; Unwin et al., 1969; Yuan, 2006) since their usage became popular.

AA5XXX-series aluminum alloys, which contain more than 3 wt% Magnesium (Mg), are used in the application of the formal methodology. These alloys are supersaturated with respect to Mg at ambient temperatures, leading to a large driving force for the formation of an intermetallic compound called β (Al_2Mg_3). The β usually preferentially precipitates along grain boundaries when exposed to temperatures higher than 70°C, according to Dix et al. (1958). The β corrodes rapidly in most solutions, including seawater, leading to materials that are susceptible to intergranular corrosion (IGC). Evaluation of the intergranular attack can be difficult. For AA5XXX-series alloys, the susceptibility to IGC is quantitatively measured by ASTM-G67 standard test, known as the Nitric Acid Mass Loss Test (NAMLT). Pan et al. (1996) and Unwin and Nicholson (1969) used transmission electron microscopy (TEM) to study the properties of grain boundary precipitation as a function of quench rate and treatment temperature. They concluded that the precipitation was strongly affected by local chemistry and atomic structures. The only grain boundary characteristic (GBC) parameter considered for correlation to IGC was grain misorientation angle (denoted as θ). Unwin and Nicholson (1969) found that there was a large amount of β along grain boundaries with θ in the range of 2° to 15°. Yuan (2006) had shown that in addition to misorientation angle, grain boundary plane also had an influence on IGC growth preferences. β phase preferentially grew on $\langle 111 \rangle$ with $\langle 112 \rangle$ and boundaries with θ

less than 20° were resistant to IGC. Kim et al. (2001) showed that grain boundaries with θ below 15° were hardly attacked by IGC in pure aluminum in 8%, 16% and 38% HCI solutions. Although many efforts and studies have been made to explore the correlations between GBCs and IGC in several alloy systems (Briant, 1980; Shimada et al., 2002; Tedmon Jr et al., 1971), the physical origins underlying relationships between GBCs and IGC are still unknown and the utilized analysis approaches are very limited (Lo et al., 2009). Some studies also found that low-angle grain boundaries and low-coincidence site lattice (CSL) boundaries have increased resistance to carbide precipitation for different alloys (Chan et al., 2008; Jakupi et al., 2010; Kim et al., 2001; Pan et al., 1996; Unwin et al., 1969; Yuan, 2006); no such analysis exists for IGC of AA5XXX-series alloys. In order to explain the underlying relationships between IGC and GBCs, and to predict future IGC based on given GBCs, a rigorous statistical model is needed.

7.2 Problem Definition

This application study begins with electrochemical experiments, of which the outputs are in two forms of images. One type of images is known as the optical microscope image and another type of known as the Electron Backscatter Diffraction (EBSD) image. Due to the limitation of the experiments, two images are distorted from each other, with noises distributed on them. Since images are the only data source for modeling, an image fusion model is necessary to extract information from them with a high accuracy. Therefore, the developed image fusion model is applied to both types of corrosion images for collecting enough relevant data for modeling use.

The noises on the images will be extracted via the image fusion model as well unavoidably. If these noises are added to the model along with other data, the prediction results that the model presents will be misleading, sometimes even against the truth. To avoid such mistakes happening, it is imperative to filter out noises from the data, and use clean data for modeling. The developed outlier detection framework is built based on such a need, and is applied to the corrosion data after being extracted from corrosion images with the image fusion model.

In this dissertation, electrochemical experiments have been done to simulate the real corrosion phenomenon, so the amount of β coverage along each grain boundary is quantitatively described by a percentage value. Each grain boundary could have one of three cases when being exposed to the etching solution: no corrosion, totally corroded and being corroded at some degree. Therefore, the variable which describes each grain boundary's degree of corrosion is a semi-continuous variable. It has many occurrences of single 0's and 1's showing no corrosion or totally corroded respectively, in addition to continuous positive values indicating the exact degrees of corrosion on grain boundaries. Since the distribution of such a variable is strongly skewed from the normal distribution, most classical regression models fail to provide a high prediction accuracy. In addition, the causality relationship between the degree of corrosion and a group of grain boundary characteristics is also demanded by material science engineers. Thus, the developed two-part generalized hierarchical model for such semi-continuous data is applied to this problem, aiming to provide material scientists and engineers with a high prediction accuracy of future intergranular corrosion events as well as good interpretability of causal mechanisms between various grain boundary characteristics and the IGC growth preferences.

7.3 Experimental Approach

7.3.1 Sample Preparation and Etching

AA5083-H131 alloy has been initially solution heat-treated and quenched (SHT/Q), and then sensitized at 100°C for a period from 3 days to 45 days. Then a range of samples have been cut and mounted. The ST surface of each sample is etched with solution of pH 1.2 ammonium persulfate ($(NH_4)_2S_2O_8$), because it is susceptible to the intergranular corrosion (IGC), due to the sensitization caused by β (Al_3Mg_2) which precipitates along grain boundaries. This solution is highly selective towards β dissolution (Allen, 2010). The Degree of Sensitization (DoS) is determined by the ASTM G67 Nitric Acid Mass Loss Test (NAMLT). In this application study, levels of DoS are obtained as follow: 2 mg/cm^2, 9 mg/cm^2, 24 mg/cm^2, 39 mg/cm^2, 49 mg/cm^2 and 57 mg/cm^2.

In preparation for electrochemical analysis and EBSD imaging, the mounted AA5083-H131 samples are firstly ground up to 1200 grit size with silicon carbide (SiC) paper in the presence of water. Samples are rinsed with water and dried with compressed air after each polishing step. Then samples are polished with 1.0 μm diamond suspension and finished with 0.02 μm silica for 5 min respectively. At last, samples are rinsed with alcohol and ion-etched for 15 min for the purpose of EBSD imaging.

Samples are exposed in ammonium persulfate solution which is of 200 ml water and 20 g ammonium persulfate ($(NH_4)_2S_2O_8$) (Allen, 2010). Different etching times from 0 min to 70 min by 10 min are tested on two samples with DoS = 49 mg/cm^2. Test images are shown in Figure 7.1 to Figure 7.4. It is found that the percent β coverage on grain boundaries for a sample plateaus by 60 min.

Therefore, etching time is chosen as 60 min for each sample at room temperature. In this way, the β phase is revealed by the precipitation, while the matrix of the sample is unattacked. Samples are rinsed with water after etching. Optical microscope images are taken after etching is completed.

7.3.2 Electron Backscatter Differaction Imaging

With the prepared samples, JEOL JSM-840 Scanning Electron Microscope is used to conduct the electron backscatter differaction imaging. The specimen is tilted 70°

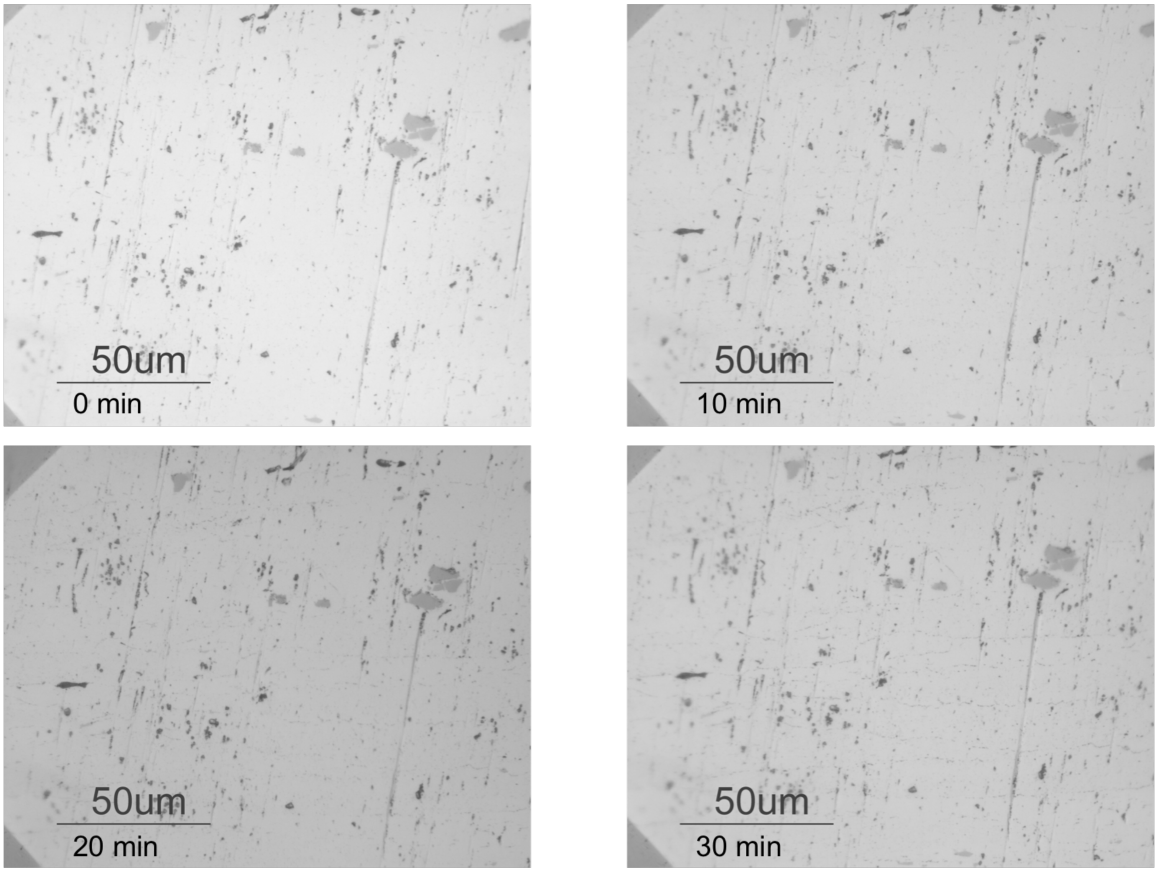


Figure 7.1: Ammonium persulfate etching comparison of different etching times on AA5083-H131 for sample 1: 0 min to 30 min. Magnification = 500X. DoS = 49 mg/cm^2.

to the horizontal in the chamber. The working distance for EBSD scanning is 21mm and the accelerated voltage is set to 20 KV. Magnification of images in this study is set to 200X. The filament current is adjusted to 200 A. The EBSD acquisition software named as *HKL CHANNEL 5 Flamenco* is utilized to obtain EBSD images and relevant grain boundary characteristics from the scanning results.

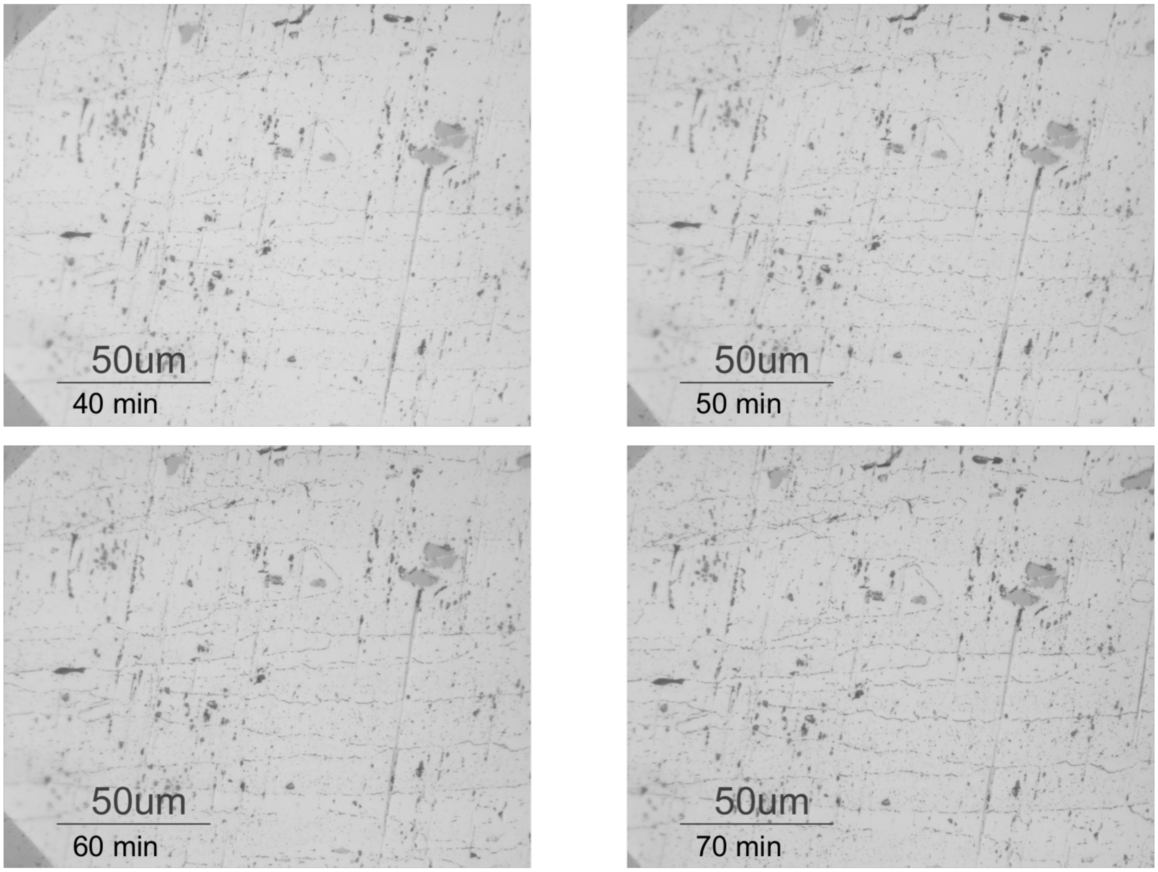


Figure 7.2: Ammonium persulfate etching comparison of different etching times on AA5083-H131 for sample 1: 40 min to 70 min. Magnification = 500X. DoS = 49 mg/cm^2.

7.4 Data Description

7.4.1 Intergranular Corrosion Quantitation

In this application study, the integrated effects of five grain boundary characteristics on the growth of intergranular corrosion have been examined using the developed formal methodology. The response variable in the data set is called *percent β coverage*. It is defined as the ratio of the corroded length to the entire grain boundary length, which is a quantitative way to describe the degree of intergranular corrosion along individual grain boundaries. Image processing algorithms are utilized to quantify this variable in this application study. If it is 0 for one grain boundary, then it

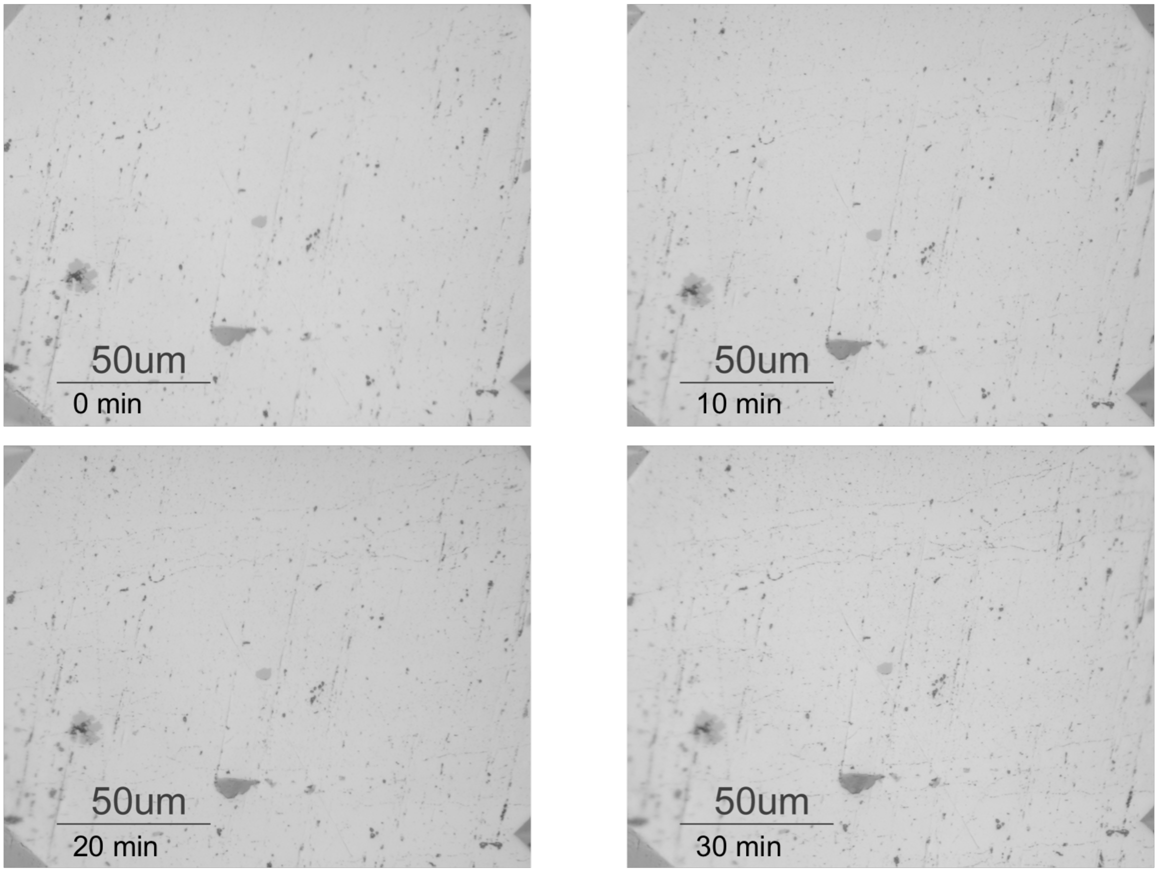


Figure 7.3: Ammonium persulfate etching comparison of different etching times on AA5083-H131 for sample 2: 0 min to 30 min. Magnification = 500X. DoS = 49 mg/cm^2.

shows no corrosion. Similarly, if it is equal to 1 for a grain boundary, then it means that this grain boundary is totally corroded. Any value between 0 and 1 indicates the specific degree of corrosion quantitatively. Therefore, according to Definition 1 in Chapter 6, the variable *percent β coverage* is a semi-continuous variable. In the data set for modeling, about 2000 grain boundaries are included and their distribution plot is shown in Figure 7.16(a).

7.4.2 Grain Boundary Characteristics

As predictors, five grain boundary characteristics are explained as follow. Grain boundary misorientation angle (θ) is defined as the orientation angle required to rotate

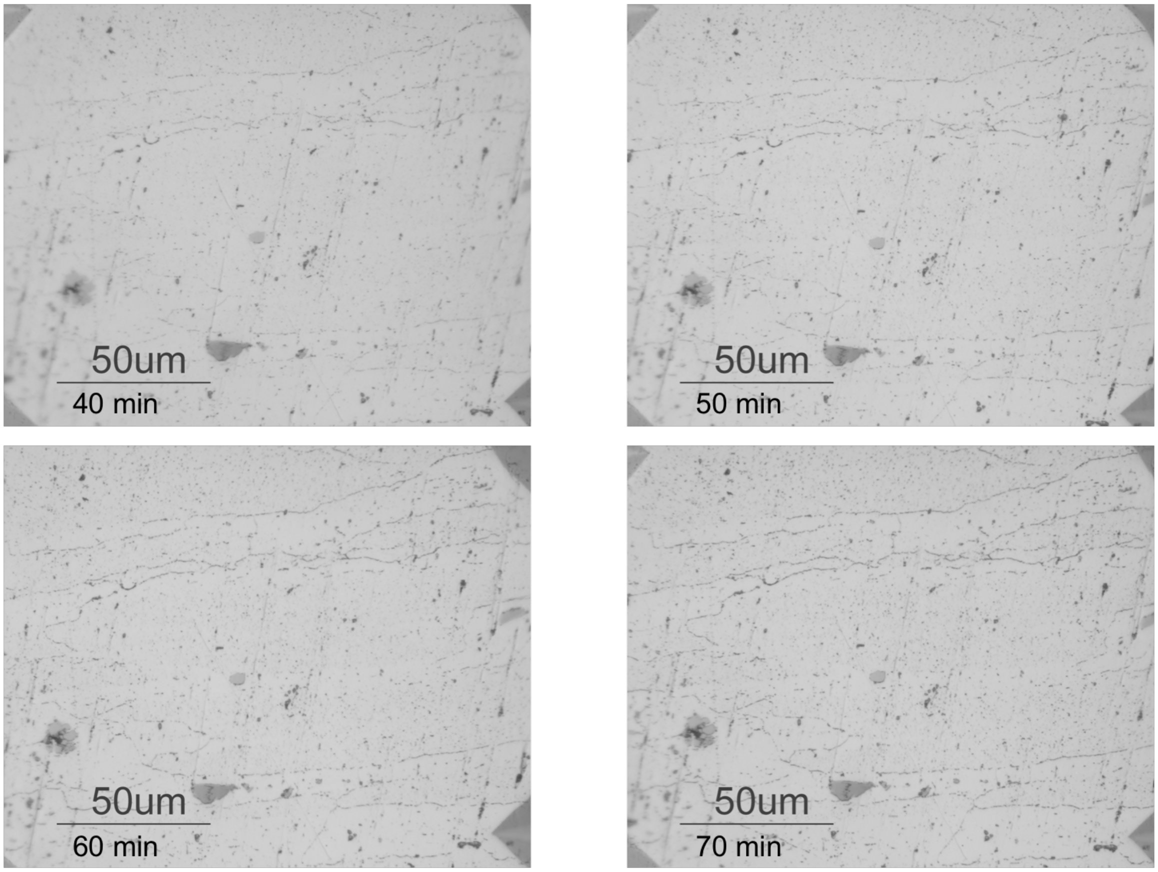


Figure 7.4: Ammonium persulfate etching comparison of different etching times on AA5083-H131 for sample 2: 40 min to 70 min. Magnification = 500X. DoS = 49 mg/cm^2.

the set of crystal axes of grain 1 into coincidence with grain 2 in its neighborhood. The demonstration of orientation angles is shown in Figure 7.7. Orientation difference angles are the differences of orientation angles between two neighboring grains grain 1 and grain 2. The illustration of orientation difference angle is shown in Figure 7.8. They are denoted as $\Delta\phi_1$,$\Delta\Phi$ and $\Delta\phi_2$, and

$$\Delta\phi_1 = |\phi_{11} - \phi_{21}| \tag{7.1}$$

$$\Delta\Phi = |\Phi_1 - \Phi_2| \tag{7.2}$$

$$\Delta\phi_2 = |\phi_{12} - \phi_{22}| \tag{7.3}$$

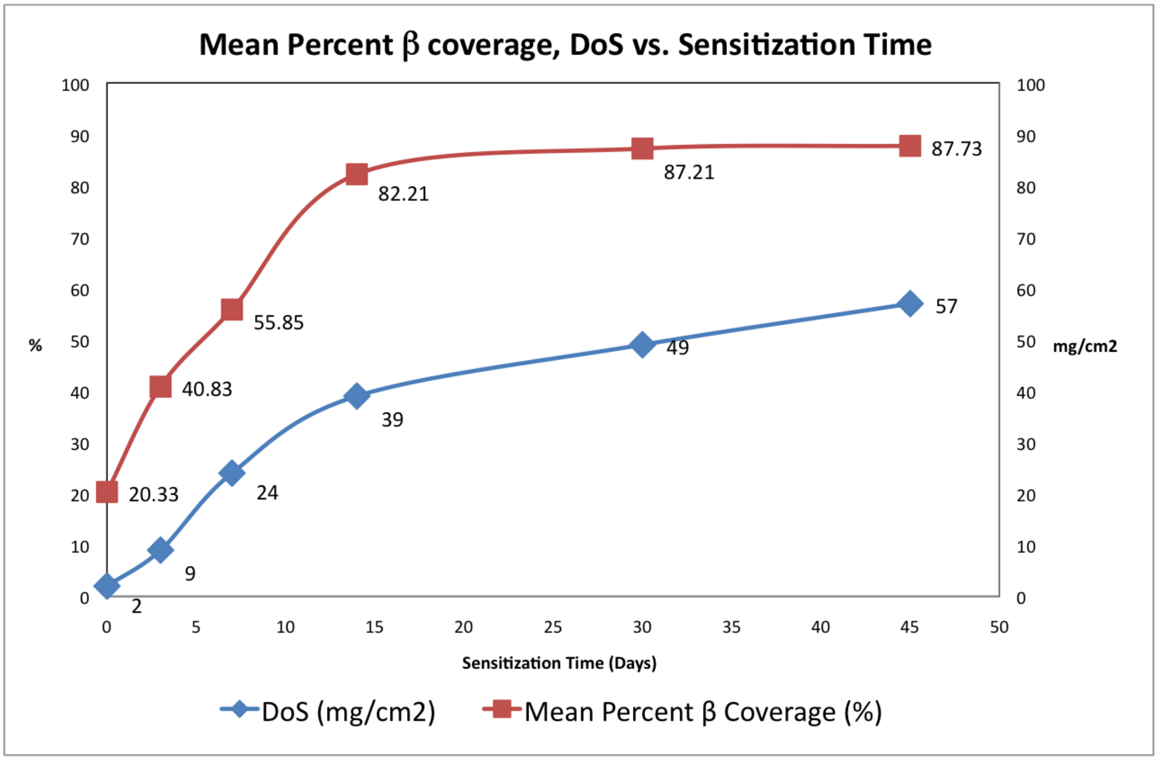


Figure 7.5: Mean percent β coverage, DoS vs. sensitization time

where $(\phi_{11}, \Phi_1, \phi_{12})$ and $(\phi_{21}, \Phi_2, \phi_{22})$ are the orientation angle sets on three directions of grain 1 and grain 2 in the Euler space respectively. An Electron Backscatter Diffraction (EBSD) system called *Channel5* from HKL Technology is used to obtain grain misorientation angles about $\langle 111 \rangle$ axis and orientation angles.

The fifth grain boundary characteristic considered is the grain boundary length l. Using a especially designed image processing algorithm, individual lengths of all grain boundaries within the sampled areas have been detected. In addition to these five grain boundary characteristics, *Degree of Sensitization* (DoS) is also included as a predictor for modeling, which is believed to have a significant impact on the growth of intergranular corrosion. In Figure 7.9, the bottom image is an EBSD image of the sample with DoS of 39mg/cm^2 and the top one is the microscope image of its corresponding corroded area on the same sample. Values of variables $\theta, \Delta\phi_1$,$\Delta\Phi$,$\Delta\phi_2$ and l are derived from the Channel5 Tango software suites (Day and Trimby, 2004).

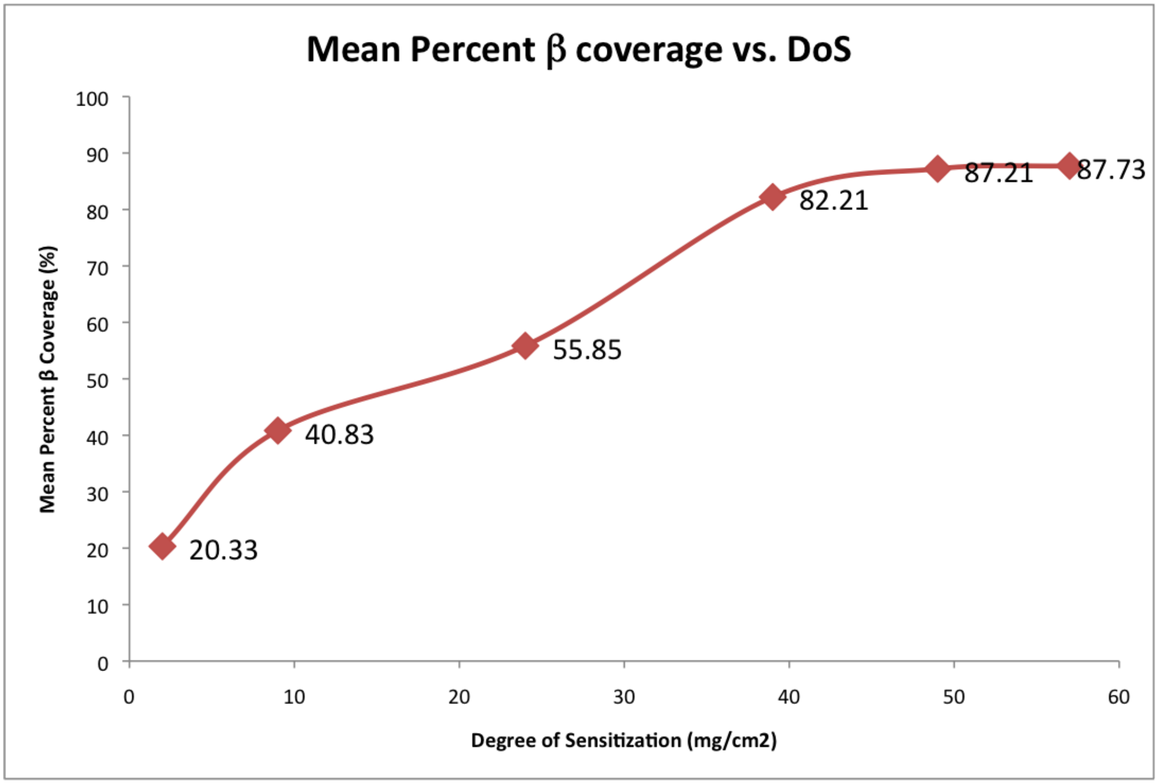


Figure 7.6: Mean Percent β coverage vs. DoS

7.5 Methodology Application

The application of the developed formal methodology to the intergranular corrosion problem is accomplished in three steps. The procedures are shown in Figure 7.10 with details. Starting with the image fusion model, relevant data are extracted from experiment images, and then noises within the data are filtered out by the outlier detection framework. At last, with the clean data, the two-part generalized hierarchical model for semi-continuous data is constructed to make predictions for future intergranular corrosion, and provide the causal mechanisms between different factors. Each section below describes one step of the application procedures.

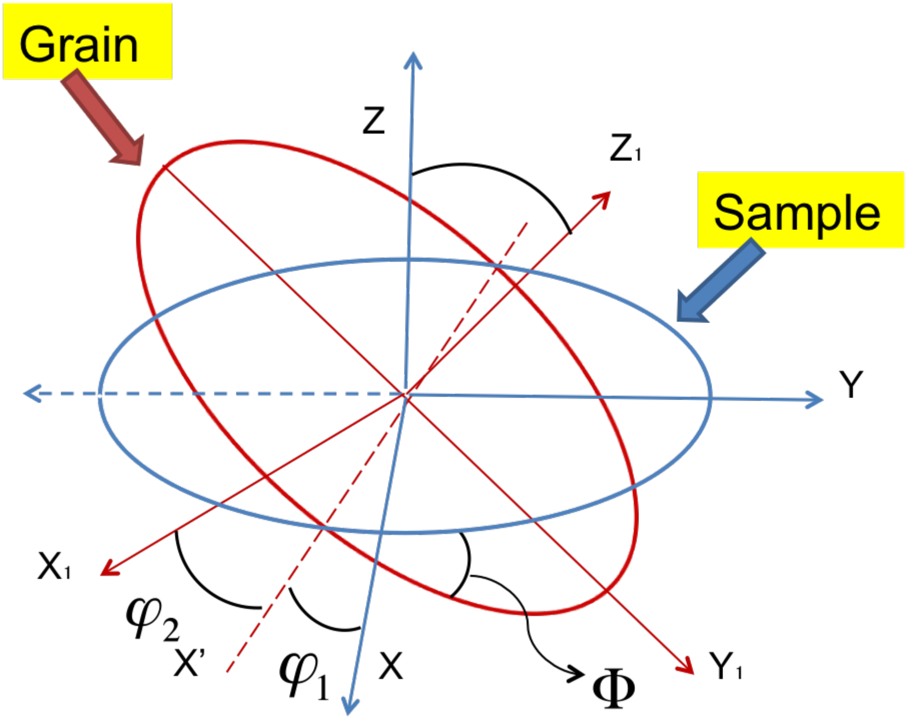


Figure 7.7: Grain Orientation (ϕ_1, Φ, ϕ_2) in the Euler Space

7.5.1 Image Fusion

Two types of images are collected from corrosion experiments, the EBSD image and the optical microscope image (as shown on the left panel of Figure 7.11). The EBSD images map the orientation of individual grains in crystalline materials (Dingley and Randle, 1992; Randle et al., 1992). They are acquired by the HKL Channel EBSD acquisition system when samples are tilted at about 70° in the chamber. The optical microscope images are taken by optical microscopes directly above samples. They present the intergranular corrosion on the material surface. Given a pair of such images, the application objective is to quantify the percentage of β coverage on each grain boundary, whose position is identified by the EBSD image.

About 2000 grain boundaries on six samples have been detected by the EBSD acquisition system. With the developed image registration algorithms, the optical microscope images are aligned with EBSD images of the same observed areas. Fig-

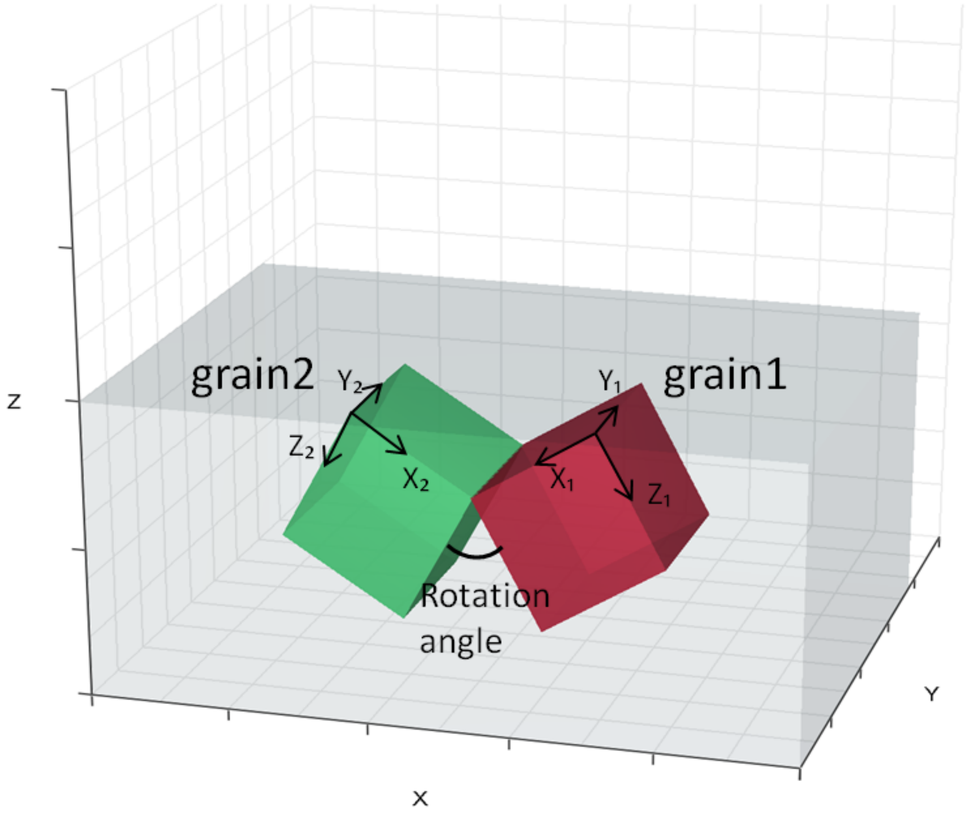


Figure 7.8: Demonstration of Orientation Difference Angles

ure 7.11 shows a pair of such images with a demonstration of the registration. As seen from Figure 7.11, two images are aligned pretty well.

The growth behavior of intergranular corrosion on each grain boundary is quantified by the percentage of the boundary length that is corroded after a 60min etching in ammonium persulfate solution with pH = 1.2. Figure 7.12 shows the illustration of percent β coverage calculation for individual grain boundaries.

To evaluate the performance of our image registration algorithms, they were tested on an EBSD image and a microscope image of the intergranular corrosion from the aluminum alloy sample of DoS (degree of sensitization) = 49 mg/cm^2. In order to measure the performance of the image fusion algorithms, intergranular corrosion was manually outlined on all grain boundaries (252 grain boundaries), and the average percentage of β coverage (E_{igc}) was calculated as $E_{igc} = 0.7693$. Our fusion algorithm gave an average result of mean percent β coverage on 252 grain boundaries as $E_p = 0.7273$ (sd = 0.0774). Using the widely used Control Point Registration method (Ingle and Proakis, 1999), the average value of mean percent β coverage of 252 grain boundaries this method output was $E_m = 0.1056$ (sd = 0.0543).

Also, the distributions of the percent β coverage calculated from the manually outlined image and the fused image were compared. Distribution plots are shown in Figure 4.2. Similar patterns are displayed on both of the distributions. Furthermore, the two-sample Kolmogorov-Smirnov test (K-S test) (Lilliefors, 1967) has been utilized to test whether the two datasets are significantly different from each other. The null hypothesis of the K-S test is that the two samples are drawn from the same distribution. The p-value of this test was 0.1108, which means we failed to reject the null hypothesis at the 5% significance level.

The evaluations show that our image fusion model performs better than the Control Point Registration, in terms of the accuracy of the data extraction, and our image registration algorithms are more suitable for image registration problems with images from the EBSD experiment and the optical microscope.

7.5.2 Outlier Detection Framework

Since the purpose of the ODF is to enhance the prediction accuracy of a regression model, the first comparison is between the ODF based on a linear regression model and three common robust regression models, robust linear model with M-estimation (RLM), least median of squares (LMS) and least trimmed of squares (LTS). The predicted mean absolute error (PMAE) is used as the criterion to compare their prediction performances. The intergranular corrosion data extracted from noisy images are used here for the comparison. With a test set evaluation method with 50 replicates, the ODF with a linear regression model has a mean PMAE of 0.249 with a confidence interval of (0.247, 0.252). The mean PMAE is 0.265 (0.263, 0.266) for RLM, 0.306 (0.293, 0.320) for LMS, 0.349 (0.334, 0.364) for LTS. These results show that the ODF with an ordinary least square regression outperforms the three major robust regression models in terms of PMAE, and it is effective to enhance the prediction accuracy of a regression model with the presence of outliers. Figure 7.13 shows that the comparison of the PMAE with the 95% confidence interval of all the tested models. All PMAE values with the 95% confidence intervals are shown in Table 7.1.

Models	PMAE (CI)
LM	0.274 (0.273, 0.276)
ODF+LM	**0.249 (0.247, 0.252)**
RLM	0.265 (0.263, 0.266)
LMS	0.306 (0.293, 0.320)
LTS	0.349 (0.334, 0.364)

Table 7.1: Mean PMAE values with 95% confidence intervals of ODF and major robust regression models

The second comparison is to test the effectiveness of the ODF by comparing the PMAE of each of four different supervised learning methods with and without the ODF. In this evaluation, four widely used supervised learning methods are applied. They are support vector machine with a radial basis (SVMR), support vector machine with a linear basis (SVML), random forest (RF) and the linear hierarchical model (MLM). The test is still taken on the intergranular corrosion data from noisy images. The regression results show that with ODF, the prediction performances of all four models are improved significantly because of a lower value of PMAE. Figure 7.14 shows that using the developed outlier detection framework improves the performance of all the tested models. All PMAE values with the 95% confidence intervals are shown in Table 7.2.

Furthermore, the Multiple Response Permutation Procedure (MRPP) in Mielke

Basis Model	PMAE (CI) without ODF	PMAE (CI) with ODF
SVMR	0.2455 (0.2429, 0.2481)	0.2359 (0.2348, 0.2370)
SVML	0.2511 (0.2482, 0.2540)	0.2456 (0.2443, 0.2469)
RF	0.2421 (0.2399, 0.2443)	0.2311 (0.2302, 0.2319)
MLM	0.2701 (0.2683, 0.2719)	0.2420 (0.2398, 0.2442)

Table 7.2: PMAE Comparison of Different Models for ODF Test

et al. (1976), Good and Wang (2005) and Park et al. (2009) are used to test whether there is a significant difference between the outlier set and the majority set. Test procedures are as follow: Generate a new data set by permuting each of the six predictors in the original data set while keeping each observation's outlier label unchanged. The outlier label is obtained from ODF. This new data set is fitted by different regression models and the regression results are shown below. Classification errors are calculated. Adaboosting is used to classify if an observation is an outlier or from the majority set on the training data. The classification error is the ratio of the sum of false positives and false negatives to the size of the training set. Figure 7.15 shows that after randomizing, the classification error increases, and the PMAEs of regression models are going up. This suggests that there exist clusters in the original data set. Outliers are grouped in the cluster that is different from other observations. The cluster that is including outliers are detected by the ODF. But the permutation disrupts the clustering patterns, which leads to a poor prediction performance for the randomized data. All PMAE values for this comparison are shown in Table 7.3.

7.5.3 Two-part Generalized Hierarchical Model for Semi-continuous Data

A range of sensitized AA5083-H131 specimens are cut, mounted and prepared using standard metallographical techniques. About 2000 grains have been detected from 6 samples with degree of sensitization equal to 2 mg/cm^2, 9 mg/cm^2, 24 mg/cm^2,

Regression Models	Data	PMAE	95% CI
SVMR	Original Data	0.2359	(0.2348, 0.2370)
	Randomized Data	0.4146	(0.4093, 0.4198)
SVML	Original Data	0.2456	(0.2443, 0.2469)
	Randomized Data	0.4349	(0.4290, 0.4408)
RF	Original Data	0.2311	(0.2302, 0.2319)
	Randomized Data	0.4031	(0.3985, 0.4078)
MLM	Original Data	0.2420	(0.2398, 0.2442)
	Randomized Data	0.4101	(0.4057, 0.4145)

Table 7.3: PMAE Comparison for Randomized Data and Original Data

39 mg/cm^2, 49 mg/cm^2 and 57 mg/cm^2, using the developed image fusion model. These extracted data are cleaned by the outlier detection framework for a better prediction performance of the regression modeling. The entire data set for modeling is collected using methods described in Section 7.4. Histograms of variables *percent β coverage*, *misorientation angle* (θ), *orientation angle differences* ($\Delta\phi_1, \Delta\Phi, \Delta\phi_2$) and *grain boundary length* (l) are shown in Figure 7.16(a) - Figure 7.18(b).

As seen in Figure 7.16(a), the distribution of *percent beta coverage* includes two single values 0 and 1 clustered on two ends, and a continuous distribution of values between them. This is a typical characteristic of semi-continuous data. Classic regression models usually fail to fit such data due to their skewness from the normal distribution. Therefore, the developed two-part generalized hierarchical model is applied to them, in order to obtain good prediction and interpretability performances. Distributions of other variables in the corrosion data set are also shown in Figure 7.16(b) - 7.18(b). Figure 7.19 shows the scatter plot matrices of all variables in the data set. As being observed, the relationship between *percent β coverage* and DoS seems to be positively linear, but relationships between other pairs do not fit a linear model quite well. As a result, a generalized hierarchical model, instead of a linear model, is inspired because of such observations.

This section focuses on the application of the two-part generalized hierarchical model for clean semi-continous data for good prediction and interpretability performance. It is evaluated by being compared its predicted mean absolute error (PMAE) with different link functions and with that of several other classic models using the same corrosion data. These classic models include linear regression model, generalized additive model, support vector machine, random forest, multivariate additive regression splines and boosted generalized linear and additive models. More details about these models are shown in Table 7.4. The results from using these models are shown in Table 7.5.

The 10-fold cross-validation technique is used to choose the optimal parameters δ_l and δ_u in our model: $\delta_l^* = 0.2$ and $\delta_u^* = 0.95$. Then two parts of the two-part generalized hierarchical model $f_{classify}$ and $f_{regression}$ are fitted using GAM with the "mgcv" package in R. The GAM plots of each predictor in model part I are shown in Figure 7.20, and GAM plots of model part II are shown in Figure 7.21. As observed in these two groups of figures, predictors $\Delta\phi_1$ and $\Delta\Phi$ are significant in model part I, and predictors θ and l are significant in model part II. These results have important meanings for material scientists and engineers because they are looking for physical causalities behind these variables. Such nonlinear impact patterns from each predictor on the response variable give them hints to consider the internal physical characters of such material.

In order to evaluate the prediction performance of this developed model, the mean of predicted mean absolute error (PMAE) has been utilized as a comparison criteria with the 100-trial of testset method. Results of the two-part generalized hierarchical model with three different link functions, along with 8 classic models using corrosion data are shown in Table 7.5. Figure 7.22 shows means of PMAE of all tested models with 95% confidence intervals. From the results, it can be concluded that the three two-part generalized hierarchical models outperform all other tested models in terms of the mean PMAE. For the different link functions, the logit and the probit link function are able to constrain the response to (C_1, C_2), so we prefer these two functions to the identity link function, and the 2P-GHM with the logit link function gives a slightly better prediction result than the one with the probit link function. In addition, since the generalized additive model is chosen as the base model in both parts, the two-part generalized hierarchical model has a good interpretability performance as well. Nonlinear impact patterns from each feature have been captured and displayed to the users clearly.

Model Name	**Explanation**	**Estimation Package in R**
LM	Linear regression model	*stats*
GAM	Generalized additive model estimated by REML; smooth terms are chosen by GCV	*mgcv*
SVMR	Support vector machine with a Gaussian radial basis function	*e1071*
SVML	Support vector machine with a linear kernel	*e1071*
RF	Random forest with 100 trees	*randomForest*
MARS	Multivariate additive regression splines	*earth*
GAMB	Boosted generalized additive model	*mboost*
GLMB	Boosted generalized linear model	*mboost*
2P-GHM-identity	Two-part generalized hierarchical model with an identity link function	Our algorithm with GAM estimated by *mgcv*
2P-GHM-logit	Two-part generalized hierarchical model with a logit link function	Our algorithm with GAM estimated by *mgcv*
2P-GHM-probit	Two-part generalized hierarchical model with a probit link function	Our algorithm with GAM estimated by *mgcv*

Table 7.4: Details of tested methods for model comparison

Model	Mean of PMAE	95% CI of mean PMAE
LM	0.237	(0.236, 0.238)
GAM	0.231	(0.230, 0.232)
SVMR	0.206	(0.205, 0.207)
SVML	0.222	(0.220, 0.223)
RF	0.215	(0.214, 0.216)
MARS	0.229	(0.227, 0.230)
GAMB	0.234	(0.231, 0.236)
GLMB	0.238	(0.237, 0.239)
2P-GHM-identity	0.198	(0.196, 0.199)
2P-GHM-logit	0.198	(0.196, 0.199)
2P-GHM-probit	0.199	(0.198, 0.201)

Table 7.5: Mean PMAE of 9 tested models for the corrosion example

7.6 Discussion

In grain boundary engineering, the Coincident Site Lattice (CSL) theory shows that the degree of fit (Σ) between the structures of the two neighboring grains can be represented by the ratio of coincidence sites to the total number of sites (Randle, 1996). Also, grain boundary energy is defined as the excess free energy associated with the presence of a grain boundary, with the perfect lattice as the reference point. Numerous studies and experiments have shown that there exists a strong correlation between the amount of grain boundary energy and the microstructure of some special grain boundaries, denoted as coincidence site lattices (Hasson et al., 1972; Kavner and Devine, 1997; Li et al., 2009; Tschopp and McDowell, 2007), which leads to the key of corrosion resistance solutions. For example, boundary with high Σ might be expected to have a higher energy than the one with low Σ.

This application study has a contribution to the exploration of the relationship between general high angle boundaries (between 30° and around 62°) and the intergranular corrosion. The result of GAM model II in the developed generalized hierarchical model shows that there exists a significant relationship between misorientation angle and *Percent β Coverage*, which is shown on the first subfigure in Figure 7.21. As observed, there are two peaks on that figure. One is at about 30° and the other one is at around 55°. This plot suggests that there exists a negative correlation between *Percent β Coverage* and misorientation from about 30° to 45°, and a positive correlation when misorientation is from 45° to 55°.

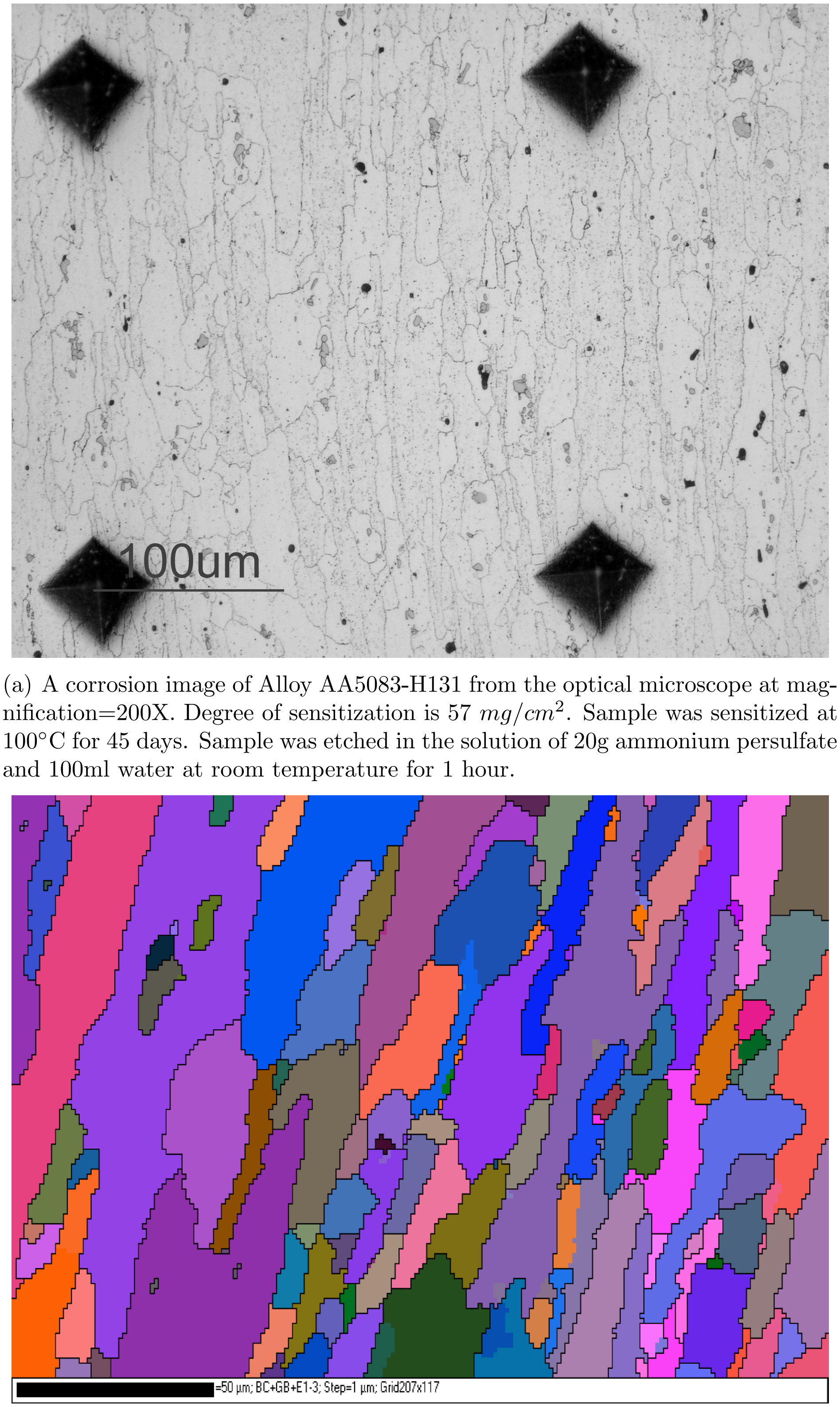
When comparing these conclusions with previous work on the grain boundary energy, it is found that the influential pattern of misorientation angle (between 15° and about 60°) with $\langle 111 \rangle$ axis on the growth of β phase performs similarly as that of the CSL on the grain boundary energy. Skidmore et al. (2004) have shown that grain boundary energy increases sharply from low angles, then gradually decreases from its highest point at 15° to 60°, with some high grain boundary energy values corresponding to angles around 30°. Their observation is partially consistent with that of this work. Since Skidmore et al. (2004) only used about 50 data points on the trend plot, while there are over 2000 data points in this work, the conclusion from this work is better supported by the larger data size. To my best knowledge, only few literatures have considered the $\langle 111 \rangle$ misorientation axis, which has been studied in this dissertation, and they did not provide a clear trend plot about $\langle 111 \rangle$ for small angles below 20°.

In the future, low angle boundaries with misorientation angles from 0° to 15° should be included in the data set as well. The Read-Shockley equation shows that low angle grain boundary energy steadily increases with misorientation angles from 0° to 15° (Read and Shockley, 1950). With the developed image fusion model, a large number of low angle boundaries are available for modeling. Therefore, it is possible to test if the relationship between low angle boundaries and *Percent β Coverage* is consistent with that between low angle boundaries and grain boundary energy, according to Read and Shockley (1950). According to Kim et al. (2006), two more grain boundary parameters should be considered when defining a grain boundary. These two parameters are utilized to describe the orientation of a grain boundary plane.

7.7 Summary

In summary, the semi-continuous data is introduced at beginning of this chapter and a two-part generalized hierarchical model is built for these data. Then the developed model is applied to the intergranular corrosion prediction problem because the response variable in this problem is a semi-continuous variable. At last, in order to evaluate the model performance, the developed model is compared with 8 other classic regression models in terms of the predicted mean absolute error and model interpretability. The developed model outperforms all its competitors and has a good interpretability performance as well.

Material scientists and engineers would benefit from the modeling results which provide clear nonlinear impact patterns from grain boundary characteristics on the intergranular corrosion growth. For example, it is known from the results that grain boundary characteristics $\Delta\phi1$ and $\Delta\Phi$ are significant in model part I, which means they mainly affect the β coverage in the classification of 0 or 1. Grain boundary length l and misorientation angle θ are significant variables affecting the percent β coverage within the range of 0 to 1, which might lead to the importance of grain shapes in determining how intergranular corrosion growth preferentially. Future work might focus on extending this modeling framework by implementing it with more complicated base models, and looking for more efficient and reliable model estimation methods for a better prediction performance.



(a) A corrosion image of Alloy AA5083-H131 from the optical microscope at magnification=200X. Degree of sensitization is 57 mg/cm^2. Sample was sensitized at 100°C for 45 days. Sample was etched in the solution of 20g ammonium persulfate and 100ml water at room temperature for 1 hour.

(b) A microstructure image from Electron Backscatter Diffraction system at magnification=200X. Degree of sensitization is 24 mg/cm^2. Sample was sensitized at 100°C for 7 days.

Figure 7.9: Examples of Images with Grain Boundary Characteristics

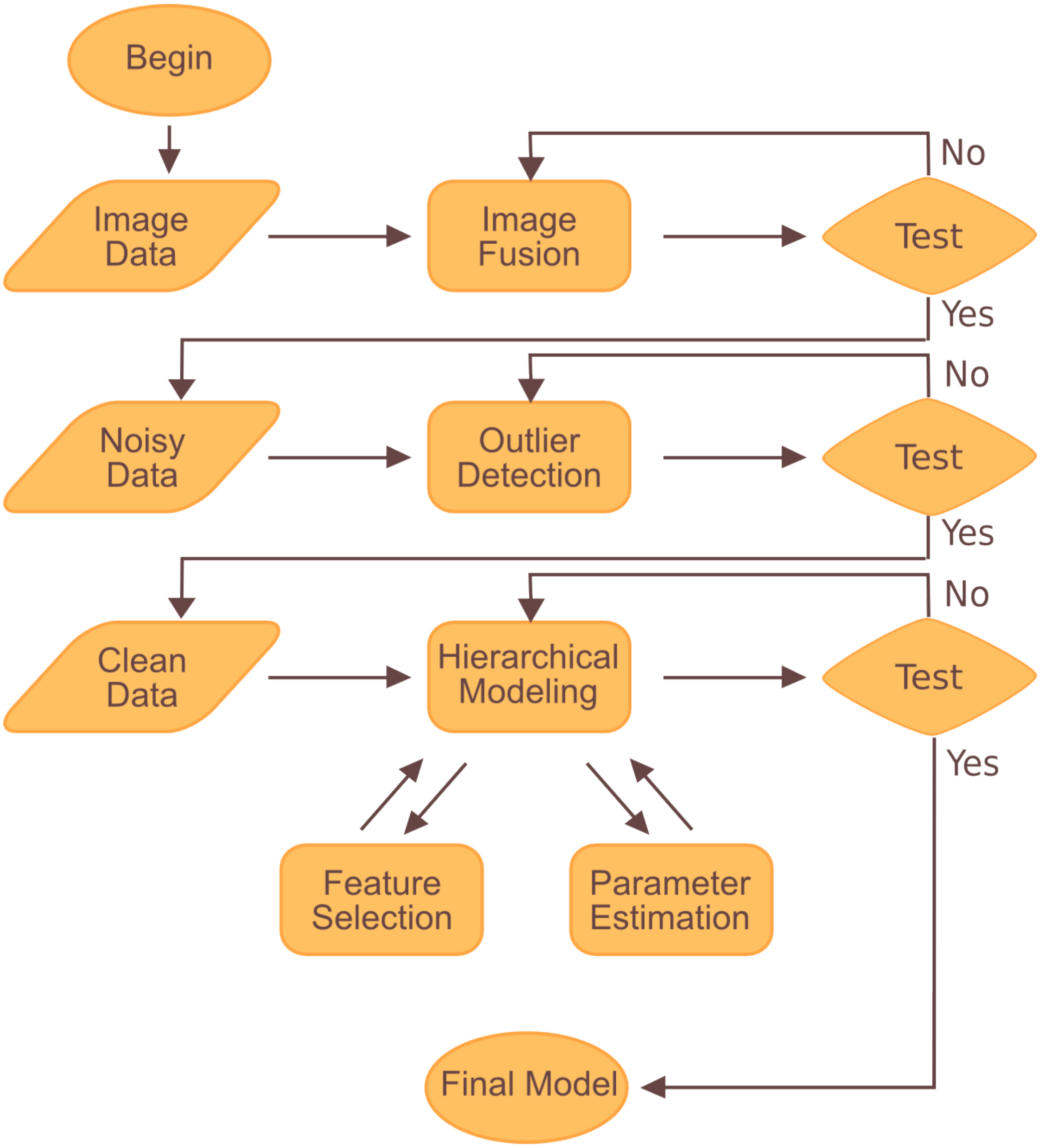


Figure 7.10: Flowchart of the methodology application: an intergranular corrosion example

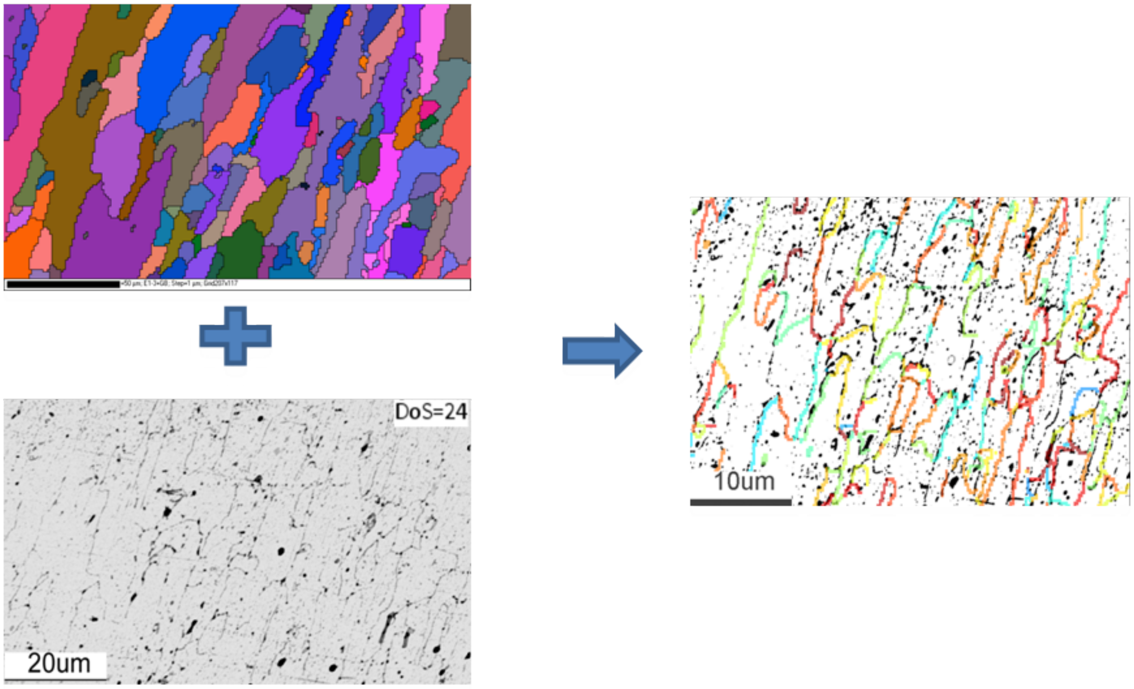
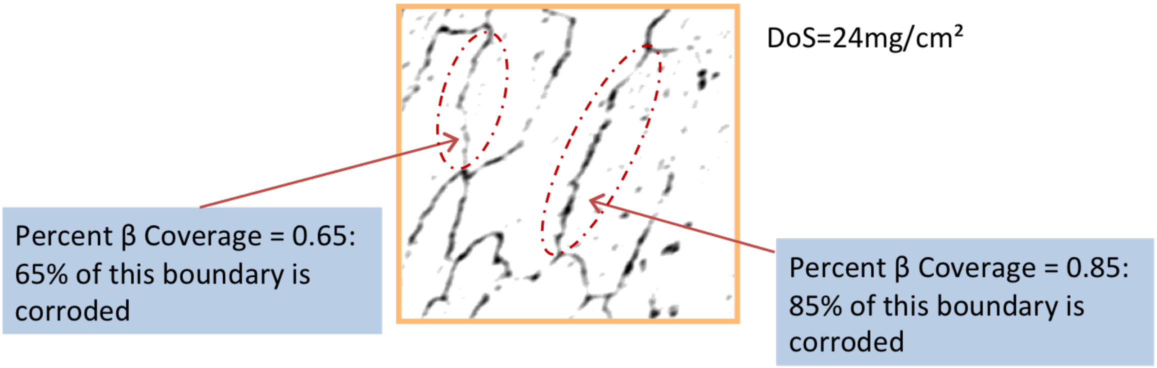


Figure 7.11: Image registration demonstration. Left: an EBSD image and an optical microscope image; Right: a registered image with colored boundaries for different orientations


DoS=24mg/cm²
Percent β Coverage = 0.65: 65% of this boundary is corroded
Percent β Coverage = 0.85: 85% of this boundary is corroded

$$\textbf{Percent } \beta \textbf{ coverage} = \frac{\textbf{length of corroded boundary}}{\textbf{length of entire boundary}}$$

where length is measured by pixels on digital images.

Figure 7.12: Degree of IGC calculation

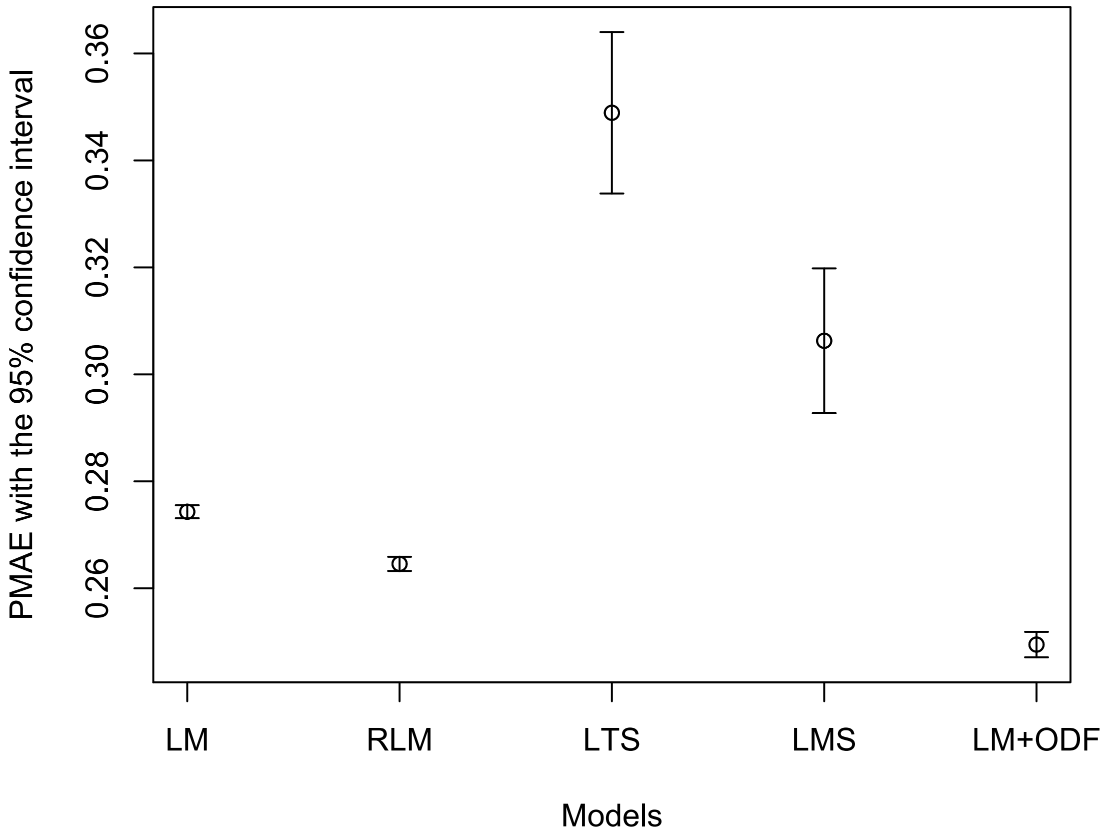


Figure 7.13: Mean PMAE comparison between ODF and major robust regression models

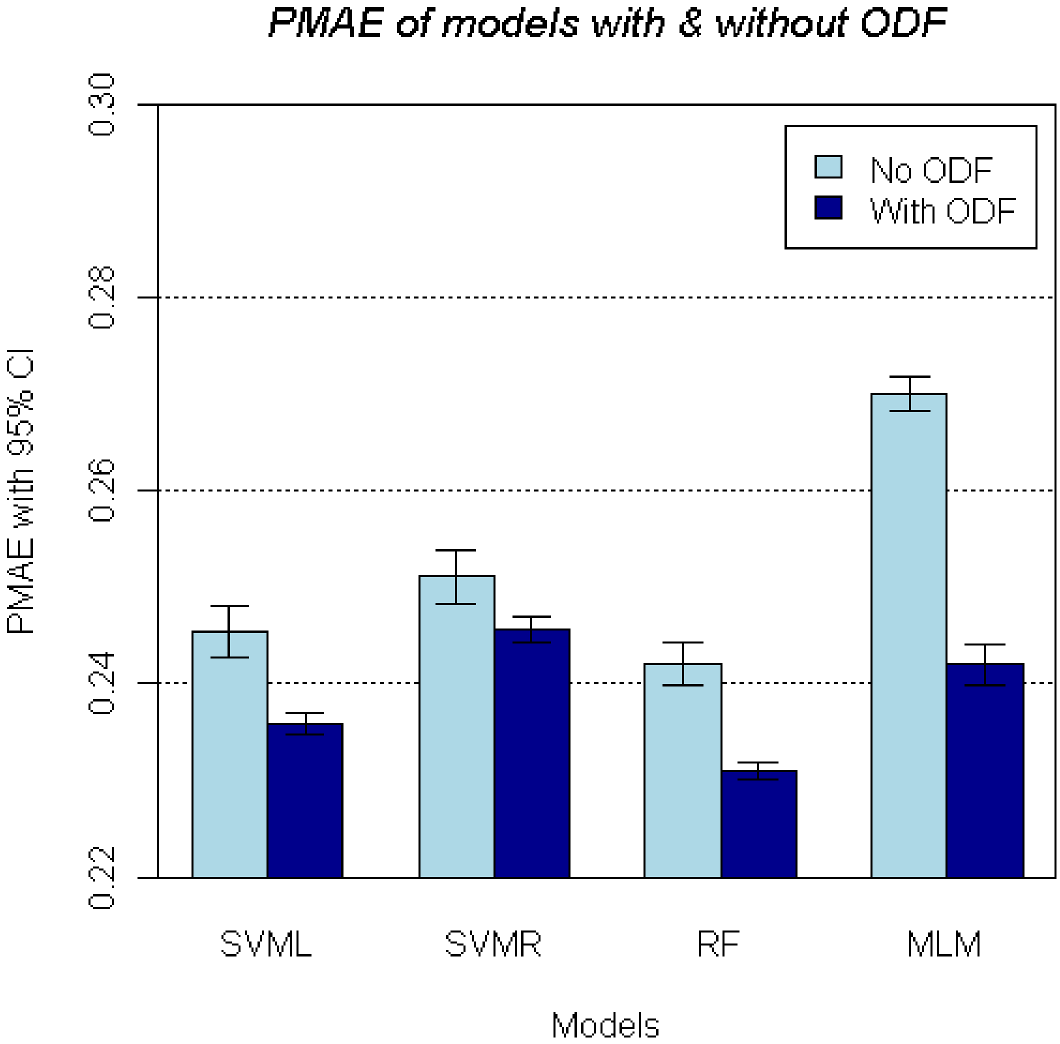


Figure 7.14: PMAE comparison of different models for ODF application

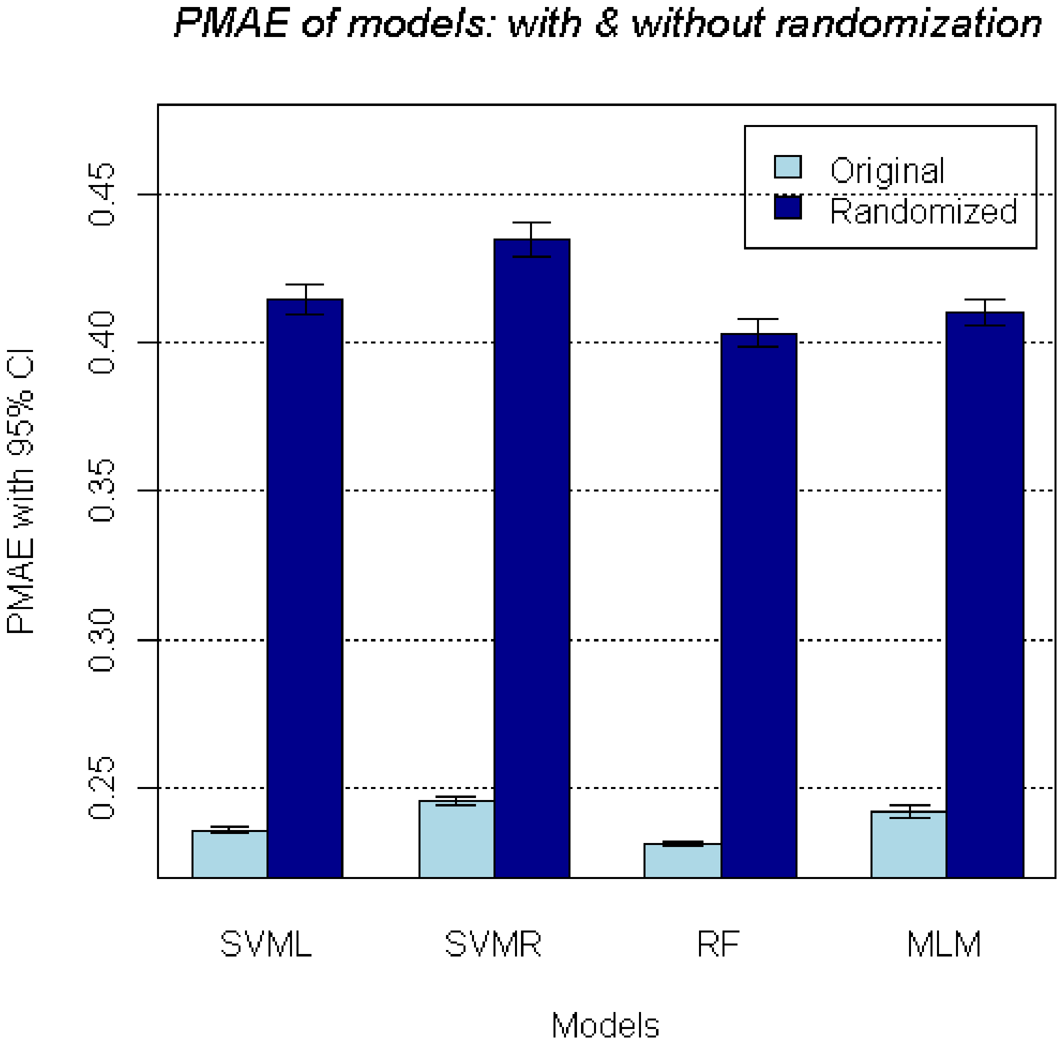
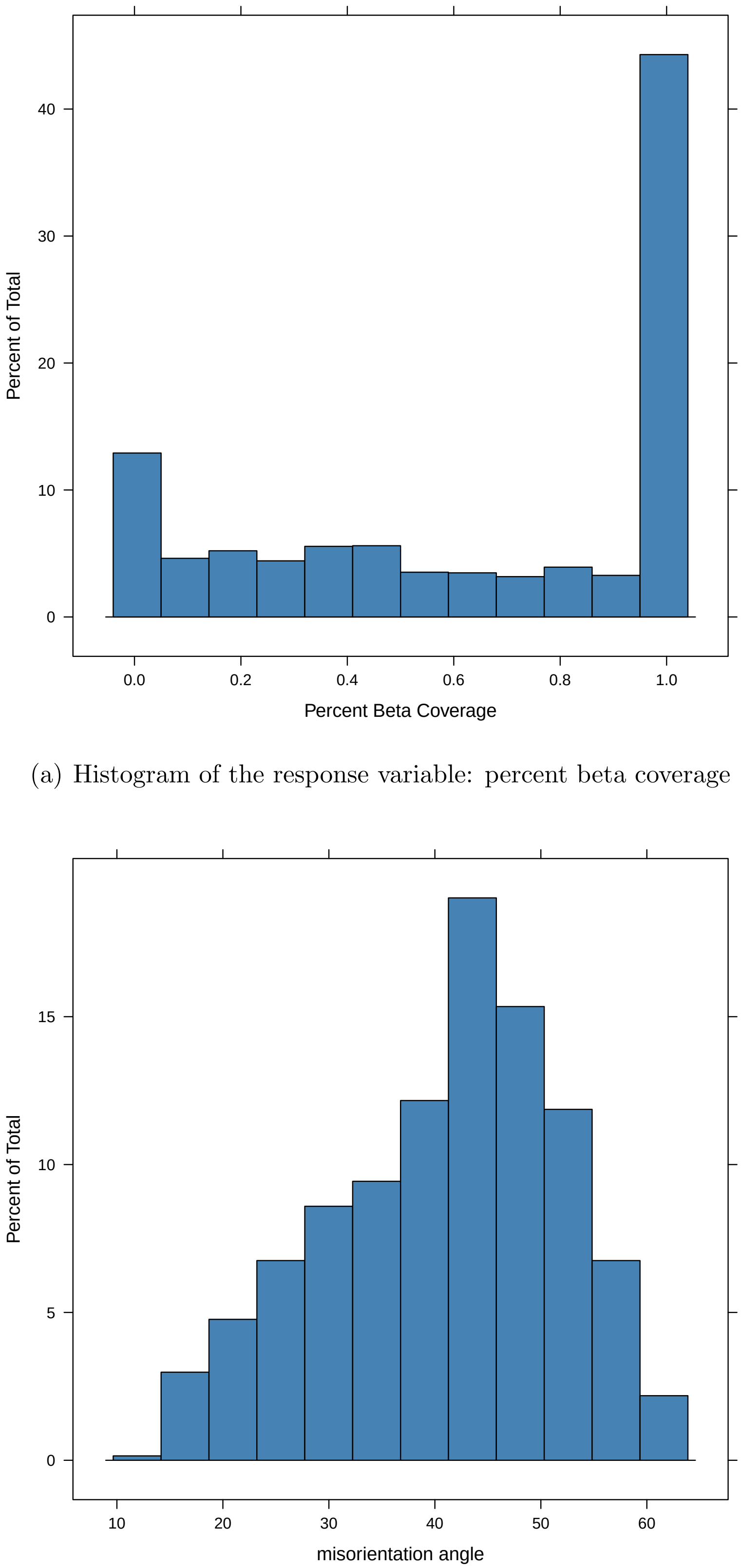


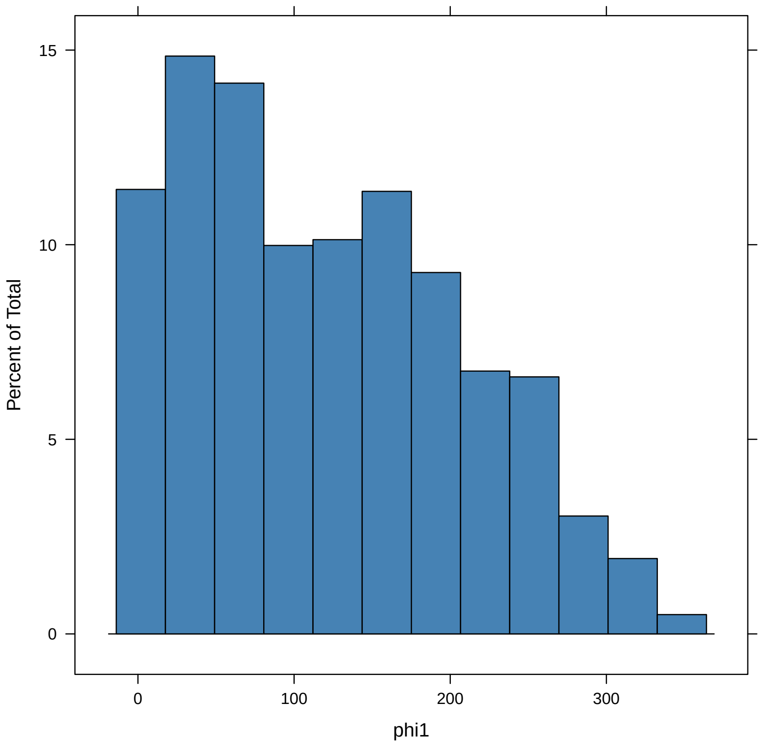
Figure 7.15: PMAE comparison for randomized data and original data



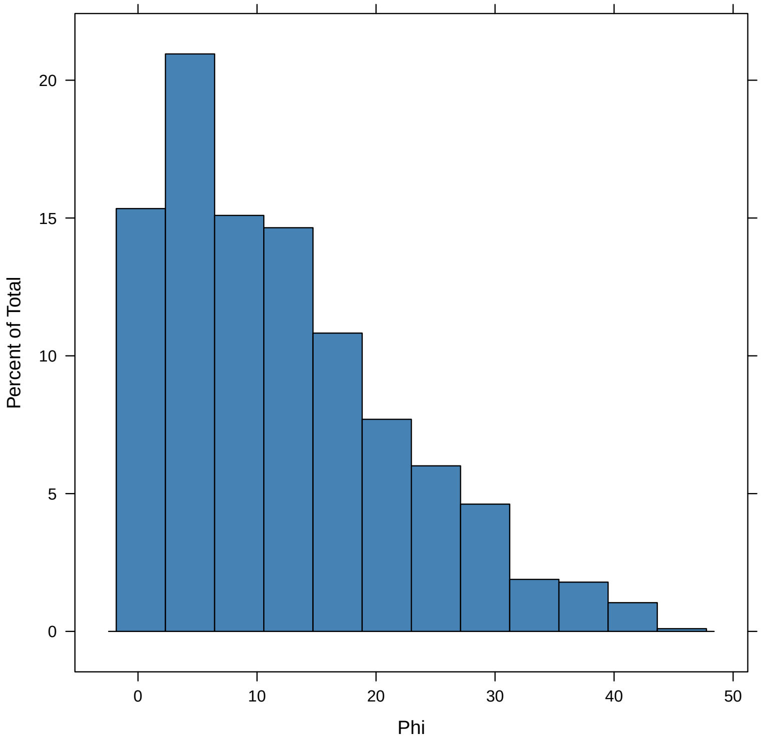
(a) Histogram of the response variable: percent beta coverage

(b) Histogram of the predictor: misorientation angle θ

Figure 7.16: Histograms of variables in the corrosion example

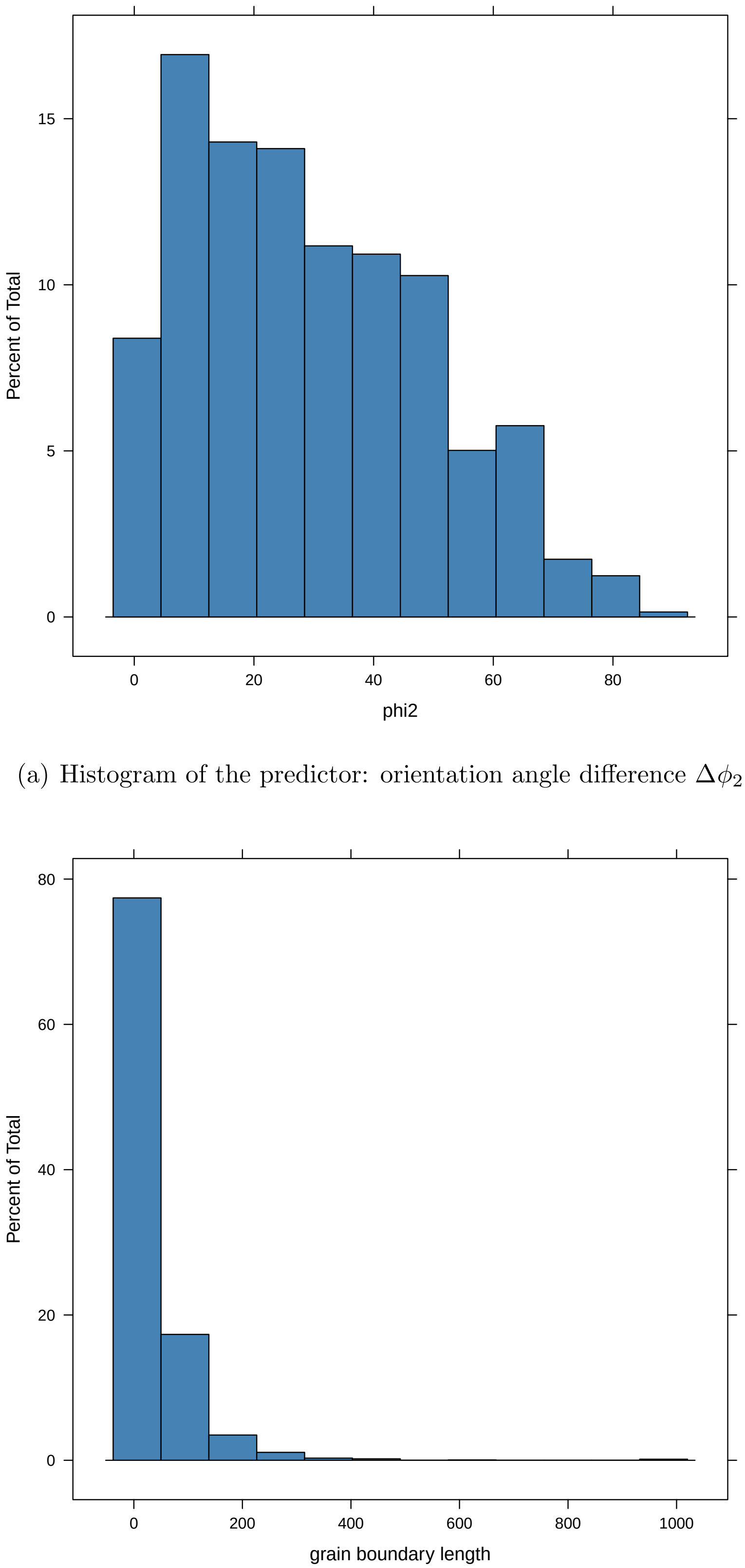


(a) Histogram of the predictor: orientation angle difference $\Delta\phi_1$



(b) Histogram of the predictor: orientation angle difference $\Delta\Phi$

Figure 7.17: Histograms of variables in the corrosion example - continued



(a) Histogram of the predictor: orientation angle difference $\Delta\phi_2$

(b) Histogram of the predictor: grain boundary length

Figure 7.18: Histograms of variables in the corrosion example - continued

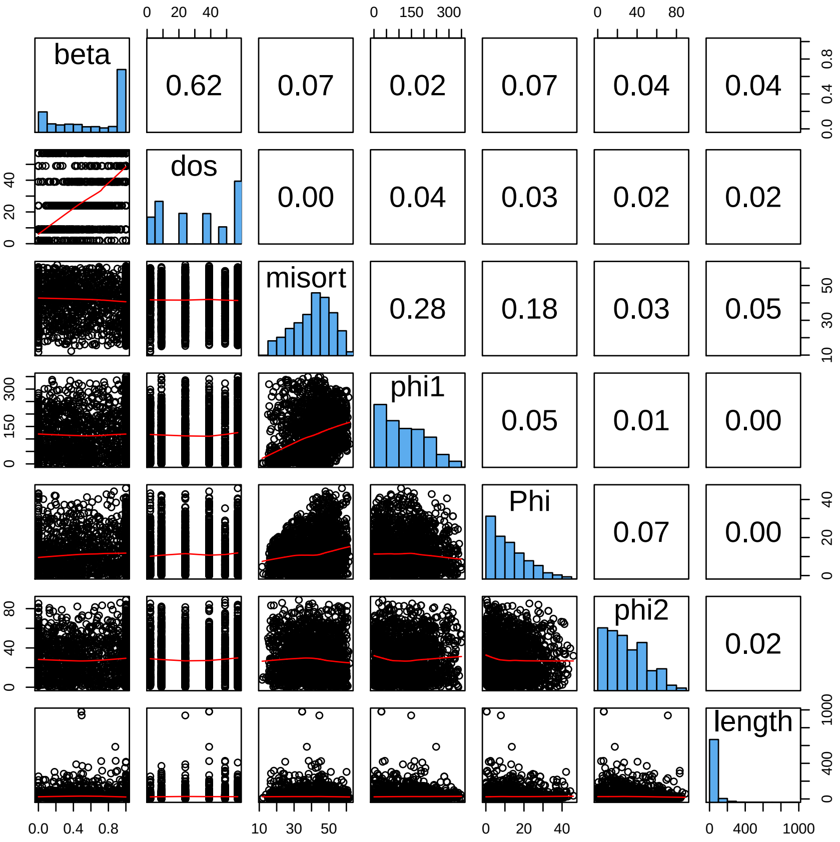


Figure 7.19: Scatter plot matrices of all variables in the corrosion data set

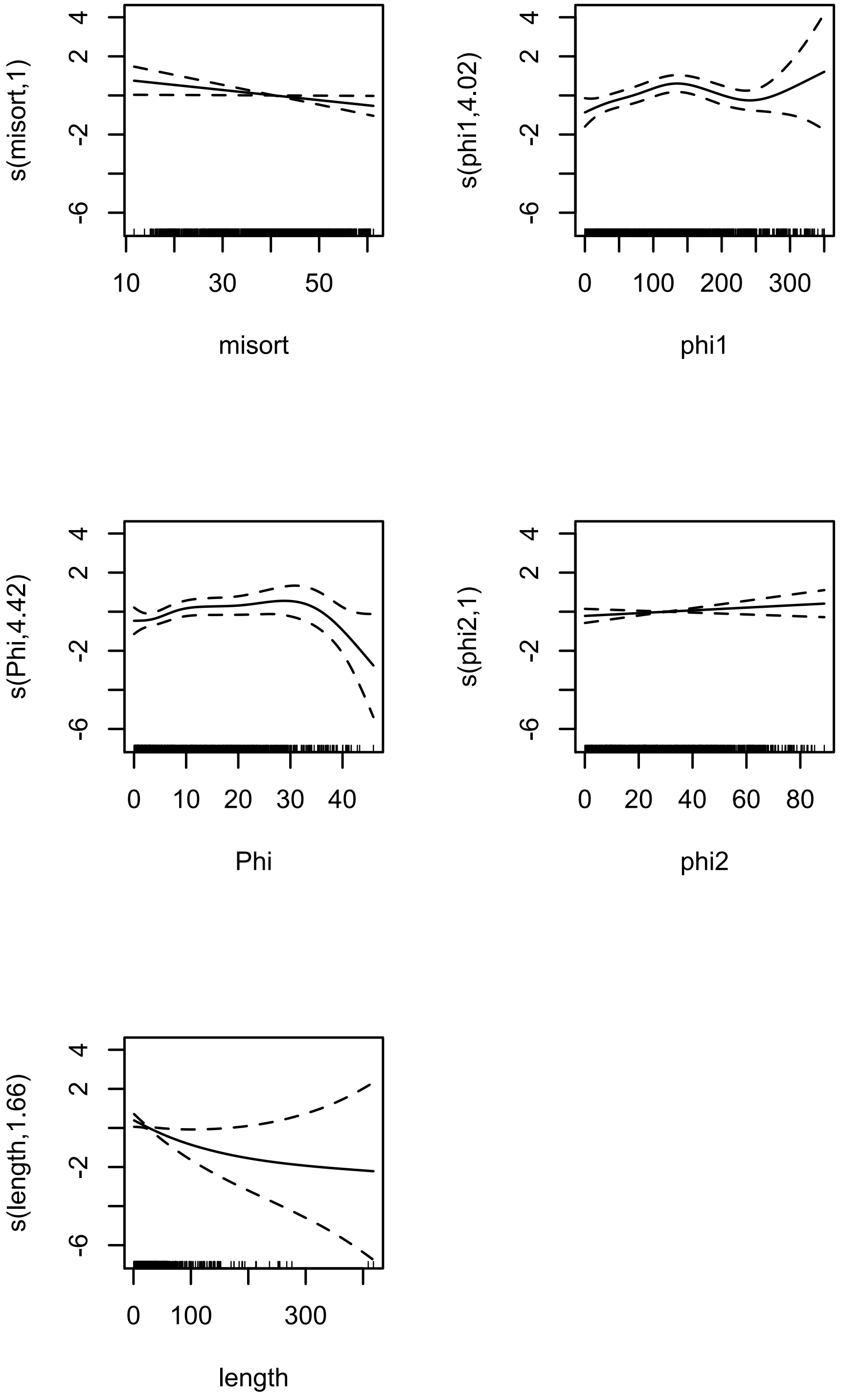


Figure 7.20: GAM plots of model part I

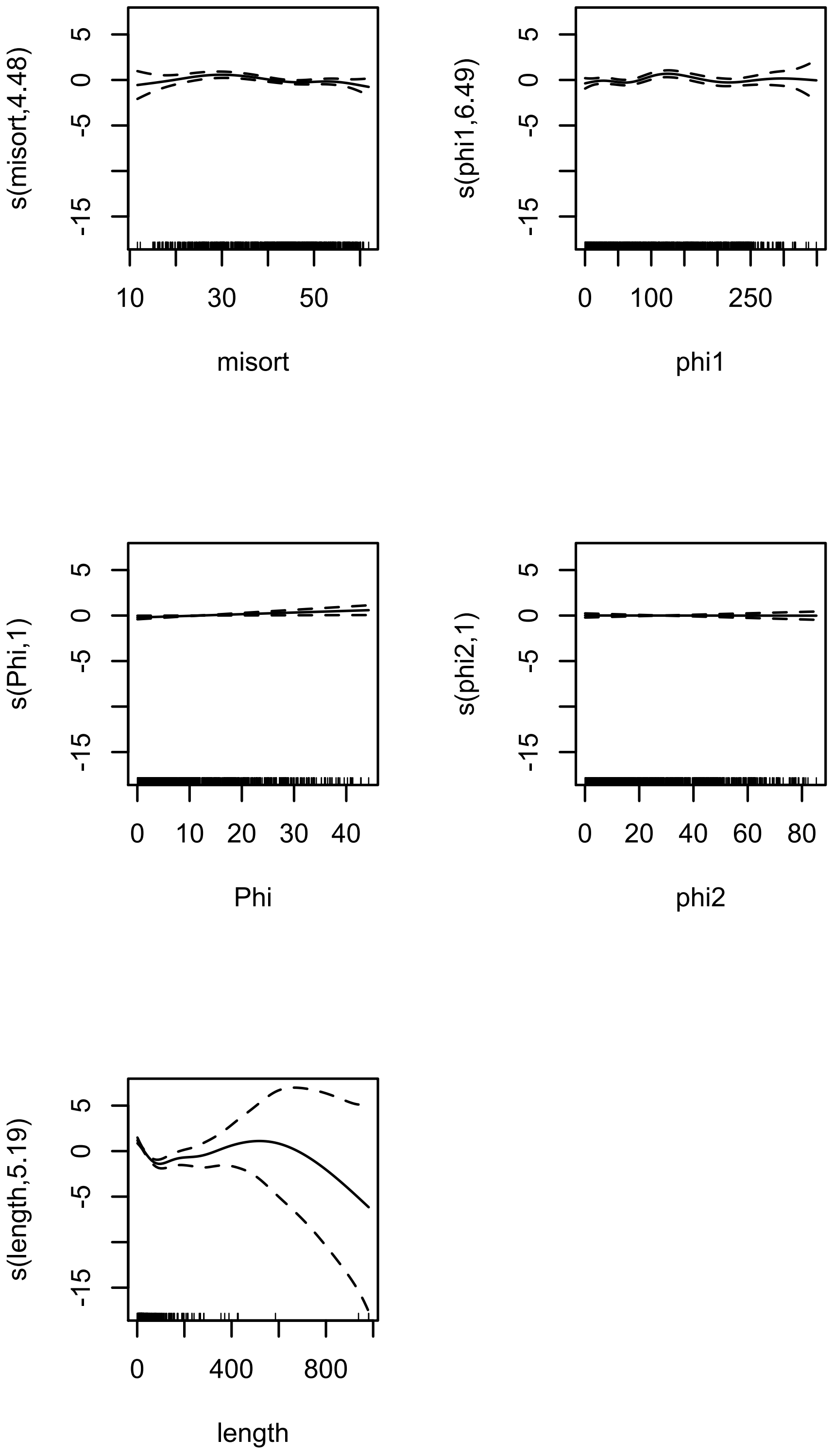


Figure 7.21: GAM plots of model part II

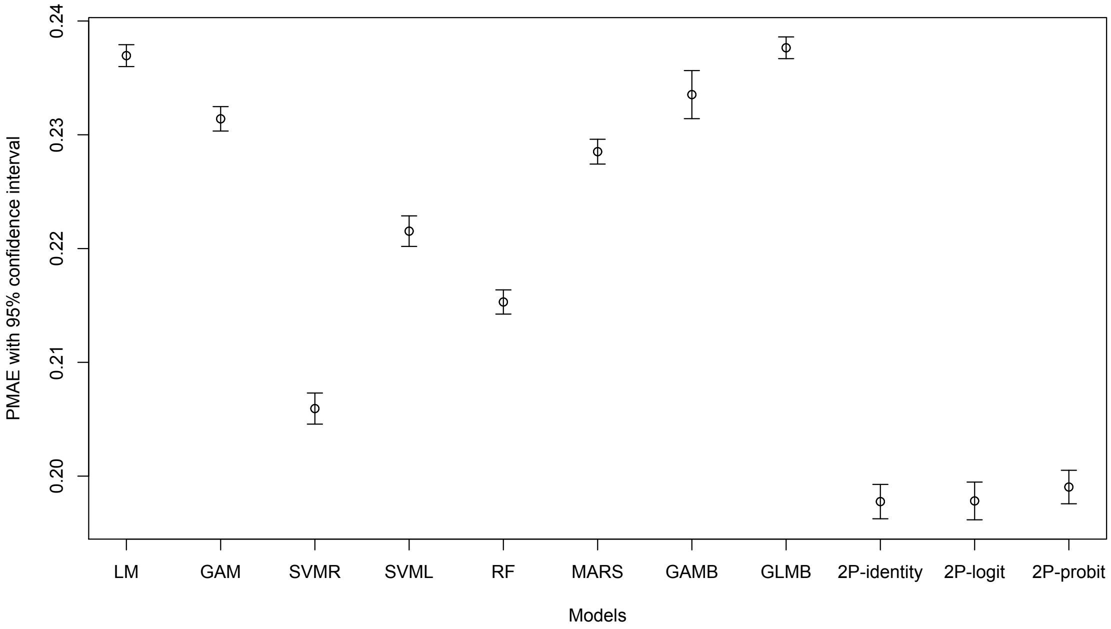


Figure 7.22: Mean PMAE comparison of 11 tested regression models with 95% confidence intervals for the corrosion example

CHAPTER 8

CONCLUSIONS

This chapter summarizes the work this dissertation has accomplished, addresses the conclusions drawn from the results of this research, and summarizes the contributions this dissertation has presented. Suggestion on future work has also been discussed in the last section.

8.1 Summary

In summary, this dissertation provides a formal supervised learning methodology when: (1) relevant information is stored in a series of images; (2) images are noisy and distorted from each other; (3) the data set for modeling is a data set with a semi-continuous response variable; and (4) the modeling goal is to understand the causal mechanism between variables and to predict future events accurately.

Many efforts have been made for problems with one of these characteristics, but there is no effective method that works as a whole system for such a type of problems. In terms of these characteristics, three separate models have been constructed to form the developed methodology, and work as a whole system. More importantly, the three models are able to function independently with good performance, compared with previous research works. The first model is an image fusion algorithm, which is designed to extract relevant information from the image data sources. This algorithm is able to deal with images with noises as well as distortions, which provides an accurate data set for supervised learning. The second model is an outlier detection framework, which is built to filter out both outliers and influential observations from the data set. This framework ensures that the supervised learning model built based on these data describes the relationships between variables precisely. The last model is a two-part generalized hierarchical model for semi-continuous data. It is especially designed for supervised learning with semi-continuous data, which are widely existing in many scientific research fields, such as medicine, material science and economics. This model is able to deal with the semi-continuous data by providing a high prediction accuracy and good model interpretability, compared with many other classical regression models.

The developed formal methodology is applied to a real practical problem which needs to predict the growth of the intergranular corrosion for 5XXX-series alloys, based on several grain boundary characteristics, as well as explain the causal mechanisms between model factors. This intergranular corrosion prediction problem has the exact four characteristics described above. Images from electrochemical experiments are the major data source for modeling, and those images show a large amount of noises. Due to the limitation of the experiment, some of those images are even distorted from others. Also, the response variable in the model is a semi-continuous variable, so most classical regression models fail to fit such data. Therefore, this is a typical problem for the developed methodology in this dissertation. By solving this problem, the developed three models have been applied to the problem in the form of three subsystems. The output of the former model is the input of the next one. Each model is evaluated and tested before proceeding to the next one. At last, this supervised learning methodology shows a high prediction accuracy of the intergranular corrosion, given a group of grain boundary characteristics. It also provides explains for the causal mechanisms behind those factors. Conclusions drawn from results of this application are able to guide material scientists and researchers about how to explain the effects of the physical property of materials on the growth pattern of corrosion. Furthermore, these explanations will help with the research development of the corrosion resistance of aluminum alloys, in order to prevent corrosion damages in construction and industrial fields.

8.2 Conclusions

The following conclusions are drawn from the results of the developed methodology, which is applied to an intergranular corrosion prediction problem with data from 5XXX-series alloys. These conclusions are based on the three separate models which are evaluated and tested individually with the provided sample data. This section also discusses results and conclusions from the modeling for this application.

(1) The regression results of the two-part generalized hierarchical model for semi-continuous data imply that grain boundary characteristics $\Delta\phi 1$ and $\Delta\Phi$ are significant in model part I, which means they mainly affect the β coverage in the classification of 0 or 1. Grain boundary length l and misorientation angle θ are significant variables affecting the percent β coverage within the range of 0 to 1, which might lead to the importance of grain shapes in determining how intergranular corrosion grows preferentially along grain boundaries. The prediction accuracy of this model ranks the highest, compared with other eight classical supervised learning methods.

(2) The developed image fusion model shows a fast yet efficient method to register a group of distorted images with noises, and extract valuable information from the registered image. In the material science area, image registration work is usually done by hand. Thus, image registration is often time-consuming and the results are objective, depending on the operator's criteria. This developed image fusion model is not only able to save tons of processing time, but also can provide a standard method with a high accuracy. To evaluate this method, it is compared with a widely used algorithm in Matlab. The registration results from these two

methods show that the developed model provides a significantly higher accuracy, because it has a noise reduction function with the distortion correction.

(3) The outlier detection framework is designed to filter out two kinds of outliers. One is the noisy outlier which would mislead the regression result if it is kept in the data set. The other kind is the influential observation which perform differently from regular observations but can provide important information to the regression model. Thus, it is critical to distinguish between these two kinds of outliers and treat them differently. The developed outlier detection framework is compatible to various supervised learning models, and the results of the applied corrosion problem indicate that after using this framework, the prediction accuracy of the regression model is significantly higher than the one without it. Such an improvement on the prediction accuracy is attributed to the distinguishing between two different kinds of outliers in the developed framework.

8.3 Contributions

The developed formal methodology is applied to an intergranular corrosion problem in this dissertation, but it is also applicable to similar phenomena, such as weather forecasting of different areas based on satellite cloud pictures, tumor growth prediction based on MRI images and epidemic disease spreading modeling. This dissertation provides four major contributions:

(1) A supervised learning methodology is constructed to model processes with information stored in both semi-continuous data and a series of noisy images. Corrosion prediction problems can be solved by this methodology and we provide a case study.

(2) An outlier detection framework for supervised learning is constructed to enhance prediction accuracy. It is applied to a generalized hierarchical model for the

corrosion prediction problem. It is applicable to other supervised learning models, such as linear regression, random forest and generalized additive model.

(3) An image fusion algorithm is presented that can extract and combine information from multiple noisy images. Corrosion images taken by different equipment are fused by this algorithm to collect data for modeling. It might also be applicable to a wider range of areas such as magnetic resonance imaging (MRI) for tumor growth monitoring and geographic information integrating.

(4) The estimation of the generalized hierarchical model can help material scientists to reveal the causal mechanisms between grain boundary characteristics and the intergranular corrosion, as well as to predict future corrosion occurrences.

8.4 Suggestions for Future Work

(1) In the application of this dissertation, five grain boundary characteristics have been collected for modeling the growth pattern of the intergranular corrosion in 5XXX-series alloys. In order to explore the causal mechanisms between IGC and grain boundary characteristics more precisely, more features should be considered and added to the current data set, such as CSL, grain boundary plane and grain boundary energy, in addition to the five grain boundary characteristics.

(2) Significant grain boundary characteristics affecting the growth of intergranular corrosion are given by the GAM model in the third subsystem. Investigation on the physical property of the relationship between significant variables and β phase is needed, in order to improve the understanding of grain boundary energy as well as to enhance the corrosion resistance of aluminum alloys in the future.

(3) The developed two-part generalized hierarchical model outperforms most of the popular supervised learning models for semi-continuous data, when applied to

the corrosion data. A theoretical proof is needed to show the effectiveness of such a model, from the aspect of statistical learning theory.

(4) The developed formal methodology is a whole system for problems with specific characteristics. A software can be designed and implemented with this methodology, so that users are able to complete the analysis efficiently and quickly.

(5) In this dissertation, all of the three models are applied to one problem, which is the intergranular corrosion prediction problem. However, the three subsystems are able to function as independent models for various problems in different fields. To evaluate the subsystems, it is necessary to apply them to different problems with various data set.

BIBLIOGRAPHY

Abe, N., Zadrozny, B., and Langford, J. (2006). Outlier detection by active learning. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 504–509. ACM.

Aggarwal, C. and Yu, P. (2001). Outlier detection for high dimensional data. *ACM Sigmod Record*, 30(2):37–46.

Aggarwal, C. and Yu, P. (2008). Outlier detection with uncertain data. In *Proc. SIAM Intĺ Conf. Data Mining (SDM)*. Citeseer.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, volume 1, pages 267–281. Springer Verlag.

Allen, T. (2010). Personal Communication. `http://www.engr.wisc.edu/ep/faculty/allen_todd.html`.

Amit, Y. and Geman, D. (1997). Shape quantization and recognition with randomized trees. *Neural computation*, 9(7):1545–1588.

Bajcsy, R. and Kovacic, S. (1989). Multiresolution elastic matching. *Computer vision, graphics, and image processing*, 46(1):1–21.

Barnea, D. and Silverman, H. (1972). A class of algorithms for fast digital image registration. *IEEE Transactions on Computers*, 21(2):179–186.

Barnett, V. and Lewis, T. (1994). Outliers in statistical data.

Beal, S. and Sheiner, L. (1980). The NONMEM system. *American Statistician*, 34(2):118–119.

Bennett, J., Racine-Poon, A., and Wakefield, J. (1996). MCMC for nonlinear hierarchical models. *Markov chain Monte Carlo in practice*, pages 339–57.

Boser, B., Guyon, I., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM.

Bracewell, R. (2000). The fourier transform & its applications 3rd Ed.

Breiman, L. (1996). Bagging predictors. *Machine learning*, 24(2):123–140.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, pages 9–25.

Briant, C. (1980). The effects of sulfur and phosphorus on the intergranular corrosion of 304 stainless steel. *Corrosion*, 36(9):497–508.

Brossart, D., Parker, R., and Castillo, L. (2011). Robust regression for single-case data analysis: How can it help? *Behavior research methods*, pages 1–10.

Brown, L. (1992). A survey of image registration techniques. *ACM computing surveys (CSUR)*, 24(4):325–376.

Butler, E. and Swann, P. (1976). In situ observations of the nucleation and initial growth of grain boundary precipitates in an Al-Zn-Mg alloy. *Acta Metallurgica*, 24(4):343–352.

Cain, S., Hayat, M., and Armstrong, E. (2002). Projection-based image registration in the presence of fixed-pattern noise. *Image Processing, IEEE Transactions on*, 10(12):1860–1872.

Caner, G., Tekalp, A., Sharma, G., and Heinzelman, W. (2006). Local image registration by adaptive filtering. *Image Processing, IEEE Transactions on*, 15(10):3053–3065.

Cerioli, A. (2010). Multivariate outlier detection with high-breakdown estimators. *Journal of the American Statistical Association*, 105(489):147–156.

Chan, L., Weiland, H., Cheong, S., Rohrer, G., and Rollett, A. (2008). The correlation between grain boundary character and intergranular corrosion susceptibility of 2124 aluminum alloy. *Applications of Texture Analysis: Ceramic Transactions*, page 261.

Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, 41(3):1–58.

Chatterjee, S. and Hadi, A. (1986). Influential observations, high leverage points, and outliers in linear regression. *Statistical Science*, 1(3):379–393.

Chi, E. and Reinsel, G. (1989). Models for longitudinal data with random effects and AR (1) errors. *Journal of the American Statistical Association*, 84(406):452–459.

Chu, W., Ma, L., Song, J., and Vorburger, T. (2010). An Iterative Image Registration Algorithm by Optimizing Similarity Measurement. *Journal of Research of the National Institute of Standards and Technology*, 115(1):1–6.

Crump, S. (1946). The estimation of variance components in analysis of variance. *Biometrics Bulletin*, 2(1):7–11.

Daubechies, I. (1992). *Ten lectures on wavelets*. Society for Industrial Mathematics.

Davidian, M. and Gallant, A. (1992). Smooth nonparametric maximum likelihood estimation for population pharmacokinetics, with application to quinidine. *Journal of Pharmacokinetics and Pharmacodynamics*, 20(5):529–556.

Day, A. and Trimby, P. (2004). Channel 5 Manual HKL Technology Inc. *Hobro, Denmark*.

de Boor, C. (1978). A practical guide to splines. 1978. *Applied mathematical sciences*.

De Castro, E. and Morandi, C. (2009). Registration of translated and rotated images using finite Fourier transforms. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (5):700–703.

Dingley, D. and Randle, V. (1992). Microtexture determination by electron backscatter diffraction. *Journal of Materials Science*, 27(17):4545–4566.

Diniz, P. (2010). *Digital signal processing*. Cambridge University Press.

Dix, E., Anderson, W., and Shumaker, M. (1958). *Development of Wrought Aluminum-Magnesium Alloys*. Aluminum Co. of America.

Duda, R., Hart, P., and Stork, D. (2001). *Pattern classification*, volume 2. Citeseer.

Ehlers, M. and Fogel, D. (1994). High-precision geometric correction of airborne remote sensing revisited: the multiquadric interpolation. In *Proceedings of SPIE*, volume 2315, page 814.

Ertoz, L., Eilertson, E., Lazarevic, A., Tan, P., Dokas, P., Kumar, V., and Srivastava, J. (2003). Detection of novel network attacks using data mining. In *Proc. of Workshop on Data Mining for Computer Security*. Citeseer.

Filzmoser, P., Maronna, R., and Werner, M. (2008). Outlier identification in high dimensions. *Computational Statistics & Data Analysis*, 52(3):1694–1711.

Fonseca, L. and Manjunath, B. (1996). Registration techniques for multisensor remotely sensed imagery. *Photogrammetric Engineering and Remote Sensing*, 62(9):1049–1056.

for Testing, A. S. and Materials (2002). Annual book of ASTM standards. American Society for Testing and Materials, PA.

Freund, Y. and Schapire, R. (1995). A desicion-theoretic generalization of on-line learning and an application to boosting. In *Computational learning theory*, pages 23–37. Springer.

Friedman, J. (1991). Multivariate adaptive regression splines. *The annals of statistics*, pages 1–67.

Friedman, J. (1994). An overview of predictive learning and function approximation. *NATO ASI SERIES F COMPUTER AND SYSTEMS SCIENCES*, 136:1–1.

Friedman, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, 38(4):367–378.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*. Springer Series in Statistics.

Gallant, A. and Corporation, E. (1987). Nonlinear statistical models.

Garg, A. and Howe, J. (1992). Grain-boundary precipitation in an Al—4.0 Cu—0.5 Mg—0.5 Ag alloy. *Acta Metallurgica et Materialia*, 40(9):2451–2462.

Gelfand, A., Hills, S., Racine-Poon, A., and Smith, A. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, pages 972–985.

Gelman, A., Hill, J., and Corporation, E. (2007). *Data analysis using regression and multilevel/hierarchical models*, volume 625. Cambridge University Press Cambridge.

Geman, S., Geman, D., and Relaxation, S. (1984). Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(2):721–741.

Goldstein, H. (1995). Multilevel statistical models. *Kendall's library of statistics.*

Golub, G., Heath, M., and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, pages 215–223.

Good, P. and Wang, R. (2005). *Permutation, parametric and bootstrap tests of hypotheses*. Springer New York:.

Goshtasby, A. (1986). Piecewise linear mapping functions for image registration. *Pattern Recognition*, 19(6):459–466.

Goshtasby, A. (2005). *2-D and 3-D image registration for medical, remote sensing, and industrial applications*. Wiley-Interscience.

Goshtasby, A. and Stockman, G. (1985). Point pattern matching using convex hull edges. *IEEE TRANS. SYST. MAN CYBER.*, 15(5):631–636.

Green, P. and Silverman, B. (1994). *Nonparametric regression and generalized linear models: a roughness penalty approach*. Chapman & Hall/CRC.

Grossman, M. and Koops, W. (1988). Multiphasic analysis of growth curves in chickens. *Poultry science*, 67(1):33.

Hartley, H. and Rao, J. (1967). Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika*, 54(1-2):93.

Harville, D. (1974). Bayesian inference for variance components using only error contrasts. *Biometrika*, 61(2):383.

Hasson, G., Boos, J., Herbeuval, I., Biscondi, M., and Goux, C. (1972). Theoretical and experimental determinations of grain boundary structures and energies: Correlation with various experimental results. *Surface Science*, 31:115–137.

Hastie, T. and Tibshirani, R. (1987). Generalized additive models: some applications. *Journal of the American Statistical Association*, pages 371–386.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. Chapman & Hall/CRC.

Hawkins, D. (1980). *Identification of outliers*. Chapman & Hall.

Henderson, C. (1953). Estimation of variance and covariance components. *Biometrics*, 9(2):226–252.

Ho, T. (1998). The random subspace method for constructing decision forests. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(8):832–844.

Hoaglin, D. and Welsch, R. (1978). The hat matrix in regression and ANOVA. *American Statistician*, 32(1):17–22.

Huber, P. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Ingle, V. and Proakis, J. (1999). *Digital signal processing using MATLAB*. Brooks/Cole Publishing Co. Pacific Grove, CA, USA.

Jakupi, P., Noël, J., and Shoesmith, D. (2010). Intergranular Corrosion Resistance of $\Sigma 3$ Grain Boundaries in Alloy 22. *Electrochemical and Solid-State Letters*, 13:C1.

Johnson, R. and Wichern, D. (2002). *Applied multivariate statistical data analysis.* Prentice Hall: Upper Saddle River, NJ.

Kapur, J. and Casasent, D. (2000). Geometric Correction of SEM images. In *Proceedings of SPIE, the International Society for Optical Engineering*, volume 4044, pages 165–176. Society of Photo-Optical Instrumentation Engineers.

Kavner, A. and Devine, T. (1997). Effect of grain boundary orientation on the sensitization of austenitic stainless steel. *Journal of materials science*, 32(6):1555–1562.

Kim, C., Rollett, A., and Rohrer, G. (2006). Grain boundary planes: New dimensions in the grain boundary character distribution. *Scripta materialia*, 54(6):1005–1009.

Kim, S., Erb, U., Aust, K., and Palumbo, G. (2001). Grain boundary character distribution and intergranular corrosion behavior in high purity aluminum. *Scripta materialia*, 44(5):835–840.

Knorr, E. and Ng, R. (1998). Algorithms for mining distance-based outliers in large datasets. In *Proceedings of the International Conference on Very Large Data Bases*, pages 392–403. Citeseer.

Koch, G. et al. (2002). *Corrosion cost and preventive strategies in the United States.* Turner-Fairbank Highway Research Center.

Kooperberg, C., Bose, S., and Stone, C. (1997). Polychotomous regression. *Journal of the American Statistical Association*, pages 117–127.

Kou, Y., Lu, C., Sirwongwattana, S., and Huang, Y. (2005). Survey of fraud detection techniques. In *Networking, Sensing and Control, 2004 IEEE International Conference on*, volume 2, pages 749–754. IEEE.

Kriegel, H. et al. (2008). Angle-based outlier detection in high-dimensional data.

In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 444–452. ACM.

Kriegel, H., Kröger, P., Schubert, E., and Zimek, A. (2009). Outlier detection in axis-parallel subspaces of high dimensional data. *Advances in Knowledge Discovery and Data Mining*, pages 831–838.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Lam, L., Lee, S., and Suen, C. (2002). Thinning methodologies-a comprehensive survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 14(9):869–885.

Latecki, L., Lazarevic, A., and Pokrajac, D. (2007). Outlier detection with kernel density functions. *Machine Learning and Data Mining in Pattern Recognition*, pages 61–75.

Le Moigne, J., Xia, W., Chalermwat, P., El-Ghazawi, T., Mareboyana, M., Netanyahu, N., Tilton, J., Campbell, W., and Cromp, R. (2002). First evaluation of automatic image registration methods. In *Geoscience and Remote Sensing Symposium Proceedings, 1998. IGARSS'98. 1998 IEEE International*, volume 1, pages 315–317. IEEE.

Lester, H. and Arridge, S. (1999). A survey of hierarchical non-linear medical image registration. *Pattern Recognition*, 32(1):129–149.

Li, J., Dillon, S., and Rohrer, G. (2009). Relative grain boundary area and energy distributions in nickel. *Acta Materialia*, 57(14):4304–4311.

Liang, K. and Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13.

Lilliefors, H. (1967). On the kolmogorov-smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, pages 399–402.

Lindsey, J. and Aickin, M. (1994). *Models for repeated measurements*. Clarendon Press Oxford.

Liu, L., Ma, J., and Johnson, B. (2008). A multi-level two-part random effects model, with application to an alcohol-dependence study. *Statistics in medicine*, 27(18):3528–3539.

Lo, K., Shek, C., and Lai, J. (2009). Recent developments in stainless steels. *Materials Science and Engineering: R: Reports*, 65(4-6):39–104.

Loader, C. (1999). *Local regression and likelihood*. Springer Verlag.

Longford, N. (1993). *Random coefficient models*. Oxford University Press, USA.

Mahadevan, S. and Casasent, D. (2003). Automated image processing for grain boundary analysis. *Ultramicroscopy*, 96(2):153–162.

Maintz, J. and Viergever, M. (1998). A survey of medical image registration. *Medical image analysis*, 2(1):1–36.

Mallet, A., Mentré, F., Steimer, J., and Lokiec, F. (1988). Nonparametric maximum likelihood estimation for population pharmacokinetics, with application to cyclosporine. *Journal of Pharmacokinetics and Pharmacodynamics*, 16(3):311–327.

Mardia, K., Kent, J., and Bibby, J. (1980). Multivariate analysis.

Mielke, P., Berry, K., and Johnson, E. (1976). Multi-response permutation procedures for a priori classifications. *Communications in Statistics-Theory and Methods*, 5(14):1409–1424.

Min, Y. and Agresti, A. (2002). Modeling nonnegative data with clumping at zero: A survey. *Journal of the Iranian Statistical Society*, 1(1-2):7–33.

Mondolfo, L. (1976). *Aluminum alloys: structure and properties*, volume 5. Butterworths London.

Naseem, I., Togneri, R., and Bennamoun, M. (2011). Robust regression for face recognition. *Pattern Recognition*.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, pages 370–384.

Nguyen, T. and Welsch, R. (2009). Outlier detection and least trimmed squares approximation using semi-definite programming. *Computational Statistics & Data Analysis*.

Olsen, M. and Schafer, J. (2001). A two-part random-effects model for semicontinuous longitudinal data. *Journal of the American Statistical Association*, 96(454):730–745.

Ozdemir, S. and Casasent, D. (1999). Scale-space median and gabor filtering and fuzzy unification for boundary detection in electron microscopy images. In *NSIP*, pages 617–621.

Pan, Y., Adams, B., Olson, T., and Panayotou, N. (1996). Grain-boundary structure effects on intergranular stress corrosion cracking of alloy X-750. *Acta Materialia*, 44(12):4685–4695.

Park, P., Manjourides, J., Bonetti, M., and Pagano, M. (2009). A permutation test for determining significance of clusters with applications to spatial and gene expression data. *Computational Statistics & Data Analysis*, 53(12):4290–4300.

Patterson, H. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58(3):545.

Pinheiro, J. (1994). *Topics in mixed effects models.* PhD thesis, UNIVERSITY OF WISCONSIN.

Pinheiro, J. and Bates, D. (2009). *Mixed-effects models in S and S-PLUS.* Springer Verlag.

Randle, V. (1996). *The role of the coincidence site lattice in grain boundary engineering.* Institute of Materials London.

Randle, V., Institute of Materials, M., and Mining (1992). *Microtexture determination and its applications.* The Institute of Materials London.

Read, W. and Shockley, W. (1950). Dislocation models of crystal grain boundaries. *Physical Review*, 78(3):275.

Riani, M., Atkinson, A., and Cerioli, A. (2009). Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):447–466.

Rousseeuw, P. (1984). Least median of squares regression. *Journal of the American statistical association*, pages 871–880.

Rousseeuw, P. and Leroy, A. (1987). *Robust regression and outlier detection*, volume 3. Wiley Online Library.

Russ, J. (2007). *The image processing handbook.* CRC.

Sabuncu, M. (2004). *Entropy-based Image Registration.* Citeseer.

Schapire, R. and Singer, Y. (1999). Improved boosting algorithms using confidence-rated predictions. *Machine learning*, 37(3):297–336.

Schapire, Y. and Freund, Y. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780.

Searle, S., Casella, G., McCulloch, C., et al. (1992). *Variance components*. Wiley Online Library.

Seber, G. (2004). *Multivariate observations*. Wiley.

She, Y. and Owen, A. (2010). Outlier detection using nonconvex penalized regression. *Arxiv preprint arXiv:1006.2592*.

Shimada, M., Kokawa, H., Wang, Z., Sato, Y., and Karibe, I. (2002). Optimization of grain boundary character distribution for intergranular corrosion resistant 304 stainless steel by twin-induced grain boundary engineering. *Acta Materialia*, 50(9):2331–2341.

Skidmore, T., Buchheit, R., and Juhas, M. (2004). Grain boundary energy vs. misorientation in inconel¡ sup¿®¡/sup¿ 600 alloy as measured by thermal groove and oim analysis correlation. *Scripta materialia*, 50(6):873–877.

Sorenson, H. (1970). Least-squares estimation: from gauss to kalman. *Spectrum, IEEE*, 7(7):63–68.

Stanghellini, E. and Gottard, A. (2011). Semicontinuous regression models with skew distributions.

Tedmon Jr, C., Vermilyea, D., and Rosolowski, J. (1971). Intergranular corrosion of austenitic stainless steel. *journal of the Electrochemical Society*, 118:192.

Thompson Jr, W. (1962). The problem of negative estimates of variance components. *The Annals of Mathematical Statistics*, 33(1):273–289.

Tschopp, M. and McDowell, D. (2007). Asymmetric tilt grain boundary structure and energy in copper and aluminium. *Philosophical Magazine*, 87(25):3871–3892.

Unwin, P., Lorimer, G., and Nicholson, R. (1969). The origin of the grain boundary precipitate free zone. *Acta Metallurgica*, 17(11):1363–1377.

Unwin, P. and Nicholson, R. (1969). The nucleation and initial stages of growth of grain boundary precipitates in Al-Zn-Mg and Al-Mg alloys. *Acta Metallurgica*, 17(11):1379–1393.

Van den Elsen, P., Pol, E., and Viergever, M. (2002). Medical image matching-a review with classification. *Engineering in Medicine and Biology Magazine, IEEE*, 12(1):26–39.

Vapnik, V. (2000). *The nature of statistical learning theory*. Springer Verlag.

Vaughan, D. (1968). Grain boundary precipitation in an Al—Cu alloy. *Acta Metallurgica*, 16(4):563–577.

Velleman, P. and Welsch, R. (1981). Efficient computing of regression diagnostics. *American Statistician*, 35(4):234–242.

Vonesh, E. and Carter, R. (1992). Mixed-effects nonlinear regression for unbalanced repeated measures. *Biometrics*, 48(1):1–17.

Wakefield, J., Smith, A., Racine-Poon, A., and Gelfand, A. (1994). Bayesian analysis of linear and non-linear population models by using the Gibbs sampler. *Applied Statistics*, 43(1):201–221.

Wand, M. and Jones, M. (1995). *Kernel smoothing*, volume 60. Chapman & Hall/CRC.

Weisberg, S. (2005). *Applied linear regression*. Wiley.

West, J., Fitzpatrick, J., Wang, M., Dawant, B., Maurer Jr, C., Kessler, R., Maciunas, R., Barillot, C., Lemoine, D., Collignon, A., et al. (1997). Comparison and evaluation of retrospective intermodality brain image registration techniques. *Journal of Computer Assisted Tomography*, 21(4):554.

Wiemker, R., Rohr, K., Binder, L., Sprengel, R., and Stiehl, H. (1996). Application of elastic registration to imagery from airborne scanners. *International Archives of Photogrammetry and Remote Sensing*, 31:949–954.

Wilcox, R. (2012). *Introduction to robust estimation and hypothesis testing*. Academic Press.

Wollny, G. and Kruggel, F. (2002). Computational cost of nonrigid registration algorithms based on fluid dynamics [mri time series application]. *Medical Imaging, IEEE Transactions on*, 21(8):946–952.

Wood, S. (2006). *Generalized additive models: an introduction with R*, volume 66. CRC Press.

Wood, S. (2007). The mgcv package. *www. r-project. org*.

Wood, S. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):495–518.

Xu, H., Yang, L., and Freitas, M. (2008). A robust linear regression based algorithm for automated evaluation of peptide identifications from shotgun proteomics by use of reversed-phase liquid chromatography retention time. *BMC bioinformatics*, 9(1):347.

Yuan, Y. (2006). *Localised Corrosion and Stress Corrosion Cracking of Aluminium-magnesium Alloys*. University of Birmingham.

Zeger, S., Liang, K., and Albert, P. (1988). Models for longitudinal data: a generalized estimating equation approach. *Biometrics*, 44(4):1049–1060.

Zhang, Y., Meratnia, N., and Havinga, P. (2010). Outlier detection techniques for wireless sensor networks: A survey. *Communications Surveys & Tutorials, IEEE*, 12(2):159–170.

Zitova, B. and Flusser, J. (2003). Image registration methods: a survey. *Image and vision computing*, 21(11):977–1000.