**Applied Data Analysis and Modeling**

**for Energy Engineers and Scientists- 2nd Ed**

**(status on March 2023)**

**Chapter 1. Mathematical Models and Data Analysis**

Abstract

Keywords

**Table of Contents**

1.1. Forward and Inverse Approaches

1.1.1 Preamble

1.1.2 The energy problem and importance of buildings

1.1.3 Forward or simulation approach

1.1.4 Inverse or data analysis approach

1.1.5 Discussion of both approaches

1.2. System Models

1.2.1 What is a system model?

1.2.2 Types of models

1.3 Types of Data

1.3.1 Classification

1.3.2 Types of uncertainty in data

1.4 Mathematical Models

1.4.1 Basic terminology

1.4.2 Block diagrams

1.4.3 Mathematical representation

1.4.4 Classification

1.4.5 Steady state and dynamic models

1.5. Mathematical Modeling Approaches

1.5.1 Broad categorization

1.5.2 Simulation or forward modeling

1.5.3 Inverse modeling

1.5.4 Calibrated simulation

1.6.Data Analytic Approaches

1.6.1 Data mining or knowledge discovery

1.6.2 Machine learning or algorithmic models

1.6.3 Introduction to big data

1.7 Data Analysis

1.7.1 Introduction

1.7.2 Basic stages

1.7.3 Example of a data collection and analysis system

1.8. Topics covered in book

Problems

References

**Chapter 2. Probability Concepts and Probability Distributions**

Abstract

Keywords

**Table of Contents**

2.1. Introduction

2.1.1 Classical concept of probability

2.1.2 Bayesian viewpoint of probability

2.1.3 Distinction between probability and statistics

2.2. Classical Probability

2.2.1 Basic terminology

2.2.2 Basic set theory notation and axioms of probability

2.2.3 Axioms of probability

2.2.4 Joint, marginal and conditional probabilities

2.2.5 Permutations and combinations

2.3. Probability Distribution Functions

2.3.1 Density functions

2.3.2 Expectations and moments

2.3.3 Function of random variables

2.3.4 Chebyshev’s theorem

2.4. Important Probability Distributions

2.4.1 Background

2.4.2 Distributions for discrete variables

2.4.3 Distributions for continuous variables

2.5. Bayesian Probability

2.5.1 Bayes’ theorem

2.5.2 Application to discrete probability variables

2.5.3 Application to continuous probability variables

2.6 Three Kinds of Probability

Problems

References

**Chapter 3. Data Collection and Preliminary Analysis**

Abstract

Keywords

**Table of Contents**

3.1. Sensors and their Characteristics

3.2. Data Collection Systems

3.2.1. Generalized measurement system

3.2.2 Types and categories of measurements

3.2.3. Data recording systems

3.3 Raw Data Validation and Preparation

3.3.1 Definitions

3.3.2 Limit checks

3.3.3 Consistency checks involving conservation balances

3.3.4 Outlier rejection by visual means

3.3.5 Handling missing data

3.4. Statistical Measures for Sample Data

3.4.1 Summary descriptive measures

3.4.2 Covariance and Pearson correlation coefficient

3.5. Exploratory Data Analysis (EDA)

3.5.1 What is EDA?

3.5.2 Purpose of data visualization

3.5.3 Static univariate graphical plots

3.5.4 Static bi- and multi-variate graphical plots

3.5.5 Interactive and dynamic graphics

3.5.6 Basic data transformations

3.6. Overall Measurement Uncertainty

3.6.1 Need for uncertainty analysis

3.6.2 Basic uncertainty concepts: Random and bias errors

3.6.3 Random uncertainty

3.6.4 Bias uncertainty

3.6.5 Overall uncertainty

3.6.6 Chauvenet’s statistical criterion of data rejection

3.7. Propagation of Errors

3.7.1. Taylor series method for cross-sectional data

3.7.2. Monte Carlo method for error propagation problems

3.8. Planning a Non-intrusive Field Experiment

Problems

References

**Chapter 4. Making Statistical Inferences from Samples**

Abstract

Keywords

**Table of Contents**

4.1 Introduction

4.2 Basic Univariate Inferential Statistics

4.2.1 Sampling distribution and confidence interval of the mean

4.2.2 Hypothesis test for single sample mean

4.2.3 Two independent sample and paired difference tests on means

4.2.4 Single and two sample tests for proportions

4.2.5 Single and two sample tests of variance

4.2.6 Tests for distributions

4.2.7 Test on the Pearson correlation coefficient

4.3 ANOVA Test for Multi-Samples

4.3.1 Single-factor ANOVA

4.3.2 Tukey’s multiple comparison test

4.4 Tests of Significance of Multivariate Data

4.4.1 Introduction to multivariate methods

4.4.2 Hotteling *T2* test

4.5 Non-parametric Tests

4.5.1 Signed and rank tests for medians

4.5.2 Kruskall-Wallis multiple samples test for medians

4.5.3 Test on Spearman rank correlation coefficient

4.6 Bayesian Inferences

4.6.1 Background

4.6.2 Estimating population parameter from a sample

4.6.3 Hypothesis testing

4.7 Some Considerations about Sampling

4.7.1 Random and non-random sampling methods

4.7.2 Desirable properties of estimators

4.7.3 Determining sample size during random surveys

4.7.4 Stratified sampling for variance reduction

4.8 Resampling Methods

4.8.1 Basic concept

4.8.2 Application to probability problems

4.8.3 Different methods of resampling

4.8.4 Application of bootstrap to statistical inference problems

4.8.5 Closing remarks

Problems

References

**Chapter 5. Linear Regression Analysis Using Least Squares**

Abstract

Keywords

**Table of Contents**

5.1 Introduction

5.2 Regression Analysis

5.2.1 Objective of regression analysis

5.2.2 Ordinary least squares

5.3 Simple OLS Regression

5.3.1 Estimation of model parameters

5.3.2 Statistical criteria for model evaluation

5.3.3 Inferences on regression coefficients and model significance

5.3.4 Model Prediction uncertainty

5.4 Multiple OLS Regression

5.4.1 Higher order linear models

5.4.2 Matrix formulation

5.4.3 Point and interval estimation

5.4.4 Beta coefficients and elasticity

5.4.5 Partial correlation coefficients

5.4.6 Assuring model parsimony—Stepwise regression

5.5 Applicability of OLS Parameter Estimation

5.5.1 Assumptions

5.5.2 Sources of errors during regression

5.6 Model Residual Analysis and Regularization

5.6.1 Detection of ill-conditioned model residual behavior

5.6.2 Leverage and influence data points

5.6.3 Remedies for non-uniform residuals

5.6.4 Serially correlated residuals

5.6.5 Dealing with misspecified models

5.7 Other Useful OLS Regression Models

5.7.1 Zero-intercept model

5.7.2 Indicator variables for local piecewise models—Linear splines

5.7.3 Indicator variables for categorical regressor models

5.8 Resampling Methods Applied to Regression

5.8.1 Basic approach

5.8.2 Jackknife and *k*-fold cross-validation

5.8.3 Bootstrap method

5.9 Case Study Example: Effect of Refrigerant Additive on Chiller Performance

5.10 Parting Comments on OLS

Problems

References

**Chapter 6. Design of Physical and Simulation Experiments**

Abstract

Keywords

**Table of Contents**

6.1 Introduction

* + 1. Types of data collection
    2. Purpose of DOE
    3. DOE terminology

6.2 Overview of Different Statistical Methods

6.2.1 Different types of ANOVA tests

* + 1. Link between ANOVA and regression
    2. Recap of basic model functional forms
  1. Basic Concepts
     1. Levels, discretization and experimental combinations
     2. Blocking
     3. Unrestricted and restricted randomization

6.4 Factorial Designs

6.4.1 Full factorial design

6.4.2 *2k* factorial designs

6.4.3 Concept of orthogonality

6.4.4 Fractional factorial designs

6.5 Block Designs

6.5.1 Complete block design

6.5.2 Latin squares

6.6 Response Surface Designs

6.6.1 Applications

6.6.2 Methodology

6.6.3 First and second order models

6.6.4 Central composite design and concept of rotation

6.7 Design of Simulation Experiments

6.7.1 Background

6.7.2 Similarities and differences between physical and simulation experiments

6.7.3 Monte Carlo and allied sampling methods

6.7.4 Sensitivity analysis for screening

6.7.5 Surrogate modeling

6.7.6 Summary

Problems

References

**Chapter 7. Optimization Methods**

Abstract

Keywords

**Table of Contents**

7.1. Introduction

7.1.1 What is optimization?

7.1.2 Simple example

7.2. Terminology and Classification

7.2.1. Definition of terms

7.2.2. Categorization of methods

7.2.3. Types of objective functions and constraints

7.2.4. Sensitivity analysis and post optimality analysis

7.3. Analytical Methods

7.3.1. Unconstrained problems

7.3.2. Direct substitution method for equality constrained problems

7.3.3 Lagrange multiplier method for equality constrained problems

7.3.4. Problems with inequality constraints

7.3.5. Penalty function method

7.4. Numerical Unconstrained Search Methods

7.4.1 Univariate methods

7.4.2 Multivariate methods

7.5 Linear Programming (LP)

7.5.1 Standard form

7.5.2 Example of a LP problem

7.5.3 Linear network models

7.5.4 Example of maximizing flow in a transportation network

7.5.5 Mixed integer linear programming (MILP)

7.5.6 Example of reliability analysis of power network

7.6. Non-linear Programming

7.6.1 Standard form

7.6.2 Quadratic programming

7.6.3 Popular numerical search algorithms

7.7. Illustrative Example: Integrated Energy System (IES) for a Campus of Buildings

7.8 Introduction to Global Optimization

7.9 Examples of Dynamic Programming

Problems

References

**Chapter 8. Analysis of Time Series Data**

Abstract

Keywords

**Table of Contents**

8.1. Basic Concepts

8.1.1. Introduction

8.1.2. Terminology

8.1.3. Basic behavior patterns

8.1.4. Illustrative data set

8.2. General Model Formulations

8.3. Smoothing Methods

8.3.1. Arithmetic moving average (AMA)

8.3.2. Exponentially weighted moving average (EWA)

8.3.3. Determining structure by cross-validation

8.4. OLS Regression Models

8.4.1. Trend modeling

8.4.2. Trend and seasonal models

8.4.3 Forecast intervals

8.4.4. Fourier series models for periodic behavior

8.4.5. Interrupted time series

8.4.5.1. Abrupt one-time change in time

8.4.5.2. Gradual change over time

8.5. Stochastic Time Series Models

8.5.1. Introduction

8.5.2. ACF, PACF and data detrending

8.5.2.1. Autocorrelation function (ACF)

8.5.2.2. Partial autocorrelation function (PACF)

8.5.2.3. Detrending data by differencing

8.5.3. ARIMA Class of Models

8.5.3.1. Overview

8.5.3.2 ARMA models

8.5.3.3. MA models

8.5.3.4. AR models

8.5.3.5. Identification and forecasting

8.5.4. Recommendations on model identification

8.6. ARMAX or Transfer Function Models

8.6.1. Conceptual approach and benefit

8.6.2. Transfer function modeling of linear dynamic systems

8.7. Quality Control and Process Monitoring Using Control Chart Methods

8.7.1. Background and approach

8.7.2. Shewart control charts for variables

8.7.2.1 Mean charts

8.7.2.2 Range charts

8.7.3. Shewart control charts for attributes

8.7.4 Practical implementation issues of control charts

8.7.5 Time weighted monitoring

8.7.5.1 Cusum Charts

8.7.5.2 EWMA process

8.7.6. Concluding remarks

Problems

References

**Chapter 9 . Parametric and Non-Parametric Regression Methods**

Abstract

Keywords

**Table of Contents**

9.1 Introduction

9.2 Important Concepts in Parameter Estimation

9.2.1 Structural identifiability

9.2.2 Ill-conditioning

9.2.3 Numerical identifiability

9.3 Dealing with Collinear Regressors: Variable Selection and Shrinkage

9.3.1 Problematic issues

9. 3.2 Principal component analysis and regression

9.3.3 Ridge and lasso regression

9.3.4 Chiller case study involving collinear regressors

9.3.5 Other multivariate methods

9.4 Going Beyond OLS

9.4.1 Background

9.4.2 Maximum likelihood estimation (MLE)

9.4.3 Generalized linear models (GLM)

9.4.4 Box-Cox transformation

9.4.5 Logistic functions

9.4.6 Error in variable (EIV) and corrected least squares

9.5 Non-linear Parametric Regression

9.5.1 Detecting non-linear correlation

9.5.2 Different non-linear search methods

9.5.3 Overview of various parametric regression methods

9.6 Non-parametric Regression

9.6.1 Background

9.6.2 Extensions to linear models

9.6.3 Basis functions

9.6.4 Polynomial regression and smoothing splines

9.7 Local Regression- LOWESS Smoothing Method

9.8 Neural Networks- Multilayer Perceptron (MLP)

9.9 Robust Regression

Problems

References

**Chapter 10. Inverse Methods for Mechanistic Models**

Abstract

Keywords

**Table of Contents**

10.1 Fundamental Concepts

10.1.1 Applicability

10.1.2 Approaches and their characteristics

10.1.3 Mechanistic models

10.1.4 Scope of chapter

10.2 Grey-box Static Models

10.2.1 Basic notions

10.2.2 Performance models for solar photovoltaic systems

10.2.2.1 Weather-based models

10.2.2.2 Weather and cell temperature-based models

10.2.3 Grey-box and black-box models for water-cooled chillers

10.2.4 Sequential stagewise regression and selection of data windows

10.2.5 Case study of non-intrusive sequential parameter estimation involving building energy flows

10.2.6 Application to policy: Dose-response

10.3 Certain Aspects of Data Collection

10.3.1 Types of data collection

10.3.2 Measures of information content

10.3.3 Functional testing and data fusion

10.4 Grey-box Models for Dynamic Systems

10.4.1 Introduction

10.4.2 Sequential estimation of thermal network model parameters from controlled tests

10.4.3 Non-intrusive identification of thermal network models and parameters

10.4.4 State space representation and compartmental models

10.4.5 Example of a compartmental model

10.4.6 Practical issues during identification

10.5 Bayesian Regression and Parameter Estimation: Case Study

10.6 Calibration of Detailed Simulation Programs

10.6.1 Purpose

10.6.2 The basic issue

10.6.3 Detailed simulation models for energy use in buildings

10.6.4 Uses of calibrated simulation

10.6.5 Causes of differences

10.6.6 Definition of terms

10.6.7 Raw input tuning (RIT)

10.6.8 Semi-analytical methods (SAM)

10.6.9 Physical parameter estimation (PPE)

10.6.10 Thoughts on Statistical Criteria for Goodness-of-Fit

Problems

References

**Chapter 11. Statistical Learning Through Data Analytics**

Abstract

Keywords

**Table of Contents**

11.1 Introduction

11.2 Distance as a Similarity Measure

11.3 Unsupervised Learning: Clustering Approaches

11.3.1 Types of clustering methods

11.3.2 Centroid-based partitional clustering by K-means

11.3.3 Density-based partitional clustering using DBSCAN

11.3.4 Agglomerative hierarchical clustering methods

11.4 Supervised Learning: Statistical-Based Classification Approaches

11.4.1 Different types of approaches

11.4.2 Distance-based classification: k Nearest Neighbors

11.4.3 Naïve Bayesian classification

11.4.4 Classical regression-based classification

11.4.5 Discriminant function analysis

11.4.6: Neural networks: Radial basis function (RBF)

11.4.7 Support vector machines (SVM)

11.5 Decision Tree-Based Classification Methods

11.5.1 Rule-based method and decision-tree representation

11.5.2 Criteria for tree splitting

11.5.3 Classification and regression trees (CART)

11.5.4 Ensemble method: Random forest

11.6 Anomaly Detection Methods

11.6.1 Introduction

11.6.2 Graphical and statistical methods

11.6.3 Model-based methods

11.6.4 Data Mining methods

11.7 Applications to Reducing Energy Use in Buildings

Problems

References

**Chapter 12. Decision-Making and Sustainability Assessments**

Abstract

Keywords

**Table of Contents**

12.1. Introduction

12.1.1. Types of decision-making problems and applications

12.1.2 Purview of reliability, risk analysis and decision-making

12.1.3. Example of discrete decision-making

12.1.4 Example of chiller FDD

12.2. Single Criterion Decision-Making

12.2.1. General framework

12.2.2. Representing problem structure: Influence diagrams and decision trees

12.2.3. Single and multi-stage decision problems

12.2.4. Value of perfect information

12.2.5 Different criteria for outcome evaluation

12.2.6 Discretizing probability distributions

12.2.7. Utility value functions for modeling risk attitudes

12.2.8. Monte Carlo simulation for first-order and nested uncertainties

12.3. Risk Analysis

12.3.1. The three aspects

12.3.2 The empirical approach

12.3.3 Context of environmental risk to humans

12.3.4 Other areas of application

12.4 Case Study: Risk Assessment of an Existing Building

12.5. Multi-criteria Decision-Making (MCDM) Methods

12.5.1 Introduction and description of terms

12.5.2 Classification of methods

12.5.3 Basic mathematical operations

12.6 Single Discipline MCDM Methods: Techno-economic Analysis

12.6.1 Review

12.6.2 Consistent attribute scales

12.6.3 Inconsistent attribute scales: Dominance and Pareto frontier

12.6.4 Case study of conflicting criteria: Supervisory control of an engineered system

12.7 Sustainability Assessments: MCDM With Multi-Discipline Attribute Scales

12.7.1 Definitions and scope

12.7.2 Indicators and metrics

12.7.3 Sustainability assessment frameworks

12.7.4 Examples of non-, semi- and fully-aggregated assessments

12.7.5 Two case studies: Structure-based and performance-based

12.7.6 Closure

Problems

References