Water Science and Technology Library

Shahab Araghinejad

Data-Driven Modeling: Using MATLAB® in Water Resources and Environmental Engineering

Springer

Shahab Araghinejad

Data-Driven Modeling: Using MATLAB® in Water Resources and Environmental Engineering

Springer

Shahab Araghinejad
College of Agriculture and Natural Resources
Irrigation and Reclamation Engineering
University of Tehran
Tehran, Iran

Additional material to this book can be downloaded from http://extras.springer.com

ISSN 0921-092X
ISBN 978-94-007-7505-3 ISBN 978-94-007-7506-0 (eBook)
DOI 10.1007/978-94-007-7506-0
Springer Dordrecht Heidelberg New York London

Library of Congress Control Number: 2013954990

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Dedicated to my wife Neda
and my daughter Roz

Preface

The purpose of writing this book has been to give a systematic account of major concepts and methodologies of data-driven models and to present a unified framework that makes the subject more accessible and applicable to researchers and practitioners. The book is structured to integrate important theories and applications on data-driven models and to use them in a wide range of problems in the field of water resources and environmental engineering. The presented models are useful for various applications, namely, hydrological forecasting, flood analysis, water quality monitoring, quantitative and qualitative modeling of water resources, regionalizing climatic data, and general function approximation. This book addresses the issue of data-driven modeling in two contexts. Theoretical background of the models and techniques is presented and discussed in a comparative manner, briefly. Also the source files of relative programs demonstrating how to use the explained models are presented with practical advice on how to advance them. The programs have been developed within the unified platform of MATLAB. The proposed models are applied in various illustrative examples as well as several workshops. The focus of the book remains a straightforward presentation of explained models by discussing in detail the necessary components and briefly touching on the more advanced components.

The book is served as a practical guide to the main audience of graduate students and researchers in water resources engineering, environmental engineering, agricultural engineering, and natural resources engineering. This book may also be adapted for use as a senior undergraduate and graduate textbook by selective choice of topics. Alternatively, it may also be used as a resource for practicing engineers, consulting engineers, and others involved in water resources and environmental engineering.

The book contains eight chapters; except first and last, each was developed in two parts of theory and practice to achieve the aim of the book.

Chapter 1 lays the foundation for the entire book with a brief review on different types of models that could be used for modeling water resources and environmental problems as well as the process of model selection for a specific problem. Furthermore, the general approach of using data-driven models is reviewed in this chapter.

Chapter 2 presents discrete and continues probability distribution functions. Since one of the most applicable fields of distribution functions is frequency analysis, dealing with this issue is also presented in this chapter. The hypothetical tests on the average and variance of one and two populations are reviewed in the chapter. Furthermore, two famous tests of *chi-square* and *Kolmogorov–Smirnov* are presented to decide on the best distribution function for a specific random variable. Each of the above calculations is supported by related commands provided in MATLAB.

Chapter 3 presents models for point and interval estimation of dependent variables using different regression methods. Multiple linear regression model, conventional nonlinear regression models, logistic regression model, and K-nearest neighbor nonparametric model are presented in different sections of this chapter. Each model is supported by related commands and programs provided in MATLAB.

Chapter 4 focuses on methods of time series analysis and modeling. Preprocessing of time series before being used through modeling containing assessment of different components is discussed in this chapter. Autoregressive (AR), autoregressive moving average (ARMA), autoregressive integrated moving average (ARIMA), and autoregressive moving average with eXogenous data (ARMAX) models are presented and discussed in this chapter. A review on the multivariate analysis of time series modeling is added. Two major applications of time series modeling, namely, forecasting and synthetic data generation, are presented, supported by the related syntaxes and programs in MATLAB.

Chapter 5 deals with the artificial neural networks (ANNs). It presents basic definition on the "Components of an ANN," "Training Algorithm," and "Mapping by ANNs." The chapter also deals with the introduction of famous ANN models, which are widely used in different fields. Theoretical background, network architecture, their training and simulation methods, as well as the codes necessary for applying the networks are presented in this section. After presenting each network sample, applications are discussed in different illustrative examples. In this chapter, static and dynamic networks and also statistical networks are those which are described and modeled in MATLAB.

Chapter 6 presents the concept of support vector machine (SVM) to analyze data and recognize patterns, required for classification and regression analysis. Two applications of SVMs including classification and regression are discussed in this chapter, and examples on using SVM for both mentioned purposes are presented. Examples and models of SVM are presented in MATLAB.

Chapter 7 is concerned with the fuzzy logic. Basic information in fuzzy logic, fuzzy clustering, fuzzy inference systems, and fuzzy regression are the main subjects presented in this chapter. Obviously, the related MATLAB commands are presented to support the models reviewed in this chapter.

Chapter 8 begins with a summary on the characteristics of the models presented in the previous chapters. The models are compared based on different criteria to give the readers ideas on how to take advantages of the models' strengths and avoid their weaknesses through the hybrid models and multi-model data fusion approach. The chapter continues with the examples of hybrid models and general techniques of multi-model data fusion.

The book also contains an appendix that helps readers to use MATLAB.

I would like to express my gratitude to my colleagues and students who helped me to complete and enhance this book. I very much thank Neda Parvini for providing artworks and graphics of the book and also her invaluable support and encouragement during this project. I would like to thank Shahrzad Farzinpak who has contributed effectively by reviewing the book. I greatly appreciate the patience and concern of Petra van Steenbergen, senior publishing editor, and Hermine Vloemans at Springer. The comments on different chapters of the book received from M. Moghaddas, H. Rousta, E. Meidany, and M. Farhangi are highly acknowledged.

Acknowledgement Constructive comments received from Mr. E. Bozorgzadeh are highly appreciated.

Tehran, Iran
2013

Shahab Araghinejad

Contents

Chapter 1
Introduction

Abstract Problems involving the process of water resources and environmental management such as simulation of natural events, warning of natural disasters, and impact analysis of development scenarios are of significant importance in case of the changing environment. Considering the complexity of natural phenomena as well as our limited knowledge of mathematical modeling, this might be a challenging problem. Recently, development of data-driven models has improved the application of specific tools to be used through the complex process of real-world modeling. Soft computing and statistical models are two common groups of data-driven models that could be employed to solve water resources and environmental problems. Data-driven models are among mathematical models, which use experimental data to analyze real-world phenomena. In contrast to physical models, they do not need a specific laboratory setup so are significantly cheaper. Also, in contrast to the analytical models, data-driven models can be used for the problems where we do not have enough knowledge about the intrinsic complexity of the phenomena. This chapter presents a brief review of different types of models that could be used for modeling water resources and environmental problems, reviews the process of model selection for a specific problem, and investigates the general approach of using data-driven models. The advanced stage of developing a model is discussed in the last section.

Keywords Data-driven models • Model selection • Type of models • Type of data • Decision support systems

1.1 Introduction

Modeling a system is one of the most significant challenges in the field of water resources and environmental engineering. That rise as a result of either physical complexity of a natural phenomenon or the time-consuming process of analyzing different components of a system. Data-driven models have been found as very

S. Araghinejad, *Data-Driven Modeling: Using MATLAB® in Water Resources and Environmental Engineering*, Water Science and Technology Library 67,
DOI 10.1007/978-94-007-7506-0_1, © Springer Science+Business Media Dordrecht 2014

powerful tools to help overcoming those challenges by presenting opportunities to build basic models from the observed patterns as well as accelerating the response of decision-makers in facing with the real-world problems. Since they are able to map causal factors and consequent outcomes of an event without the need for a deep understanding of the physical process surrounding the occurrence of an event, these models have become popular among water resources and environmental engineers. Also, as recent progresses in soft computing have enriched the collection of data-driven techniques by presenting new models as well as enhancing the classic ones, continuity of such popularity is expected.

Data-driven models, as it is understood by its name, refer to a wide range of models that simulate a system by the data experienced in the real life of that system. They include different categories generally divided into statistical and soft computing (also known as artificial-intelligent) models. Data-driven models are often inexpensive, accurate, precise, and more importantly flexible, which make them able to handle a wide range of real-world systems with different degrees of complexity based on our level of knowledge and understanding about a system. As far as the statistical type of them is concerned, these models could be considered among the very primary models in the life of modern engineering. However, they could be categorized as brand new models with regard to soft computing type. Data-driven modeling could be defined as a solution defined by the paradigm of "engineering thinking and judgment" to the world of modeling to deal with the problems, which are considered too complex by our knowledge of mathematical equations. Data-driven models have been brought up to their present form by the ideas and applications from different fields of engineering.

As complexity of a system increases, efficiency of offered models by data-driven methods rises in modeling the system. For systems with little complexity, analytical models based on mathematical equations provide precise descriptions, but for the ones with significant complexity, data-driven models are more useful to define the patterns within the behavior of the system.

Data-driven models can be applied in a wide range of problems including simulation of natural phenomena, synthetic data generation, forecasting and warning of extreme events, developing decision-making rules, and many others. Generally, the purpose of data-driven modeling includes but is not limited to:

- Data classification and clustering
- Function approximation
- Forecasting
- Data generation
- General simulation

As far as problems in the field of water resources and environmental engineering are concerned, application of data-driven models may include:

- Water quality simulation and prediction
- Extreme value prediction with emphasis on floods and droughts

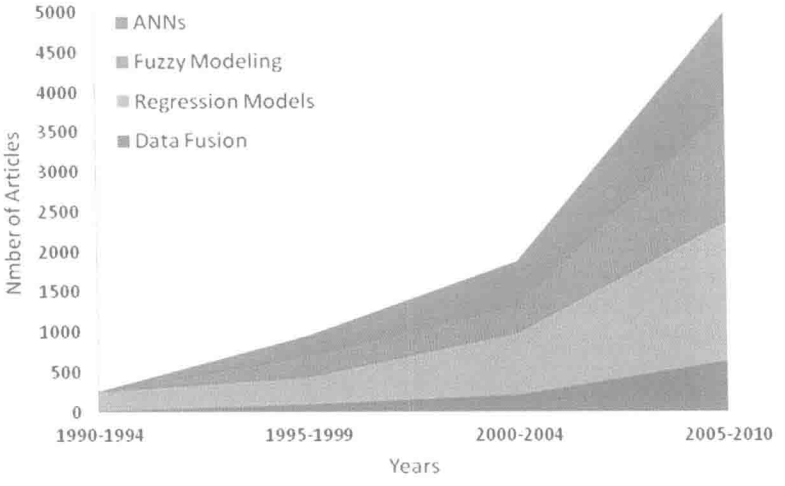

Fig. 1.1 Number of articles published in the selected data-driven techniques (Google Scholar)

- Modeling water balance concerning different components of a hydrological system
- Extending the length of hydroclimatological data from the historical ones
- Estimating censored data

The inexpensive process of developing data-driven models makes them a good choice as either the main tool for modeling a system or an alternative for the baseline model to be compared with the results obtained by analytical and physical models to validate them or to provide useful data and information to enhance them.

There has been an increasing interest on the data-driven modeling in the field of water resources and environmental engineering during the recent decade. Figure 1.1 demonstrates the trend of the number of articles published in the selected data-driven techniques including artificial neural networks (ANNs), fuzzy modeling, regression models, and data fusion in the field of water resources and environmental engineering in the period of 1990–2010. However, the presented statistic might not represent the actual number of researches in those fields; their relative changes demonstrate the fact of an increasing demand on the application of those models. The figure shows a considerable ascending trend in all classic and modern techniques.

This introductory chapter starts with a section discussing different types of models and their complexity. The chapter continues by presenting criteria to select a model. Despite the difference in the type of data-driven models, they all follow a general approach of modeling, which is presented in another section of this chapter. The chapter ends by a discussion on the next level of developing computing tools that could be imagined in the field of modeling and simulation.

1.2 Types of Models

The term "model" refers to a wide variety of programs, softwares, and tools used to represent a real-world system in order to predict its behavior and response to the changing factors approximately. In spite of the wide number, models are divided into two main categories: physical and mathematical models. Data-driven models are classified among the latter category. Figure 1.2 shows a proposed framework for different types of models, specifically those which could be used in the field of water resources and environmental field. As the figure depicts, a mathematical model is in turn divided into three types of data-driven, conceptual, and analytical models. The next expressions present a brief description of shown models in Fig. 1.2.

1.2.1 Physical Model

A physical model is a smaller or larger physical copy of a system. The geometry of the model and the object it represents are often similar in the sense that one is a rescaling of the other.

1.2.2 Mathematical Model

A mathematical model is based on mathematical logic and equations, which benefits from mathematical knowledge to simulate a system in an explicit or implicit manner. It is presented in the forms of analytical, conceptual, and data-driven models.

1.2.3 Analytical Model

An analytical model is a model that represents a system by mathematical equations explicitly. These models are applied in cases that are not too complex comparing to our knowledge of mathematics. Models developed for porous media and groundwater environment are examples of this kind of model.

1.2.4 Data-Driven Model

Data-driven models also known as *experimental models* refer to a kind of models which benefit input and output data of a system to find out specific patterns to be generalized for a broader range of data. Statistical models and artificial-intelligent models are two famous types classified in this category.

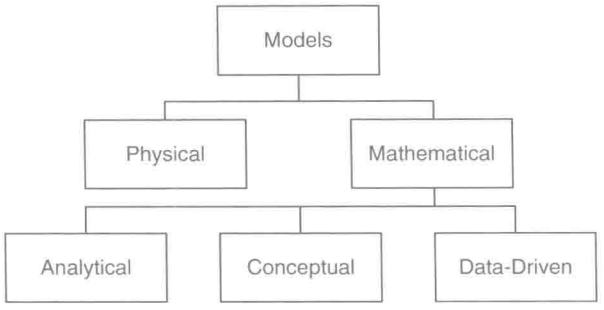

Fig. 1.2 A general classification of different types of models

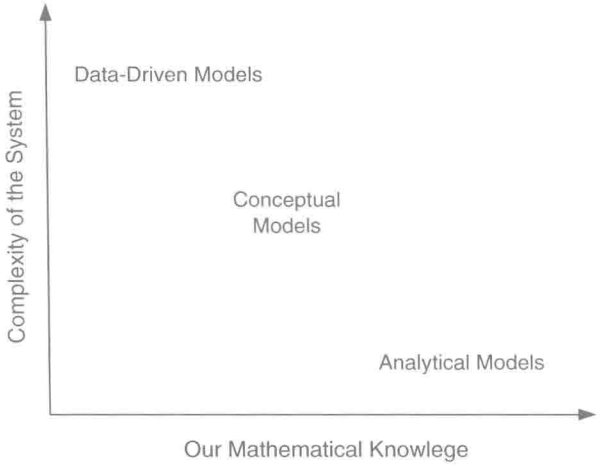

Fig. 1.3 Status of three types of mathematical models based on two characteristics of complexity and mathematical knowledge

1.2.5 Conceptual Model

A conceptual model in water resources and environmental engineering is a model that benefits from the physical definition of a system partially in cases that we do not have adequate knowledge of mathematical equation to represent the system by an analytical model. A conceptual model uses a combination of both analytical and data-driven model approaches in a way to employ the benefits of both to simulate a system.

Based on the presented definition, different types of mathematical models could be classified by the complexity of a system as well as our mathematical knowledge, as shown in Fig. 1.3.

Data-driven models as the focus of this book are divided into two general forms of statistical models and soft computing models defined as follows.

1.2.5.1 Statistical Model

The definition of a statistical model is tied to the definition of a stochastic process. A stochastic process includes both random and deterministic variables. Deterministic variables are dealt with mathematical models. Meanwhile, a random variable is represented by the theory of probability and a probabilistic model. A probabilistic model is a probability distribution function proposed as generating data. To simulate a stochastic process, statistical models, which benefit from both mathematics and probability theory, are used to model stochastic variables. These models could be parametric or nonparametric. The former has variable parameters, such as the mean and variance in a normal distribution. The latter has a distribution function without parameters, such as nearest neighborhood modeling, and only loosely confined by assumptions.

1.2.5.2 Soft Computing Model

Fuzzy logic, neuro-computing, and genetic algorithms may be viewed as the principal constituents of what might be called soft computing. Unlike the traditional hard computing, soft computing accommodates the imprecision of the real world by the rules obtained from the biological concepts.

1.3 Spatiotemporal Complexity of a Model

Regardless to the type of a model, it could be classified by its complexity in a spatial and temporal manner. The significance of dynamic change of natural phenomena related to water resources and global environment along with the need to asses them in a regional manner make these characteristics of great importance for a modeler. Even though, a fully developed dynamic and distributed model is considered as a very powerful tool in the field of water resources and environmental engineering, the limitation of data, the level of our expectation from the model, as well as the complicated and time-consuming process of modeling might prevent application of such a model in every problem we deal with. The following expressions define different types of models from the spatiotemporal complexity aspect.

1.3.1 Spatial Complexity

From the spatial complexity aspect, a model is categorized as lumped, distributed, and semi-distributed model. Variables in water resources and environmental engineering are usually defined within a specific spatial boundary such as watersheds, basins, provinces, and rivers. A lumped model is used in cases that the spatial variation of this variable within that boundary is not of interest of the modeler.

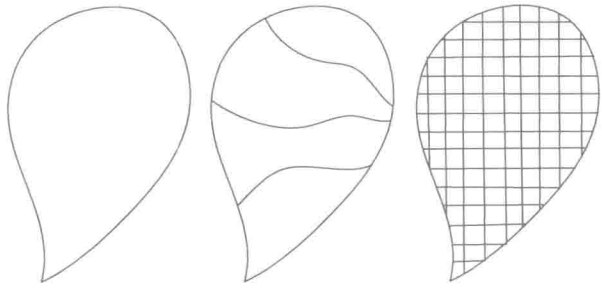

Fig. 1.4 (From *left*) Examples of lumped, semi-distributed, and distributed model

For instance, a model benefits from the average precipitation in a basin. In contrast, in some cases we need a model to be highly sensitive to the location where a variable is collected, generated, or developed through the process of modeling. This case needs a model to be considered as a distributed model within the boundary of the system. A model with the capability of interpolating/extrapolating of rainfall in each location of a watershed is considered as a distributed model. A semi-distributed model is a model which is not sensitive to each location of a region but considers variation within subdistricts of the region and acts like a lumped model within them. A hydrological model which gets different parameters for different subbasins of a river basin but follows the rules and equations of a lumped model within them is known as a semi-distributed model. Figure 1.4 shows schematically the difference between three lumped, distributed, and semi-distributed models. The figure shows the boundaries where the spatial variability of the model's parameters is concerned.

1.3.2 Temporal Complexity

Data-driven models can be either static or dynamic. Static simulation models are based on the current state of a system and assume a constant balance between parameters with no predicted changing. In contrast, dynamic simulation models rely on the detailed assumptions regarding changes in existing parameters by changing the state of a system. A rainfall forecasting model is considered as a dynamic model if its parameters change as it receives new precipitation data in its application time line. In contrast, it may be considered as a static model in case its parameters rely on the historical data with no plan to be changed in the time horizon of its application.

1.4 Model Selection

Model selection is the task of selecting a model from a set of candidate models via number of logical criteria. Undoubtedly among given candidate models of similar accuracy and precision, the simplest model is most likely to be chosen but still they need to be examined closely. Different criteria might be used in the process of selecting a model as follows.