

PAPER • **OPEN ACCESS**

Data-driven Modelling for decision making under uncertainty

To cite this article: Layla Angria S *et al* 2018 *IOP Conf. Ser.: Mater. Sci. Eng.* **300** 012013

View the [article online](#) for updates and enhancements.

You may also like

- [Preparation and application of high thermal conductivity TMBPDGE-DDM@h-BN composites](#)
Tengfei Qin, Hua Wang, Jing He *et al.*
- [The distance discordance metric—a novel approach to quantifying spatial uncertainties in intra- and inter-patient deformable image registration](#)
Ziad H Saleh, Aditya P Apte, Gregory C Sharp *et al.*
- [Estimating integrated measures of forage quality for herbivores by fusing optical and structural remote sensing data](#)
J S Jennewein, J U H Eitel, K Joly *et al.*



ECS The Electrochemical Society
Advancing solid state & electrochemical science & technology

247th ECS Meeting
Montréal, Canada
May 18-22, 2025
Palais des Congrès de Montréal

**Abstracts due
December
6th**

Showcase your science!

ECS UNITED

Data-driven Modelling for decision making under uncertainty

Layla Angria S¹, Yunita Dwi Sari², Muhammad Zarlis³, Tulus⁴

Department of Mathematics, University of Sumatera Utara, Medan, Indonesia

Email: ¹lyla_angria@yahoo.com, ²niitadwisari@gmail.com, ³m.zarlis@usu.ac.id,

⁴tulusjp@yahoo.com

Abstract. The rise of the issues with the uncertainty of decision making has become a very warm conversation in operation research. Many models have been presented, one of which is with data-driven modelling (DDM). The purpose of this paper is to extract and recognize patterns in data, and find the best model in decision-making problem under uncertainty by using data-driven modeling approach with linear programming, linear and nonlinear differential equation, bayesian approach. Model criteria tested to determine the smallest error, and it will be the best model that can be used.

1. Introduction

One of the important things in a life is the existence of a decision created by the decision maker. Simply put, in everyday life we often make decisions based on the things that support them decision-making is based on available information. Information obtained is usually in the form of data.

Soumendra Mohanty, et. al (2013) In general, data can be defined into 3 important characteristics: composition, context, and conditions. Composition refers to the data structure: what is the source, what type of data, what is the nature of the data (most static data or real time stream data) and so on. Context refers to how data is generated, what events are related to the data, how the data sensitivity and others. Conditions refer to the state of the data and whether the data can be used for analysis or need further cleaning and enrichment.

The reality is that according to Dimtris Bertsimas and Aurelie Thiele (2006) the information or data available is incomplete and strategies based on wrong inputs may not be easy to do, or produce poor results when implemented. So that will affect the decision-making.

Based on what is presented above is the necessity of decision making appropriately. Derek W. Bunn (1984) Decision-making is an ubiquitous activity attached to an individual, social group, or group of organizations. In the world of decision-making industry is very important such as in deciding how much consumer demand for a product, transportation problems, water resources and more.

But in decision making there are many parameter uncertainty. Raiffa and Schlaifer (1961) Decision-making with uncertainty is the main arena in the performance of decision-making methodology. Indeed, the main function of decision analysis, as it was originally developed, is to



assist in the evaluation of decision-making for options whose ultimate outcome is not known for certain.

Therefore, as is the uncertainty parameters that affecting the decision, must use appropriate models to determine what decisions we will gain and reduce the risk to be generated. The model to be discussed in this issue is the data-driven model. Mosaad Khadr and Mohamed Elshemy (2016) argue that data-driven models is based on analyzing the data characterizing, where a model is built on the basis of finding connections between the system state variables such as input, internal and output variables with only a limited assumption about the system.

2. Essence of data – driven modelling

Solomatine et al. (2008) a data-driven model is based on the analysis of the data about a specific system. The main concept of data-driven model is to find relationships between the system state variables (input and output) without explicit knowledge of the physical behavior of the system. Task of data-driven modelling is classification: where the task constitutes of assigning a class for an input data point. Association: where association between variables characterising the system is to be identified, which is used in subsequent prediction. Regression: where the task constitutes of predicting a real value associated with an input data point. Clustering: where groups of data points with within group similarity are to be determined.

The process of building a data driven modelling follows general principles adopted in modelling: study the problem – collect data – select model structure – build the model – test the model and (possibly) iterate. In data-driven modelling, not only the model parameters, but also the models structure, is often subject to optimisation.

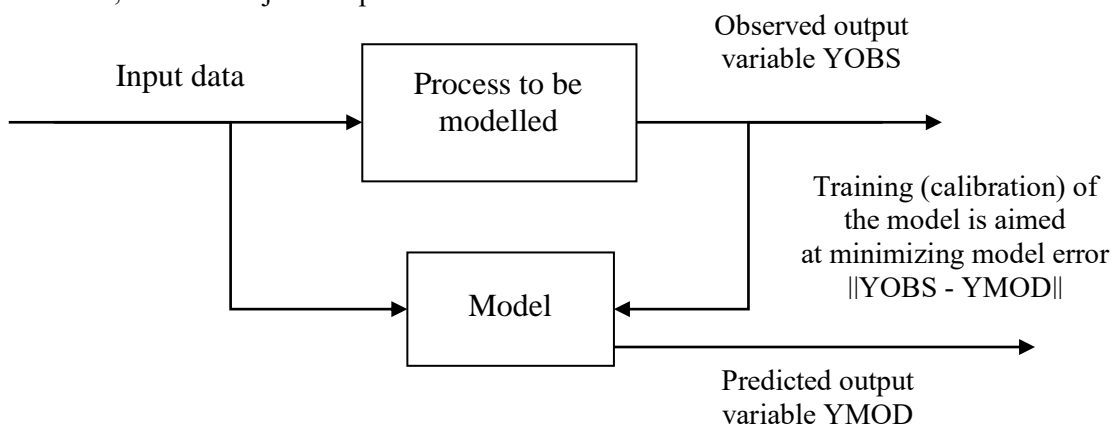


Figure 1. General approach to modelling

a. Regression and Classification

In general two major areas where engineering analysis is required are modeling or simulation of a given occurring phenomena (regression analysis), and identification or categorization of occurring events or pattern (classification analysis). These two major areas come with their unique characteristics in term of objectives to achieve, data sources, and post-calibration implementation, requiring adequate approaches for a successful model development.

b. Regression Analysis

The regression analysis includes any technique or algorithm used for mapping or linking several variables among them. The focus is to develop a relationship among one or several dependent variables (outputs) with one or more independent variables (inputs).

c. *Clasification Analysis*

The classification analysis is related to the categorization or labeling of a given group of variables values into a pre-defined possible list of groups (in some cases the pre-defined groups is not necessary or available). The variables in this case are not necessarily defined as input – outputs groups like in regression analysis.

3. Models and Method

a. *Data fitting and model selection*

The data that have been obtained are selected according to their respective characters. Data included in linear / nonlinear, stochastic or deterministic. Then modeled with linear / nonlinear equations, arima, ANFIS. Linear equation can be described by the following model:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \quad (2.1)$$

the curve is represented by the quadratic curve.

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \dots + \beta_k X^k + \varepsilon \quad (2.2)$$

Where $\beta_0, \beta_1, \dots, \beta_k$ are fixed regression parameters and ε is a random error or noise parameter.

Luthkepohl (2005) Time series is basically a measurement data taken chronologically within a certain time. ARMA is a time series forecasting technique. Sadeq (2008) ARMA model can be broken down into AR model (*Autoregressive*) and MA model (*Moving Average*).

$$y_t = b_0 + b_1 y_{t-1} + b_2 y_{t-2} + \dots + b_n y_{t-n} + \varepsilon_t \quad (2.3)$$

MA Model (*Moving Average*)

$$y_t = a_0 + a_1 \varepsilon_{t-1} + a_2 \varepsilon_{t-2} + \dots + a_n \varepsilon_{t-n} + \varepsilon_t \quad (2.4)$$

Where:

y_t : time series

$y_{t-1}, y_{t-2}, \dots, y_{t-n}$: previous time series

$\varepsilon_{t-1}, \varepsilon_{t-2}, \dots, \varepsilon_{t-n}, \varepsilon_t$ residual

$a_0, a_1, a_2, \dots, a_n$ constants dan coeffisient MA model

$b_0, b_1, b_2, \dots, b_n$ constants dan coeffisient RA model

Mosaad and Elshemy (2016) An adaptive neuro-fuzzy inference system (ANFIS) is a fuzzy inference system formulated as a feed-forward neural network. Hence, the advantages of a fuzzy system can be combined with a learning algorithm. ANFIS was introduced as an effective tool to represent simple and highly complex functions more powerfully than conventional statistical methods

b. Model and parameter identification

In the current work, developing model using an appropriate approach e.g with a bayesian approach for both model and parameters of the models were identified.

c. Model selection criteria

1. Normalised root mean squared error (NRMSE) was used to compare the models with different orders for each individual time series. Inferred parameters θ_i were applied back to the system equation to generate time series without stochastic terms, i.e deterministic solution. Note that because parameters were identified in the form of normal distribution, the most probable (mean) value of parameters were plugged into the system equation. Root mean squared error (RMSE) between the measurement time series y_t and the inferred deterministic time series \hat{y}_t can be calculated as follows:

$$RMSE = \sqrt{\frac{\sum_{t=1}^n (\hat{y}_t - y_t)^2}{n}}$$

2. *Generic form* . As the entire aim was to find a model capable of capturing the common patterns in all time series, a model that could only describe some time series was disregarded.
3. *System stability* . The model has to have a unique stable steady state, which implies that the system's response de- cays with time. This has been checked via calculations of the real parts of corresponding eigenvalues which have to be negative for stability

d. Statistical analysis

The null hypothesis of no difference between the groups of interest was tested at the 5% significance level, and the results are presented as p -values. Statistical analysis of the differences in *NRMSE* between two models was performed using the one sample t -test, which assesses the normality of data with zero mean and unknown variance at the 5% significance level. The result is presented as p - values.

4. Application

4.1 start with the first case where the decision-maker must decide how many items to order at each time period at a single store. (In mathematical terms, the state of the system can be described as a scalar variable, specically, the amount of inventory in the store.) We use the following notation:

$$x_{t+1} = x_t + u_t - w_t \quad 4.1$$

which yields the closed-form formula:

$$x_{t+1} = x_0 + \sum(u_t - w_t) \quad 4.2$$

x_t : inventory at the beginning of time period t ,
 u_t : amount ordered at the beginning of time period t ,
 w_t : demand occurring during time period t .

Demand is backlogged over time, and orders made at the beginning of a time period arrive at the end of that same period. Therefore, the dynamics of the system can be described as a *linear* equation:

4.2 In the problem of antibody transplant Yan Zhang et. Al (2016) the data-driven model is used to determine the best model. The first model that can considered is a nonlinear differential equation with coefficients in the form of a polinomial function. Model can be described by the following:

$$\frac{d^n}{dt^n} x_t + \sum_{i=0}^{n-1} f_{i-1}(x_t) \frac{d^i}{dt^i} x_t + f_0(x_t) = \eta_t \quad 4.3$$

$$y_t = x_t + \varepsilon_t \quad 4.4$$

Linear models with different system orders are considered first. Eq. (1) in Section 3.1.2 transforms into a linear differential equation when $f_{i+1}(x_t)$ are constants, with constant parameters θ_i ($i = 0, 1, \dots, n-1$) where $f_n(x_t) = \theta_n, \dots, f_2(x_t) = \theta_2, f_1(x_t) = \theta_1, f_0(x_t)$ is defined as $-\theta_0$ for convenience. A first order linear model was considered first, and it did not show good performance. Then linear models with higher system orders were investigated. In this section, we present the results of system and parameter identification by comparing solutions for linear first, second and third order dynamic equations only:

$$\text{Model 1 } (M_1) : \frac{dx_t}{dt} + \theta_1 - \theta_0 = 0 \quad 4.5$$

$$\text{Model 2 } (M_2) : \frac{d^2x_t}{dt^2} + \theta_1 \frac{dx_t}{dt} + \theta_1 x_t - \theta_0 = 0 \quad 4.6$$

$$\text{Model 3 } (M_3) : \frac{d^3x_t}{dt^3} + \theta_3 \frac{d^2x_t}{dt^2} + \theta_2 \frac{dx_t}{dt} + \theta_1 x_t - \theta_0 = 0 \quad 4.7$$

Two nonlinear models were considered: model *NM1* with nonlinearity in the damping term ($f_1(x_t) = k_1(x_t) + \theta_1, f_2(x_t) = k_2(x_t) + \theta_2$) and model *NM2* with nonlinearity in the x_t term ($f_1(x_t) = \theta_1, f_2(x_t) = k_2 x_t + \theta_2$). The corresponding system equations are as follows:

$$NM_1 : \frac{d^2x_t}{dt^2} + \theta_2 \frac{dx_t}{dt} + (k_1 x_t + \theta_1) x_t - \theta_0 = 0 \quad 4.8$$

$$NM_1 : \frac{d^2x_t}{dt^2} + (k_2 x_t + \theta_2) \frac{dx_t}{dt} + \theta_1 x_t - \theta_0 = 0 \quad 4.9$$

Table 1

	NRMSE		
	M_1	M_2	M_3
(a)	0.272	0.053	0.014
(b)	0.083	0.085	0.013
(c)	0.090	0.096	0.053
(d)	0.088	0.071	0.073

4.3 Problem of river flow prediction. In this problem Joko suryanto (2016) Consider 3 models of ARIMA, ANFIS and FFNN. From three models obtained the best model is FFNN (feed forward neural network) with the smallest error value, by follow equation:

$$y_i = f_2 \left[\sum_{k=1}^k w_{1k} f_2 \left(\sum_{j=1}^j w_{kj} x_j + b_k \right) + b_1 \right] \quad 4.10$$

$$f_2(p) = \frac{2}{(1+e^{-2p})} - 1 \quad 4.11$$

5. Conclusion

In making decisions with uncertainty, we should be able to classify the data obtained in order to build a suitable model so as not to cause too much risk. The data obtained must also be in accordance with the objectives to be achieved by the needs maker, so that in making the model does not occur a big error and has biggest correlation. E.g in second application problem, the best model is obtained by looking at the smallest error by using a linear model $\frac{d^2x_t}{dt^2} + (k_2x_t + \theta_2)\frac{dx_t}{dt} + \theta_1x_t - \theta_0 = 0$. In the third problem the best model is FFNN by looking at the smallest error and biggest correlation, with error 8,422 and correlation 0,85.

References

- [1] Ackoff, R. C. dan F. E. Emery. (1972). *On purposeful system*. Aldine-Atherton. Chicago
- [2] Bertsimas, Dimtris. dan Thiele, Aurelie. (2006). *Robust and data-driven optimization: modern decision making under uncertainty*. cambridge
- [3] Bunn, W, Derek. (1984). *Applied decision analysis*. USA
- [4] Bertsimas, Dimtris. dan Thiele, Aurelie. (2006). *A Robust Optimization Approach to Inventory Theory*. *Operation Research*. 54(1):150-168
- [5] Bertsimas dan M. Sim. (2004). *The Price of Robustness*. Operation Research
- [6] Elshemy, Mohamed. dan Khadr, Mosaad. (2016). *Data-driven modeling for water quality prediction case study: the drains system associated with manzala lake, egypt*. Tanta university
- [7] Femandes, Betina. Street Alexandre, dkk. (2016). *An adaptive robust portofolio optimization model with loss constraints based on data-driven polihedral uncertainty sets*. Pontifical catholic university of rio de
- [8] J. Daunizeau, KJ. Friston, SJ. Kiebel. (2009). *Variational bayesian identification and prediction of stochastic nonlinear dynamic causal models*, *phys. D nonlinear Phenom*. 2089-2118
- [9] Meng, Chao. Nageshwaranier, Sal Srinivas, dkk. (2013). *Data-driven modeling and simulation frame-work for material handling systems in coal mines*. University of arizona, 766-779
- [10] M.J. Beal. (2003). *Varitional Algorithms for Approximate Bayesian inference* (Ph.D. Thesis), The Gatsby Computational Neuroscience Unit. University College London
- [11] Mohanty, Soumendra. Harsa Srivatsa, dkk. (2013). *Big data imperatives*. India
- [12] Simon, H. A. (1957). *Models of man*. New York
- [13] Zhang, Yan. David Briggs, dkk. (2016). *A new data-driven model for post-trasplant antibody dynamics in high risk kidney transplantation*. University of warwick, UK
- [14] Solomatine, DP, dan Ostfeld. (2008). *A Data-driven Modelling: Some Past Experience and New Approaches*. *J of Hydroinformatics*. 3-22
- [15] Supranto. (2009). *Teknik Pengambilan Keputusan*. Rineka Cipta. Jakarta
- [16] Suryanto, Joko. (2016). *Aplication data-driven teknik for monthly river flow prediction*. Unsyiah. Indonesia