Integrating Mechanistic Modelling and Statistical Learning Theoretic-Methods: initial steps toward a strategy for model evaluation and structure selection

A. Guergachi

School of Information Technology Management, Ryerson University, Toronto, Canada

(a2guerga@ryerson.ca or aguergachi@hotmail.com)

Abstract: The traditional model validation procedure has been criticized in many areas of environmental engineering. In this paper, it is suggested to replace it by a more reasonable procedure: "model evaluation". A framework for model evaluation and structure selection (FMESS) is presented. Based on statistical learning theory, this framework considers the model identification step as a learning problem. The model evaluation process is based on one single criterion: model performance. The latter is measured by a mathematical deviation between the model prediction and reality. Although it is an exact measure of model performance, this deviation cannot be computed, but can be related to the empirical measure that system modellers have traditionally used in the steps of model identification and validation. The relationship between the exact and empirical measures is called an uncertainty model. Two uncertainty models are presented in this paper. For these models to be valid, a set of conditions with regard to the system uncertainty needs to be satisfied. Although quite weak, these conditions may not always hold true in the case of complex environmental systems. Mechanistic information on the system's dynamic processes would be required to ensure the fulfilment of these conditions.

Keywords: Statistical learning theory, Mechanistic modelling, uncertainty models, model performance, exact and empirical measures of model performance.

1. INTRODUCTION

Mathematical modelling of environmental systems is traditionally carried out in three sequential steps:

- *model structure development*: the modeller collects the available knowledge about the studied system in the form of first principles and empirical laws. Based on this knowledge, the modeller develops a set of mathematical relationships (i.e., the system model M) among the system state variables, which can generally be written in the form of a differential equation:

$$\dot{\mathbf{x}} = \mathbf{f}(t, \mathbf{x}, \mathbf{p}) \qquad (1)$$

  where $\mathbf{x}$ is the system state vector, $\mathbf{p}$ is the model parameter vector, t is the time and $\mathbf{f}$ is a vector function generally non-linear.
- *model identification*: after the model is developed, the modeller uses a set Y_N (N being a natural number) of empirical data:

$$\mathrm{Y}_N : \quad \mathbf{x}^{data}(t_1), \mathbf{x}^{data}(t_2), \ldots, \mathbf{x}^{data}(t_N)$$

  collected from the real operation of the system, to identify the model parameters. This step usually requires the minimization of an objective function $J(\mathbf{p})$ of the form:

$$J(\mathbf{p}) = \sum_{k=1}^{N} \left\| \mathbf{x}(\mathbf{p}, t_k) - \mathbf{x}^{data}(t_k) \right\|^2$$

  where $\mathbf{x}(\mathbf{p}, t_k)$ represents the solution to the model equation (1). In most cases, the data set Y_N would actually be divided into two subsets Y_{N_1} and Y_{N_2} ($N = N_1 + N_2$). The first subset (called identification sample) is used for the model parameter vector identification, and the second (called validation sample) for model validation (step below).
- *model validation*: in this step, the identified system model is tested on the validation subset Y_{N_2} that it has never "seen". If the model performs well on this sample, then it is retained. Otherwise, the model structure is

adjusted and the validation procedure repeated.

The model validation step has been criticized in many areas of environmental engineering [Jeppsson, 1996; Zheng and Bennett, 1995]. The criticisms are generally due to the highly complex nature of the environmental systems that modellers deal with nowadays, which makes the traditional modelling approach (TMA), outlined above, simply inadequate for developing useful descriptions of the behaviour of such systems. The appellation "model validation" has been considered to convey a false sense of truth and accuracy [Konikow and Bredehoeft, 1992]. It implies that the developed (and identified) model structure is already an accurate representation of reality, and the goal of the third step of the TMA is merely to confirm the quality of the model (the Oxford English dictionary defines the word "validate" as a synonym of "ratify", "confirm"). This situation would almost never occur in the case of a complex system, as in general we have no a priori proof that the developed mathematical model structure reflects all the mechanisms that govern the system dynamics in reality. Because of this, the purpose of the third step of the TMA should be to evaluate the performance of the developed model and compare it to that of other models. We shall name this step the "model evaluation" step.

The aim of this paper is to present a strategy for carrying out this step. This strategy is based on a new mathematically-based framework developed by the author for model evaluation and structure selection [Guergachi, 2000]. The development of this framework is based on Vapnik's work in the area of Statistical Learning Theory (SLT) [Vapnik, 1998].

This paper presents a qualitative overview of the framework. The basic ideas of the framework are discussed in the next section. The third section develops the concepts of 'exact' and 'empirical' measures of model performance, while the fourth section presents two expressions of uncertainty models that relate the exact measure to the empirical one. The fifth section discusses the applicability of the framework, and briefly describes a new paradigm for system modelling. The mathematics that underlies the framework is not discussed in this paper, but can be found in Guergachi [2000] and in Guergachi and Patry [2002b].

2. BASIC IDEAS OF THE FRAMEWORK FOR MODEL EVALUATION AND STRUCTURE SELECTION (FMESS)

Any mechanistic model M of a complex system S is necessarily an imperfect representation of reality (this assertion may be used as a definition of the expression "*complex system*"). Jeppsson [1996] has very rightly expressed this fact for the case of the activated sludge wastewater treatment process: "*Though available models are quite complex they are still greatly simplifying the representation of many species of organisms. As the microbial population changes this needs to be reflected in changing kinetic parameters and even adding new state variables*". In the FMESS, a mechanistic model is viewed as a learning machine that *tries* to acquire more information about the system behaviour (and, therefore, improves the model representation of reality) from the data set Y_N. The traditional model identification is the procedure that will be used to carry out the task of teaching or, more accurately (remember, the model is mechanistic, so it already contains some amount of information about the system mechanisms), improving the "model's knowledge" about the system. Model identification is, therefore, viewed as a learning problem or, equivalently, an information transfer from the data set Y_N to the model M. Depending on its complexity, the model M may or may not be able to 'absorb' all the information contained in the data set Y_N, during the identification procedure. The FMESS considers model M as an information container. The size of this container is characterized by the model complexity and is measured by a scalar number q, called Vapnik-Chervonenkis (VC) dimension [Vapnik, 1998; Guergachi, 2000; Guergachi and Patry, 2002b, 2002c]. This dimension characterizes the two aspects where model complexity is reflected: (1) the number of model parameters *and* (2) the nature of the model equations' non-linearities. The amount of information that is contained in the data set Y_N is, on the other hand, evaluated from a statistical point of view. It is characterized by both the number N of elements in the data set and the degree of statistical dependency that exists among these elements. The results presented in this paper consider only one particular case of statistical dependency: it is the *independent case*. As a result, the amount of information contained in the data set Y_N is measured by the number N only. At the end of the learning phase (or, using the traditional modelling jargon, the model identification phase), the model performance needs to be evaluated. In

the FMESS, the evaluation of models is based on one single criterion: model performance. Issues such as model identifiability, verifiability and reduction are not taken into account (directly) in the model evaluation and selection process. Model performance is measured by a mathematical deviation R between the model prediction and reality over a very long period of time that includes both the 'past' and the 'future'. Because the 'reality' is never completely known (particularly in the case of future time instants), the exact value of the deviation R cannot be computed. In the FMESS, we don't strive to compute R; we look for a (mathematically-proven) upper bound φ on the value of R. This upper bound φ is a function of the dimension q and the number N, among other variables (see discussion below). The variables on which φ is dependent are called the *model performance control variables* (MPCV), the function φ is called a *guarantee on the model performance*, and the relationship:

$$R \le \varphi \qquad (1)$$

between R and φ is called an *uncertainty model*. In the FMESS, the MPCVs and the function φ are readily determinable/computable, which renders an inequality of the type (1) very useful for environmental system modellers in many respects:

1. *For model evaluation*: when a set of values of q and N is fixed (say, for example $q = q_0$ and $N = N_0$), the inequality (1) will allow the modeller to compute an upper bound φ_0 (φ_0 being the value of φ for $q = q_0$ and $N = N_0$) on the deviation R. The modeller obtains then a quantitative guarantee on the system model quality.

2. *For model performance improvement*: when the model complexity (i.e., the dimension q) is fixed (say, for example, $q = q_0$), the inequality (1) will allow the modeller to assess the rate of model performance improvement, as the amount N of data used for model training (or identification) increases. This assessment can be done by computing the partial derivative $(\partial\varphi/\partial N)_{q_0}$ of the bound φ.

3. *For model structure selection*: when the amount of data N is fixed (say, for example, $N = N_0$), the inequality (1) will allow the modeller to select the optimal model structure complexity q_{opt}, which represents the value of q at which the upper bound function φ attains its minimum, N being set equal to N_0.

3. EMPIRICAL VERSUS EXACT MEASURE OF MODEL PERFORMANCE:

Let M be a model of the system S and suppose we are interested in the model predictions of the i_0-th state variable x_{i_0} of S. The traditional model identification and validation procedures rely exclusively on the finite-sum-based objective function:

$$J_{i_0}(\mathbf{p}) = \sum_{k=1}^{N} \left| x_{i_0}(\mathbf{p}, t_k) - x_{i_0}^{data}(t_k) \right|^2$$

or, equivalently, the arithmetic mean:

$$R_{emp}(\mathbf{p}) = \frac{1}{N} \sum_{k=1}^{N} \left| x_{i_0}(\mathbf{p}, t_k) - x_{i_0}^{data}(t_k) \right|^2 \qquad (2)$$

to evaluate the model performance [the subscript $_{emp}$ refers to "*empirical*"; $\mathbf{p}$ is the model parameter vector; $x_{i_0}(\mathbf{p}, t_k)$ is the model prediction for the variable x_{i_0} at time t_k]. The expression of this function is restricted to the examples $x_{i_0}^{data}(t_1), x_{i_0}^{data}(t_2), ..., x_{i_0}^{data}(t_N)$ that the model M had ''seen'' in the course of its development. The user, however, expects the model to produce good predictions not only for the situations that it had seen before, but also for the other unseen situations that will occur in the *future* real-world operation of the system. A more realistic and accurate measure of the model performance should then be a function of the type:

$$J_{i_0}^{P\&F}(\mathbf{p}) = \sum_{k=1}^{N} \left| x_{i_0}(\mathbf{p}, t_k) - x_{i_0}^{data}(t_k) \right|^2 + \left| x_{i_0}(\mathbf{p}, t_{N+1}) - x_{i_0}^{data}(t_{N+1}) \right|^2 + \left| x_{i_0}(\mathbf{p}, t_{N+2}) - x_{i_0}^{data}(t_{N+2}) \right|^2 + ... + \left| x_{i_0}(\mathbf{p}, t_h) - x_{i_0}^{data}(t_h) \right|^2 + ...$$

(the superscript $^{P\&F}$ refers to "*Past & Future*") or, equivalently, the expected average:

$$R(\mathbf{p}) = \frac{1}{n} \sum_{k=1}^{n} \left| x_{i_0}(\mathbf{p}, t_k) - x_{i_0}^{data}(t_k) \right|^2 \qquad (3)$$

where $n > N$ is a very large number (the time sequence t_h can be defined by $t_h = h\,\Delta t$, where Δt is a fixed time step). This function takes into account the deviation of the model predictions

from the *past* system data, as well as from the *future* system response and, as such, it is an exact measure of model performance. System modellers must aim to minimize $R(\mathbf{p})$, not $R_{emp}(\mathbf{p})$. The problem, however, is that the function $R(\mathbf{p})$ cannot be computed, for the obvious reason that the future system response $x_{i_0}^{data}(t_h)$ is not known. Therefore, we end up with the following situation:

- The function $R_{emp}(\mathbf{p})$ is merely an empirical measure of the model performance, but its numerical value is accessible to us.
- The function $R(\mathbf{p})$ is an exact measure of the model performance, but its value is inaccessible to us.

The whole issue of model evaluation is, therefore, how to infer (if it is possible) information about the exact measure $R(\mathbf{p})$ from the knowledge of the value of the empirical measure $R_{emp}(\mathbf{p})$. Addressing this issue will be done in the FMESS by developing uncertainty models of the type of inequality (1). The upper bound function φ in these models is dependent on $R_{emp}(\mathbf{p})$, in addition to the two aforementioned MPCVs, q and N, and other variables that will be introduced below. Thus, inequality (1) can be re-written in a form that shows the MPCVs introduced up until now:

$$R(\mathbf{p}) \le \varphi(q, N, R_{emp}(\mathbf{p})) \qquad (4)$$

The function $R_{emp}(\mathbf{p})$ is called *empirical risk*. It has been linked to the so-called "information conversion efficiency function" [Guergachi and Patry, 2002a]. The function $R(\mathbf{p})$ is called *expected risk*; note that its expression (3) can be re-written as:

$$R(\mathbf{p}) = \lim_{n \to +\infty} \frac{1}{n} \sum_{k=1}^{n} \left| x_{i_0}(\mathbf{p}, t_k) - x_{i_0}^{data}(t_k) \right|^2$$

The general method of inferring information about $R(\mathbf{p})$ from the knowledge of $R_{emp}(\mathbf{p})$ is called the *Inductive Principle of Empirical Risk Minimization*.

4. EXPRESSIONS OF THE UNCERTAINTY MODELS:

The FMESS uses the random variable paradigm to account for the uncertainty that underlies the behaviour of the complex system S and to develop inequalities of the type (1). It considers $x_{i_0}^{data}(t_h)$ as a random variable that arises according to some probability density function (pdf) P. Several modelling technologies, such as time series analysis and stochastic differential equations, have used the random variable paradigm to account for uncertainty, but have assumed, in most cases, that the type of the pdf P is known. In many situations, indeed, researchers as well as practitioners have taken for granted the fact that the pdf of the random variable under study is of the normal or uniform type. This is a very strong assumption about the nature of the system uncertainty, and there is generally no way of justifying it. The FMESS tries to assume the least amount of prior information about the pdf P. In particular, it does not assume anything about the distribution type of P, which can, a priori, be normal, uniform, exponential or any other density function. The function P is considered *unknown* in the FMESS. It has been shown, however, that without some additional information about the deviation:

$$\delta(t_k, \mathbf{p}) = \left| x_{i_0}(\mathbf{p}, t_k) - x_{i_0}^{data}(t_k) \right|^2$$

it is impossible to obtain any inequality of the type (1) [Vapnik, 1998]. As a result, using the work of Vapnik, the FMESS considers two conditions that are viewed by machine learning theorists as quite weak:

- *Condition C.1*: the deviation $\delta(t_k, \mathbf{p})$ is bounded, i.e., it is possible to find a positive number M such that:
$$\delta(t_k, \mathbf{p}) \le M \qquad (5)$$
- *Condition C.2*: the ratio:
$$\frac{\left(\mathbf{E}[\delta^s(t_k, \mathbf{p})]\right)^{1/s}}{\mathbf{E}[\delta(t_k, \mathbf{p})]}$$
is bounded ($\mathbf{E}[X]$ represents the expectation of the random variable X, and s is any number greater than 2) i.e., there exists a positive number τ such that:
$$\frac{\left(\mathbf{E}[\delta^s(t_k, \mathbf{p})]\right)^{1/s}}{\mathbf{E}[\delta(t_k, \mathbf{p})]} < \tau \qquad (6)$$

The condition *C.1* is simpler, but more stringent, while the condition *C.2* looks complex, but it is a lot weaker than *C.1*. Both conditions, however, are a lot weaker than assuming that the pdf P is known. A detailed discussion of the mathematics, as well as the qualitative meaning of these conditions can be found in Guergachi [2000] and in Guergachi and Patry [2002c].

Based on the work of Vapnik [1998], the following uncertainty models (7) and (8) can be obtained [Guegachi, 2000; Guergachi and Patry, 2002b]:

- In case the condition *C.1* holds true, we get:

$$R(\mathbf{p}) \le R_{emp}(\mathbf{p}) + \frac{M\zeta}{2}\left(1+\sqrt{1+\frac{4R_{emp}(\mathbf{p})}{M\zeta}}\right) \quad (7)$$

- In case the condition *C.2* holds true, we get:

$$R(\mathbf{p}) \le \frac{R_{emp}(\mathbf{p})}{(1-\gamma(s)\ \tau\sqrt{\zeta})_+} \quad (8)$$

The objects used in the inequalities (7) and (8) are as follows:

- The number ζ is:

$$\zeta = 4\frac{\left[q\left(\ln\left(\frac{2N}{q}\right)+1\right)-\ln\left(\frac{\eta}{4}\right)\right]}{N}$$

- The inequality (7) holds true with a probability of at least $1-\eta$, where the number η is within the interval]0,1[. In general, η is set equal to 0.05 so that $1-\eta = 0.95$. Vapnik [1998, page 523] suggested to make η inversely proportional to the square root of N. These comments apply to the inequality (8) as well.
- The numbers M, s and τ are as specified in the conditions (5) and (6).
- $\gamma(s) = \sqrt[s]{\frac{1}{2}\left(\frac{s-1}{s-2}\right)^{s-1}}$ and $(a)_+ = \max(a,0)$

5. APPLICABILITY OF THE FMESS AND PROPOSAL OF A NEW PARADIGM FOR SYSTEM MODELLING

The inequalities (7) and (8) represent the most significant results of the FMESS. They allow the modeller to evaluate the model performance and to select the system's optimal model structure. They have been rigorously proven in mathematical statistics and, therefore, the modeller should not have any doubt about their validity. However, like almost any result in science and engineering, there is a set of conditions $\{C.i\}$ that need to be satisfied, in order for the inequalities (7) and (8) to hold true. The modeller needs to understand these conditions and be aware of them, whenever the uncertainty models (7) and (8) are used. The implementation of these uncertainty models and the conclusions that are drawn from them must take into account the degree of truthfulness of the conditions $\{C.i\}$. For a complete discussion of the nature of these conditions and how they affect the validity of the uncertainty models (7) and (8), the reader is referred to Guergachi [2000] and Guergachi and Patry [2002c]. In what follows, we provide the list of these conditions:

- The *Condition C.1* and/or *Condition C.2* is/are satisfied (see above section).
- *Condition C.*3: the system data are independent and identically distributed (*i.i.d.*), and arise according to *one fixed* pdf P.
- *Condition C.*4: the VC dimension q of the model M is finite.

It should be noted that these conditions are quite weak and, therefore, they tend to hold true for a wide range of situations. Pattern recognition (e.g., hand-written postal code recognition) is one of the most popular of these situations. In the case of a continuous system that is dynamic, stochastic and highly non-linear, such as the biological wastewater treatment systems or the climate systems for example, the fulfilment of these conditions is not as obvious. It is not guaranteed that the system data would keep arising according to *one* fixed pdf, in an *i.i.d.* fashion during the system's entire life. In fact, the data will be dependent and the pdf that underlies them will keep changing whenever the system's operating mode changes. The condition *C.1* will definitely not hold true if the model is a black-box one, such as a neural network for example, because of the a priori wide range of variations of the network weights.

It is suggested here that the integration of the mechanistic modelling approach (which relies mostly on the use of the system's domain knowledge) and the SLT (which relies on the use of the system's empirical data) is the key to resolve the issue of FMESS's conditions fulfilment. The author [Guergachi, 2000] had showed that replacing the series of random variables $x_{i_0}^{data}(t_1), x_{i_0}^{data}(t_2),..., x_{i_0}^{data}(t_N)$ by the series $\delta(t_1,\mathbf{p}), \delta(t_2,\mathbf{p}),..., \delta(t_N,\mathbf{p})$ ($\mathbf{p}$ being the parameter vector of a *mechanistic* model) would lead to the partial fulfilment of the condition *C.3*. Also, the use of the prediction $x_{i_0}(\mathbf{p}, t_k)$ of a mechanistic (as opposed to a black-box) model in inequality (5) can guarantee the fulfilment of the condition *C.1*.

The intuitive idea behind the integration of mechanistic modelling and SLT lies in the fact that the more *domain knowledge* (or mechanistic information, i.e., information about the system mechanisms) a system model M has, the easier for M to see a pattern in the system data and, therefore, the more plausible the realization of the FMESS's conditions. At a more fundamental level, the task of integrating mechanistic modelling and SLT finds its justification (and abstraction) in a new paradigm for system modelling that uses the

concepts of '*system information content, transfer* and *conversion*', and that was introduced by the author in earlier publications [Guergachi, 2000; Guergachi and Patry, 2002a]. In this paradigm, information, like energy, can be converted from one form to another, and transferred from one system to another. The conditions $\{C.i\}$, if they hold true for a given physical system, would represent one form of useful information about this system (we will call this form "*statistical information*"), in the same way as mechanistic information (mass balance principle, kinetic laws, etc.) represent another form of information about the same system. Because of the very nature of complex environmental systems, mechanistic information is very hard to obtain and, in most cases, all that we can extract is a partial amount of it (otherwise, the system would become a simple one). However, if we manage to convert this partial mechanistic information into useful statistical information that allows us to guarantee the realization of the conditions $\{C.i\}$, then we will have won the battle of complex system model evaluation [Guergachi, 2001].

5. CONCLUSION

An overview of the framework for model evaluation and structure selection (FMESS) developed by the author was presented. The most significant results of this framework are the relationships between the empirical and exact measures of model performance. For these relationships to be valid, a set of conditions on the system uncertainty needs to be satisfied. It was suggested that the integration of the mechanistic modelling approach and the statistical learning theory is the key to resolve the issue of FMESS's conditions fulfilment. This integration finds its justification in a new paradigm for system modelling that uses the concepts of '*system information content*, *transfer* and *conversion*'.

6. ACKNOWLEDGEMENTS

The author wishes to thank the Faculty of Business of Ryerson University for its kind financial support.

7. REFERENCES

Guergachi, A., "*Uncertainty Management in the Activated Sludge Process - Innovative Applications of Computational Learning Theory*", Ph.D. thesis, University of Ottawa, Ottawa, Canada.

Guergachi, A., "Integrating Domain Knowledge and Machine Learning Theoretic-Methods for Modelling Complex Physical Systems", Research Proposal submitted to the National Science and Engineering Research Council, November 2001.

Guergachi, A. and G. Patry, "Statistical Learning Theory, Model Identification and System Information Content", *International Journal of General Systems*, accepted for publication, 2002a.

Guergachi, A. and G. Patry, "Using Statistical Learning Theory to Rationalize System Model Identification and Validation. Part I: Mathematical Foundations", *Complex Systems*, in review, 2002b.

Guergachi, A. and G. Patry, "Using Statistical Learning Theory to Rationalize System Model Identification and Validation. Part II: Application to Biological Treatment of Wastewater", *Complex Systems*, to be submitted, 2002c.

Jeppsson,U., "*Modelling Aspects of Wastewater Treatment Processes*", Lund Institute of Technology, Department of Industrial Electrical Engineering and Automation, Lund, Sweden, 1996.

Konikow, L. and Bredehoeft, J., "Ground-water models cannot be validated", *Adv. Water Resour.*, 19(2) 75-83, 1992.

Vapnik, N., "*Statistical Learning Theory*", Wiley, NY, 1998.

Zheng, C., and G. Bennett, "*Applied Contaminant Transport Modeling – Theory and Practice*", Van Nostrand Reinhold, NY, 1995.