Input Variable Selection for Data-driven Models of Coriolis Flowmeters for Two-phase Flow Measurement

Lijuan Wang[1], Yong Yan[1], Xue Wang[2], Tao Wang[3]

[1] School of Engineering and Digital Arts, University of Kent, Canterbury, Kent CT2 7NT, U.K.

[2] School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury, Kent CT2 7NF, U.K.

[3] KROHNE Ltd., 34-38 Rutherford Drive, Wellingborough NN8 6AE, U.K.

Abstract:

Input variable selection is an essential step in the development of data-driven models for environmental, biological and industrial applications. Through input variable selection to eliminate the irrelevant or redundant variables, a suitable subset of variables is identified as the input of a model. Meanwhile, through input variable selection the complexity of the model structure is simplified and the computational efficiency is improved. This paper describes the procedures of the input variable selection for the data-driven models for the measurement of liquid mass flowrate and gas volume fraction under two-phase flow conditions using Coriolis flowmeters. Three advanced input variable selection methods, including Partial Mutual Information (PMI), Genetic Algorithm - Artificial Neural Network (GA-ANN) and tree-based Iterative Input Selection (IIS) are applied in this study. Typical data-driven models incorporating Support Vector Machine (SVM) are established individually based on the input candidates resulting from the selection methods. The validity of the selection outcomes is assessed through an output performance comparison of the SVM based data-driven models and sensitivity analysis. The validation and analysis results suggest that the input variables selected from the PMI algorithm provide more effective information for the models to measure liquid mass flowrate while the IIS algorithm provides a fewer but more effective variables for the models to predict gas volume fraction.

Keywords: Input variable selection; Data driven model; Coriolis flowmeter; Two-phase flow measurement; Artificial neural network; Support Vector Machine

1. Introduction

Coriolis flowmeters are capable of measuring two-phase flows due to the recent advances in sensor and transmitter technologies [1]. However, the accuracy of Coriolis flowmeters for liquid flow measurement with entrained gas is still unsatisfactory. For single-phase flow measurement theoretical analysis has been conducted to assess the errors of Coriolis flowmeters in the measurement of mass flow rate and density [2-7]. Keita [2] assessed the effects of pressure and fluid compressibility on the meter readings through modelling of Coriolis flowmeters. Anklin [3] derived an equation to quantify the effects of compressibility and non-linear flow velocity on the measurement of compressible flow. Kutin et. al. [4] developed a mathematical model of a straight, slender-tube Coriolis meter to study the effects of axial force, added mass, damping and excitation on the measurement accuracy. Afterwards, they applied the weight vector theory to evaluate the effects of velocity profile on the sensitivity of Coriolis mass flowmeters [5, 6]. The velocity profile effects were found to depend on the configuration of Coriolis tubes. Enz [7] undertook experimental investigations into twin bent tubes to study the influence of asymmetrical and symmetrical pipe damping, added mass and increasing

ambient temperature of phased shifts in the case of zero mass flow. Extensive research has been undertaken over the past two decades to investigate the performance of Coriolis flowmeters under two-phase flow conditions. The compressibility effects on the readings of Coriolis flowmeters for gas-liquid two-phase flow measurement were discussed by Gysling [8] and Rieder [9], respectively. Subsequently, Hemp et. al. [10] conducted more theoretical investigations into straight beam-type Coriolis flowmeters, aiming to quantify the errors in the measurement of fluid density and mass flow rate due to the compressibility of the fluid. The effects of asymmetric damping of the meter tube and bubble buoyancy were mentioned in [10]. Basse [11] analyzed the measurement errors with the combined effects of compressibility and phase decoupling. Since there were some assumptions in developing the mathematical models and certain important factors such as volume fraction, pressure loss and pipe geometry were not taken into consideration, there were significant differences between the theoretical analysis and experimental results. Thanks to the new generation of flow transmitters, experimental tests prove that the liquid mass flow errors of Coriolis flowmeters under two-phase flow conditions are generally reproducible [12]. In this case, data-driven models incorporating soft computing algorithms can be used to correct the mass flow errors and predicting gas volume fraction under two-phase or multi-phase flow conditions. Henry and Liu [13] used a neural network to correct the mass flow error of a Coriolis flowmeter installed on a horizontal two-phase flow pipeline. For the same purpose, another model established on fuzzy systems was put forward by Safarinejadian *et. al*. [14]. However, the input variables in the proposed models were selected by experience and may underestimate the performance of dada-driven models. Although certain theoretical knowledge about the factors affecting the performance of Coriolis flowmeters is available, parametric dependency between the input variables is not studied and their significance and sensitivity to the desired outputs are not quantified, either. Moreover, input variable selection methods which aim to extract the information from the available data provide a new approach to determining the inputs of data-driven models for two-phase or multiphase flow measurement. For these reasons, input variable selection as a fundamental consideration in identifying the optimal structure of a data-driven model should be studied systematically before modelling a Coriolis flowmeter for two-phase flow measurement.

Input variable selection is in fact to establish the relationships within the available data and identify suitable predictors of the model output. In the case of artificial neural networks or other similar data-driven statistical models, input variable selection is indispensable for developing a suitable model. May *et. al.* summarized the input variable selection methods for artificial neural networks and gave several key considerations in determining the most appropriate approach to input variable selection for a given application [15]. The input variable selection techniques can be classified into three main categories: wrapper, embedded and filter algorithms. The first two are model-based and the latter is model-free. In order to choose an appropriate approach, the first consideration is to decide whether the case under consideration is a linear or non-linear problem. Subsequently, the computational requirements in the model-based techniques and the selection accuracy should be balanced as well. The Partial Mutual Information (PMI) approach, as an advanced model-free variable selection method, is able to minimize redundancy and maximize relevance in the available data by estimating the maximum joint mutual information. Bowden *et. al.* tested PMI and SOM-GAGRNN (Self Organizing Map - Genetic Algorithm and General Regression Neural Network) on a number of synthetic data sets and real-time forecasting simulation [16, 17]. It turned out that both approaches were acceptable when predictive performance was the primary aim. The variables determined using the PMI algorithm were most robust for the validation set and the PMI scores can reveal useful

information about the order of importance of each significant input. May *et. al.* further researched on the partial mutual information based input variable selection for non-linear artificial neural networks [18]. In order to resolve the problems of underlying assumption of linearity and redundancy within the available data, three novel termination criteria, i.e. Tabulated critical values, Akaike Information Criterion (AIC) and Hampel test criterion, were proposed to improve computational efficiency and accuracy of the algorithm. Because the data distributions in real-world applications are often unknown and the assumption of Gaussian data may be inappropriate, the AIC and Hampel criteria are recommended for wider applicability. Galelli *et. al.* proposed a hybrid approach combining both model-free and model-based methods, namely, the tree-based Iterative Input variable Selection (IIS) algorithm, in 2013 [19]. In this algorithm a tree-based ranking method is used in place of the information-theoretic measure such as PMI to estimate the information gained from the data. Results indicate that IIS is capable of selecting the most significant and non-redundant inputs in different test conditions. Subsequently, Galelli *et. al.* made an inter-comparison of four input variable selection methods, including PMI, PCIS (Partial Correlation Input Selection), IIS and GA-ANN, through testing on several datasets, and evaluated the four algorithms in terms of selection accuracy, computational efficiency and qualitative criteria [20]. Li *et. al.* presented preliminary guidelines for the selection of most appropriate methods for obtaining the required Kernel Density Estimates (KDEs) in the PMI method [21]. The use of alternative bandwidth estimators can result in significant improvements in the PMI method for non-normally distributed data.

In order to realize two-phase or multi-phase flow measurement using Coriolis flowmeters, data-driven models are to be established for the flowmeters to predict liquid mass flowrate and gas volume fraction. This paper focuses on input variable selection for the data-driven models through comparing different selection techniques. In this research, experimental data were obtained from a gas-water two-phase flow test rig with liquid mass flow rate ranging from 700 kg/h to 14,500 kg/h and gas volume fraction between 0 and 30%. In view of the gravitational effect on two-phase flows and the performance of Coriolis flowmeters, separate models should be established for vertical and horizontal installations, respectively. For each installation of the flowmeter, two independent models are to be developed, one for liquid mass flowrate and the other one for gas volume fraction. For each model, there are a total of 12 potential input variables which are direct outputs of the Coriolis flowmeters or transformed internal parameters in addition to the outputs from differential-pressure (DP) transducers and electric impedance sensors. The variable selection process is realized through analyzing the underlying relationships between the variables and desired outputs based on the observational evidence and interpreting the selection outcomes in view of physical properties of the two-phase fluid. Three input variable selection methods, PMI, GA-ANN and IIS are applied for each model. SVM (Support Vector Machine) based data-driven models are built up with the selected input candidates and the validity of the selected variables is assessed by comparing the output performance of the models. The parametric dependence, significance and sensitivity of the input variables to the desired outputs are discussed in the following sections. The most suitable subset of input variables for each data-drive model is finally recommended.

2. Input variable selection

The model-based input variable selection approach aims to select the variable set which makes the model perform well through establishing and evaluating the model based on the potential variable

combinations. The main drawback of this approach is the high computational requirement due to a large number of calibration and validation processes. Moreover, the selection result depends on the predefined model in terms of architecture and parameters. On the contrary, the model-free approach is directly based on the information (interclass distance, statistical dependence, or information theory, etc.) between the available dataset, so the computational efficiency is not an issue. However, a trade-off criterion should be defined to balance the significance measurement and the number of selected variables. In this study, three input variable selection methods, PMI, GA-ANN and IIS, are considered to identify input variables before modelling of Coriolis flowmeters for two-phase flow measurement.

2.1 Partial Mutual Information (PMI)

The PMI input variable selection method was first proposed by Sharma in 2000 for the identification of inputs for hydrological models [20]. PMI is a model-free variable selection method, which utilizes a measure of the partial dependence between a potential input variable and the output, conditional on any inputs that have already been selected. Earlier research on evaluating the performance of PMI on synthetic data sets and real-world data sets shows that PMI is a promising method [18, 21].

Given a dependent variable Y, the potential input variable pool $x=\{x_1, x_2, \dots, x_m\}$ and the already selected variables S, the PMI value of a potential input variable x_i can be formulated as:

$$PMI_i = \frac{1}{n}\sum_{j=1}^{n}\log\left[\frac{\hat{f}(v_j,u_j)}{\hat{f}(v_j)\hat{f}(u_j)}\right] \tag{1}$$

$$u = Y - \hat{m}_Y(S) \tag{2}$$

$$v = x_i - \hat{m}_{x_i}(S) \tag{3}$$

where $\hat{m}_Y(S)$ is the conditional expectation of Y given an observed S. $\hat{m}_{x_i}(S)$ is the conditional expectation of x_i given an observed S. $u=\{u_1,u_2,\dots,u_n\}$ and $v=\{v_1,v_2,\dots,v_n\}$ represent the residual information in dependent variable Y and potential input variable x_i once the effect of the already selected inputs S has been taken into consideration. n is the number of observations. $\hat{f}(v_j)$, $\hat{f}(u_j)$ and $\hat{f}(v_j,u_j)$ are the estimated marginal and joint probability density functions which are realized by kernel density estimation.

Akaike information criterion (AIC) is considered as the termination criterion of the PMI algorithm, as it can provide a general measure of the trade-off between information gain and the complexity introduced to the model by the addition of input variables. It is based on the analysis of the output variable residual u. Once there is no further reduction in the information contained in u, the optimal input variable set is reached and the selection is terminated.

$$AIC = n\log_e\left(\frac{1}{n}\sum_{j=1}^{n}u_j^2\right) + 2p \tag{4}$$

where p is the number of model parameters. u is the residual of the desired and estimated outputs and $\boldsymbol{u}$ can be calculated from equation (2).

The search strategy is a forward selection procedure which is described as follows:
(1) Initialize the selected input set S with null.

(2) Calculate the PMI value (eq.(1)) for each potential variable which is beyond *S*.
(3) Find out the variable x_i with the highest PMI value.
(4) Calculate the AIC score (eq.(4)), assuming x_i is included in *S*.
(5) If AIC score decreases, variable x_i can be added to *S* and go to step (2).
Otherwise variable x_i is rejected and the selection is terminated.

2.2 Genetic Algorithm - Artificial Neural Network (GA-ANN)

GA-ANN is a kind of model-based input variable selection method. It comprises of a simple 1-hidden node multilayer perceptron neural network as a regression model and genetic algorithm as the heuristic search strategy. The potential variable subsets are generated by the genetic operations: selection, crossover and mutation. The suitability of the input variables is determined by assessing the performance of the neural network which is established based on the corresponding variables. The out-of-sample AIC is considered to quantify the accuracy of the neural network after k-fold cross validation. The termination criterion of the genetic algorithm is either the maximum number of evaluations or convergence of the fitness function is achieved [22].

A genetic algorithm usually includes five steps: population initialization, fitness function, selection, crossover and mutation. The selection procedure is as follows:
(1) Initialise a population with a random population of chromosomes.
(2) Each chromosome in the population is decoded into a solution. Calculate its fitness using an objective function.
(3) The chromosome with best fitness is selected to generate a second generation.
(4) Partially exchange genetic information between two parent chromosomes during crossover.
(5) With some low probability, a portion of the new individuals have some of their chromosomes flipped during mutation, which is used to keep the population diverse and prevent the algorithm from prematurely converging onto a local minimum.
(6) If the maximum number of evaluations or convergence of the fitness function is reached, the selection procedure is over. Otherwise, go to (2).

2.3 Iterative Input Selection (IIS)

The IIS method is a combination of model-free and model-based methods. It utilizes extra-trees to estimate the relative contribution of each candidate input. The ranking-based evaluation does not require any assumption on the statistical properties of the input data set (e.g. Gaussian distribution) and can be applied to any sort of samples. Moreover, this approach does not rely on computationally intensive methods (e.g. bootstrapping) to estimate the information content in the data and thus is generally faster and more efficient [23]. The relevance *G* of the variable x_i in explaining the output *Y* can be evaluated by:

$$G = \frac{\sum_{t=1}^{M}\sum_{j=1}^{\Omega} k \cdot \Delta_{\text{var}}(w_j)|D|}{\sum_{t=1}^{M}\sum_{j=1}^{\Omega} \Delta_{\text{var}}(w_j)|D|} \tag{5}$$

where *M* is the number of different trees and **Ω** is the number of non-terminal nodes in the tree. w_j is the j^{th} non-terminal node in the t^{th} tree. *k* is equal to 1 if the variable x_i is used to split the node w_j,

otherwise k is 0. $\Delta_{\text{var}}(w_j)$ is the variance reduction associated to node w_j. $|D|$ is the number of observations in training dataset D.

The input variables are sorted by decreasing the relevance of potential variables, and thus the first variable should be most significant. In order to reduce miss-selection and minimize the redundancy, the first p variables in the ranking are further evaluated by SISO (Single Input Single Output) and MISO (Multiple Input Single Output) models. The termination criterion is defined as either the best variable obtained in the current iteration is already in the selected variable set S, or the improvement of the model performance reaches to tolerance ε.

The search strategy is forward selection and the selection procedure is as follows:
(1) Calculate the input ranking of the potential input variables according to the explained variance and pick out the p most relevant input variables.
(2) Build SISO models for each of the first p-ranked variables. Compute the distance metric between Y and the model output, and then select the most relevant variable x_i and add it to the selected variable set S.
(3) Build a MISO model with selected variable set S and compute the distance metric between Y and the model output.
(4) Evaluate the variation of the distance metric (D) between the current and previous iteration. If D is less than tolerance ε or a variable is selected twice, the selection is terminated.

3. Experiments and results

3.1 Experimental tests

Experimental tests were conducted on a 1-inch bore, air-water two-phase flow test rig (see Fig. 1). Two independent Coriolis flowmeters were installed before the mixer to provide references for the individual mass flowrates of the liquid and gas phases. Due to the effect of gravity and buoyancy on the gas-liquid two-phase flow, bubbles distributed within the Coriolis measuring tubes are different between horizontal and vertical installations and hence Coriolis flowmeters perform differently. For this reason, two additional Coriolis flowmeters of the same type were installed in the horizontal and vertical test sections, respectively, on the test rig to measure the liquid mass flowrate and gas volume fraction. The Coriolis flowmeters under test are twin bent-tube design, as shown in Fig. 2. Two DP transducers were also used to record the DP value across each Coriolis flowmeter under test. In addition, an electric impedance sensor was installed in the upstream of the Coriolis flowmeter on the horizontal test section.

Two series of tests (Tests I and Tests II) were conducted over a range of liquid mass flowrates from 700 kg/h to 14500 kg/h with gas volume fraction from 0 to 30%. The test points with variations in liquid mass flowrate and gas volume fraction are summarized Fig. 3. For the purpose of training a data-driven model, a dataset of 237 records (circular markers in Fig. 3) was collected from Tests I while the dataset of 24 records (triangular markers in Fig. 3) from Tests II for testing the model. Each dataset represents the average of all recorded values within an approximate window of 100 seconds. The fluid temperature during the tests was between 18°C~24°C, which varied with the ambient temperature in the laboratory. It was observed during the experiments that the flow pattern in the

horizontal pipe was mostly slug or dispersed bubbly flow whilst the flow was of bubbly and dispersed bubbly nature in the vertical pipe. However, it must be borne in mind that the observed flow patterns are different from the flow regimes within the Coriolis measuring tubes as the incoming fluid is separated and flows into two smaller bent tubes (Fig.2). Although the bubble distributions within the two tubes are unknown, the variable selection methods based on the whole experimental data will take into consideration the flow patterns in horizontal or vertical pipes.

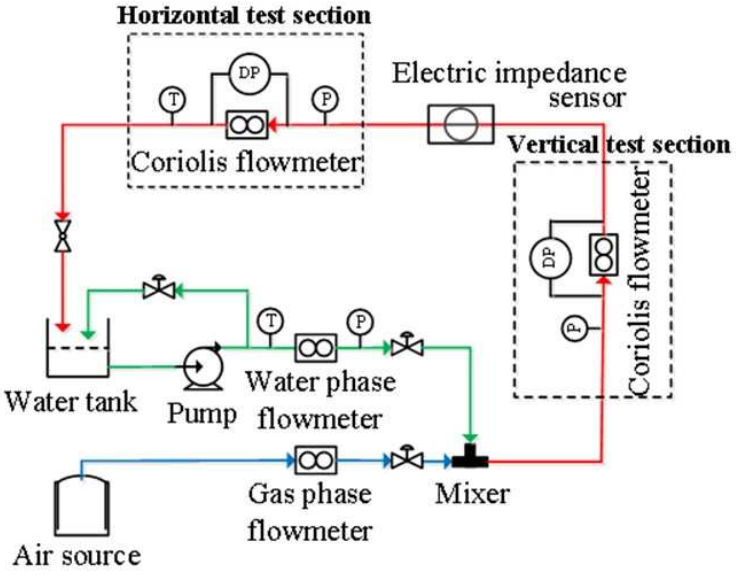


Fig. 1 Schematic of the two-phase flow test rig

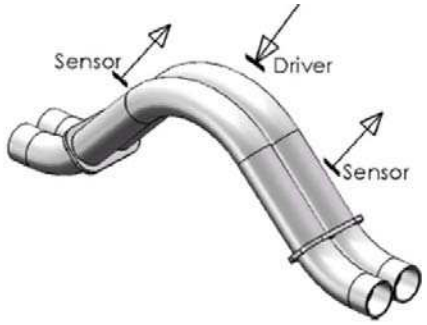


Fig. 2 Typical design of twin bent Coriolis measuring tubes [1]

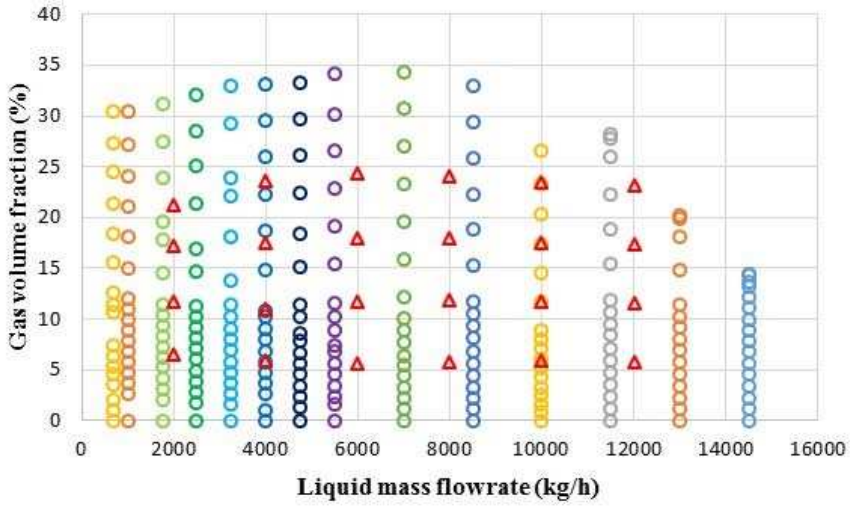


Fig. 3 Experimental test points of the gas-liquid two-phase flow

3.2 Experimental results

The variables and their corresponding physical definitions are outlined in Table I. All input variables (x_1-x_{14}) are obtained through measuring or transforming the internal parameters of the two Coriolis flowmeters, DP transducers and the electrical impedance sensor. Variables x_1-x_3 are direct measurement outputs from Coriolis flowmeters while x_4-x_{11} are their internal parameters. The measurement principle of mass flowrate ($\dot{m}$) of a single-phase flow using Coriolis flowmeters is represented as [24, 25]

$$\dot{m}=K_R[1+(K_T+e)\Delta T+C_1\Delta\sigma+C_2\Delta\tau+C_3\Delta T+C_4\Delta T^2]t_d \quad (6)$$

where K_R is a calibration factor for the measurement of the mass flowrate of a single-phase flow at a reference temperature, K_T is a material temperature dependence factor, e is a device-specific temperature dependence factor, and ΔT is a relevant temperature difference. $\Delta\sigma$ and $\Delta\tau$ indicate the relevant differences on circumferential stress and axial stress of a measuring tube. C_1, C_2, C_3 and C_4 are corresponding calibration coefficients. t_d is the time shift between the signals from motion sensors A and B.

The fluid density (ρ) from a Coriolis flowmeter is determined by

$$\rho=A_R(1+K_E\Delta T)\frac{1}{f^2}-B_R \quad (7)$$

where A_R and B_R are calibration factors to measure the density of single phase flow, K_E is a temperature dependence factor, and f is the frequency of a measuring tube.

Coriolis flowmeters are capable of providing accurate measurements of mass flowrate and density of single-phase flow. However, the mass flowrate and density readings are erroneous under the condition of two-phase or multiphase flows due to the effect of additional phase on the vibration of the measuring tube. In this case, the erroneous mass flowrate and density are defined as apparent mass flowrate and observed density (i.e. x_1 and x_2 in Table I).

x_{12} (DP) from the DP transducer represents the pressure difference across a Coriolis flowmeter. x_{13} and x_{14} are from the electric impedance sensor and represent the magnitude and phase angle of the

impedance of the flow between two electrodes. Under two phase flow conditions, it is impossible to measure the liquid mass flowrate and gas volume fraction directly using each of the instruments alone. However, the variations of the variables in Table I reflect the changing of the desired outputs Y_1 (desired liquid mass flowrate) and Y_2 (desired gas volume fraction) to some extent. Due to the reproducibility of the experimental tests, it is possible to establish data-driven models using the available experimental data to estimate liquid mass flowrate ($\hat{Y}_1$) and gas volume fraction ($\hat{Y}_2$). Because some of the variables from different sensors are independent and some are related to each other in physical sense, it is difficult to empirically determine which variable has more contribution to explain the desired outputs. It is necessary to formulate the relationship between the potential input variables and the desired outputs and then identify significant variables which are able to explain the desired outputs completely. Meanwhile, the repeated information should be eliminated to reduce the complexity of the model structure and improve the performance of the model.

Table I Variables and their corresponding physical definitions

ID	Variable name and symbol	Physical definition	Source
x_1	Apparent mass flowrate (m)	The mass flowrate reading from Coriolis flowmeters based on calibration characteristics for single-phase flows.	Coriolis flowmeter
x_2	Observed density (ρ)	The density reading from Coriolis flowmeters based on calibration characteristics for single-phase flows.	Coriolis flowmeter
x_3	Process temperature (T)	The temperature reading from Coriolis flowmeters	Coriolis flowmeter
x_4	Sensor-A amplitude (V_A)	The voltage amplitude of signals from motion sensor-A.	Coriolis flowmeter
x_5	Sensor-B amplitude (V_B)	The voltage amplitude of signals from motion sensor-B.	Coriolis flowmeter
x_6	Drive level (I_D)	The current amplitude of the driver output.	Coriolis flowmeter
x_7	Time shift (t_d)	The time delay between the signals from the two motion sensors induced by flow.	Coriolis flowmeter
x_8	Tube frequency (f)	The oscillation frequency of the Coriolis measuring tube(s) inside Coriolis flowmeters.	Coriolis flowmeter
x_9	Sensor balance (B_s)	The ratio of the voltage amplitude of sensor-A signal to sensor-B signal.	Coriolis flowmeter
x_{10}	Damping (K)	Damping factor of Coriolis measuring tube(s)	Coriolis flowmeter
x_{11}	Two phase indicator	An indicator for the detection of a two-phase or multiphase flow [26].	Coriolis flowmeter
x_{12}	Difference pressure (DP)	DP is the differential pressure across the Coriolis flowmeter.	DP transducer

x_{13}	Magnitude of Impedance ($\|Z\|$)	The ratio of the voltage difference amplitude to the current amplitude.	Electrical impedance sensor
x_{14}	Phase factor of Impedance (θ)	The phase difference between voltage and current.	Electrical impedance sensor
Y_1	Desired liquid mass flowrate	Desired liquid mass flowrate on the test section is obtained from liquid reference Coriolis flowmeter	Liquid reference Coriolis flowmeter
Y_2	Desired gas volume fraction	Desired gas volume fraction on the test section is obtained by calculation according to liquid and gas reference Coriolis flowmeters	Liquid and gas reference flowmeters
$\hat{Y}_1$	Estimated liquid mass flowrate	Output of data-driven models for the measurement of liquid mass flowrate under two-phase flow conditions.	Data-driven model
$\hat{Y}_2$	Estimated Gas volume fraction	Output of data-driven models for the measurement of gas volume fraction under two-phase flow conditions.	Data-driven model

In consideration of different performances of the Coriolis flowmeter on horizontal and vertical installations, two separate models are to be built for each position. Model-H-L and Model-H-G present the models built for the Coriolis flowmeter in the horizontal section to measure liquid mass flowrate and gas volume fraction, respectively. Similarly, Model-V-L and Model-V-G indicate the models built in the vertical section to measure liquid mass flowrate and gas volume fraction, separately. Since apparent mass flowrate (x1) and observed density (x2) are derived directly from time shift (x7) and tube frequency (x8), respectively, with temperature compensation, the possibility of replacing x7 and x8 with x1 and x2 is considered, respectively. For this reason, x7 and x8 are excluded in Table II. Regarding each model, there are 12 potential input variables and one output. The input variables and outputs from the four models are outlined in Table II.

Table II Symbols of the input variables and corresponding outputs for the four data models

Installation	Model	Potential input variables	Model Output
Horizontal pipe	H-L	x_1-x_6, x_9-x_{14}	$\hat{Y}_1$
	H-G	x_1-x_6, x_9-x_{14}	$\hat{Y}_2$
Vertical pipe	V-L	x_1-x_6, x_9-x_{14}	$\hat{Y}_1'$
	V-G	x_1-x_6, x_9-x_{14}	$\hat{Y}_2'$

3.3 Implementation of input variable selection

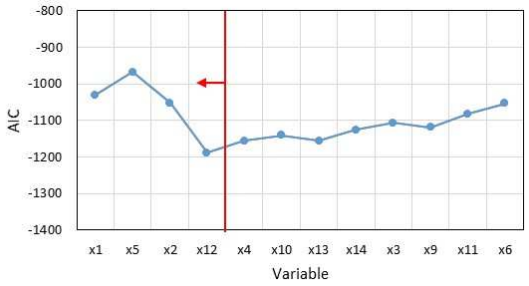
3.3.1 Implementation of PMI

The PMI algorithm was implemented on the experimental dataset for each model. The variable with the highest PMI value is selected to the candidate input variable set for each iteration. Fig. 4 shows

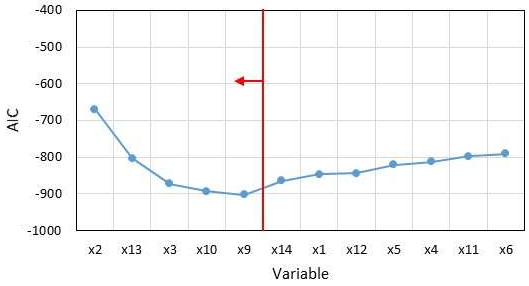
the change of AIC score when a new variable is added. As Fig. 4 (a)-(d) shown, the AIC score has a decreasing trend until a new variable is selected so that the termination criterion is reached. The variables before the red line in Fig. 4 (a)-(d) are the final selected variables for the four models.

Model-H-L and Model-V-L are applied to the Coriolis flowmeters on horizontal and vertical pipelines, respectively, for the measurement of liquid mass flowrate. The results of PMI selection (Fig. 4 (a) and (c)) show that variables x_1 (apparent mass flowrate), x_2 (observed density) and x_{12} (DP) are main factors for the two models to estimate the desired liquid mass flowrate. This means, although the Coriolis flowmeters on different orientations have different performances, the main factors affecting liquid mass flowrate are the same. Variable x_5 (sensor-B amplitude) provides additional information for Model-H-L while x_{10} (damping) is helpful for Model-V-L. The combined effect of the variables is more significant than that of an individual variable on the output. The selection results demonstrate x_1 has more contribution than the other variables to the measurement of liquid mass flowrate as one would expect from a purely physical point of view. The difference in selection sequence between the models for horizontal and vertical pipes is due to the effect of installation on the performance of Coriolis flowmeters.

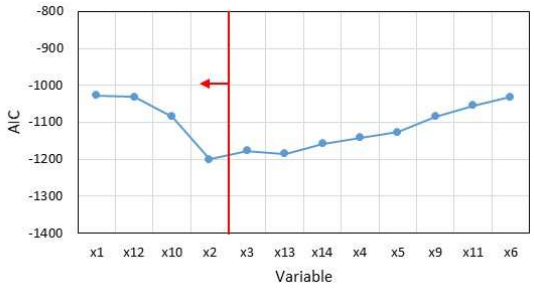
Model-H-G and Model-V-G are applied for gas volume fraction prediction of Coriolis flowmeters on horizontal and vertical pipelines, separately. Fig. 4(b) and Fig. 4(d) show that variables x_2 (observed density), x_{13} (magnitude of impedance) and x_3 (process temperature) are obtained from the PMI selection procedures for both Model-H-G and Model-V-G. Apart from the three variables, x_{10} and x_9 are also beneficial to Model-H-G for fitting the desired output. As expected, variable x_2 was first selected for Model-H-G and Model-V-G since the observed density changes significantly in relation to the amount of gas entrained in the liquid flow.



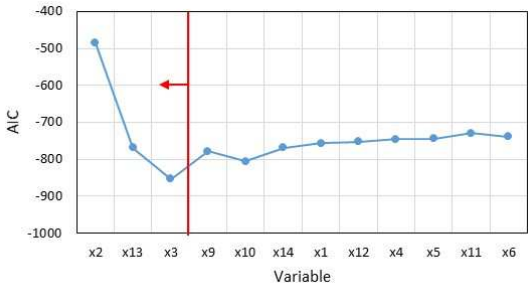
(a) PMI for Model-H-L



(b) PMI for Model-H-G



(c) PMI for Model-V-L



(d) PMI for Model-V-G

Fig. 4 Procedures of PMI input variable selection for Models H-L, H-G, V-L and V-G

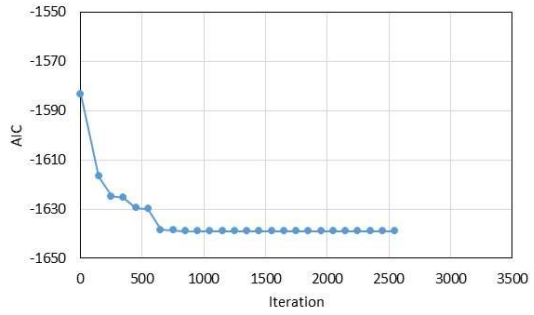
3.3.2 Implementation of GA-ANN

GA-ANN input variable selection was implemented on the dataset of Models H-L, H-G, V-L and V-G. The GA algorithm takes thousands of iterations to conduct selection, crossover and mutation before the AIC value is converged. The iteration process for each model is shown in Fig. 5. The out-of-sample AIC for the four models are individually 1639.06, 958.65, 1555.46 and 911.36.

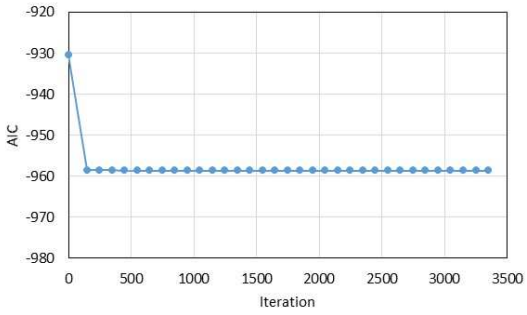
The selected input variables are x_1, x_5, x_6, x_{10} and x_{12} for Model-H-L, while x_1, x_5, x_6, x_9, x_{10} and x_{12} for Model-V-L. It is obvious that the common variables x_1 (apparent mass flowrate), x_5 (sensor-B amplitude), x_6 (drive level), x_{10} (damping) and x_{12} (DP) are important for the pre-defined ANN models to fit the desired liquid mass flowrate.

The selected inputs variables are x_2, x_3, x_4, x_5, x_{11}, x_{12}, x_{13}, x_{14} for Model-H-G, while x_2, x_3, x_4, x_9, x_{11}, x_{12}, x_{13}, x_{14} for Model-V-G. As the GA-ANN selection results suggest, the variables x_2 (observed density), x_3 (process temperature), x_4 (sensor-A amplitude), x_{11}(two phase indicator), x_{12} (DP), x_{13} (magnitude of impedance) and x_{14} (phase factor of impedance) are common important variables for the pre-defined models to predict gas volume fraction.

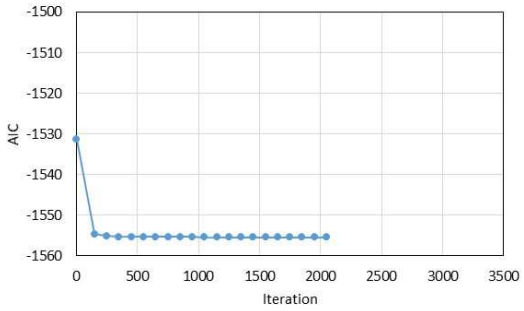
The outcomes of GA-ANN include more variables than PMI. The selected subset is restricted by the structure of ANN and the problem of redundancy in the selected variables is more serious than PMI. Moreover, it cannot provide the information of significance level of the selected input variables.



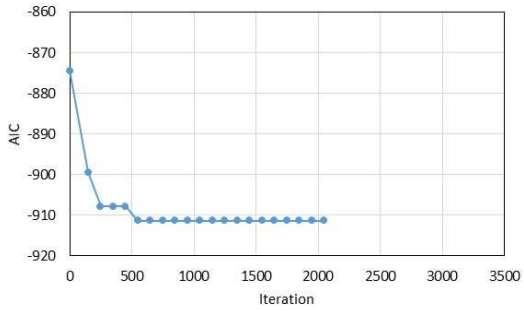
(a) GA-ANN for Model-H-L



(b) GA-ANN for Model-H-G



(c) GA-ANN for Model-V-L



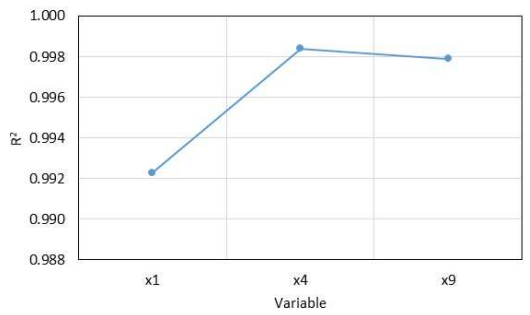
(d) GA-ANN for Model-V-G

Fig. 5 Integration of GA algorithm for Models H-L, H-G, V-L and V-G

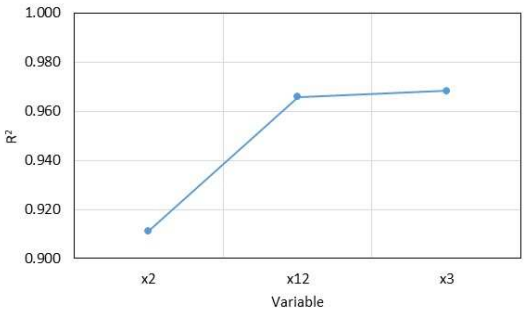
3.3.3 Implementation of IIS

IIS algorithm was executed on the dataset of Models H-L, H-G, V-L and V-G, respectively. The candidate subset was incrementally built through ranking, SISO and MISO evaluation. The selection process and the performance of MISO are shown in Fig. 6 for each model. For Model-H-L, the tolerance ε is reached after x_1, x_4 and x_9 are all included. For Model-H-G, the selection process is terminated because variable x_{12} is selected twice. For Model-V-L, the tolerance ε is reached after x_1, x_4, x_2 and x_9 are selected. For Model-V-G, input variable x_{12} is also selected twice resulting in the selection procedure terminated.

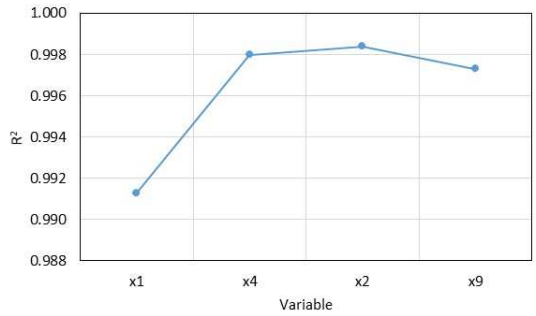
From the results of Model-H-L, the variables x_1 (apparent mass flow), x_4 (sensor-A amplitude), x_9 (sensor balance) are selected while variables x_1, x_4, x_2 (observed density), x_9 are selected for Model-V-L. This presents the information of apparent mass flow and sensor balance are two most useful variables to improve the performance of MISO models.
From the results of Model-H-G, the variables x_2, x_{12} (DP), x_3 (process temperature) were selected while variables x_2, x_1, x_{12} were selected for Model-V-G. The fitting results of MISO models with the inputs of the information of observed density and mass flowrate information (DP or apparent mass flowrate) are significantly improved.



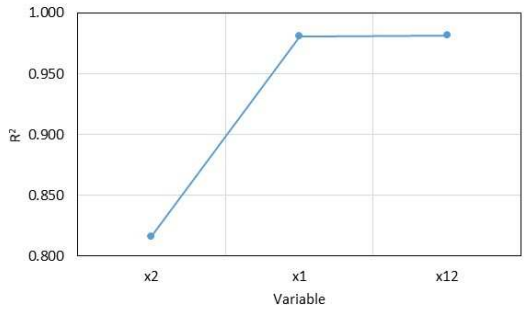
(a) IIS for Model-H-L



(b) IIS for Model-H-G



(c) IIS for Model-V-L



(d) IIS for Model-V-G

Fig. 6 Results of IIS for Models H-L, H-G, V-L and V-G

3.3.4 Comparison of PMI, GA-ANN and IIS methods

The selected variables by PMI, GA-ANN and IIS methods and corresponding running time (the laptop processor is Intel(R) Core (TM) i5-4200U CPU @ 1.60GHz) are summarized in Table III. It is clear that GA-ANN algorithm takes longer time to implement the heuristic search than the other two methods. The running time of GA-ANN is about 10 times of IIS running time and 100 times of PMI

running time. In view of time efficiency, PMI algorithm is the most effective approach and IIS is in the second place.

In terms of the number of selected input variables, GA-ANN produces more candidate variables than PMI and IIS. Moreover, there is much more redundant and unnecessary information in the selected subset. IIS generates smallest subset without such redundancy. Comparing the outcomes from IIS and PMI, the important information of the candidate variables from IIS are mostly included in the selected variables from PMI.

Table III Variable selection outcomes for PMI, GA-ANN and IIS

Model	PMI		GA-ANN		IIS	
	Selected variables	Elapsed time (s)	Selected variables	Elapsed time (s)	Selected variables	Elapsed time (s)
H-L	x_1, x_5, x_2, x_{12}	11.81	x_1, x_5, x_6, x_{10}, x_{12}	5281.93	x_1, x_4, x_9	217.33
H-G	x_2, x_{13}, x_3, x_{10}, x_9	13.10	x_2, x_3, x_4, x_5, x_{11}, x_{12}, x_{13}, x_{14}	2689.13	x_2, x_{12}, x_3	242.92
V-L	x_1, x_{12}, x_{10}, x_2	13.14	x_1, x_5, x_6, x_9, x_{10}, x_{12}	4152.62	x_1, x_4, x_2, x_9	279.63
V-G	x_2, x_{13}, x_3	12.63	x_2, x_3, x_4, x_9, x_{11}, x_{12}, x_{13}, x_{14}	721.12	x_2, x_1, x_{12}	247.57

3.4 Results and discussion

3.4.1 Validation

In order to compare the performance of the three variable selection methods, a typical data-driven model is established for each case with SVM. SVM regression performs a linear regression in the high dimensional feature space using ε-insensitive loss and tends to reduce the model complexity by minimizing the weight vector [27]. Since an SVM model is based on the statistic learning theory and structural risk minimization, the output is not affected by the initial parameters and pre-defined structure of the model. Due to the relatively constant structure and better performance of the SVM model, it is applied to establish the relationship between the input variables and the output. In this study, the parameters for building the SVM models are optimized through 10-fold cross validation. The performance of the SVM model is quantified by Normalised Root-Mean-Square Error (NRMSE), which is defined as

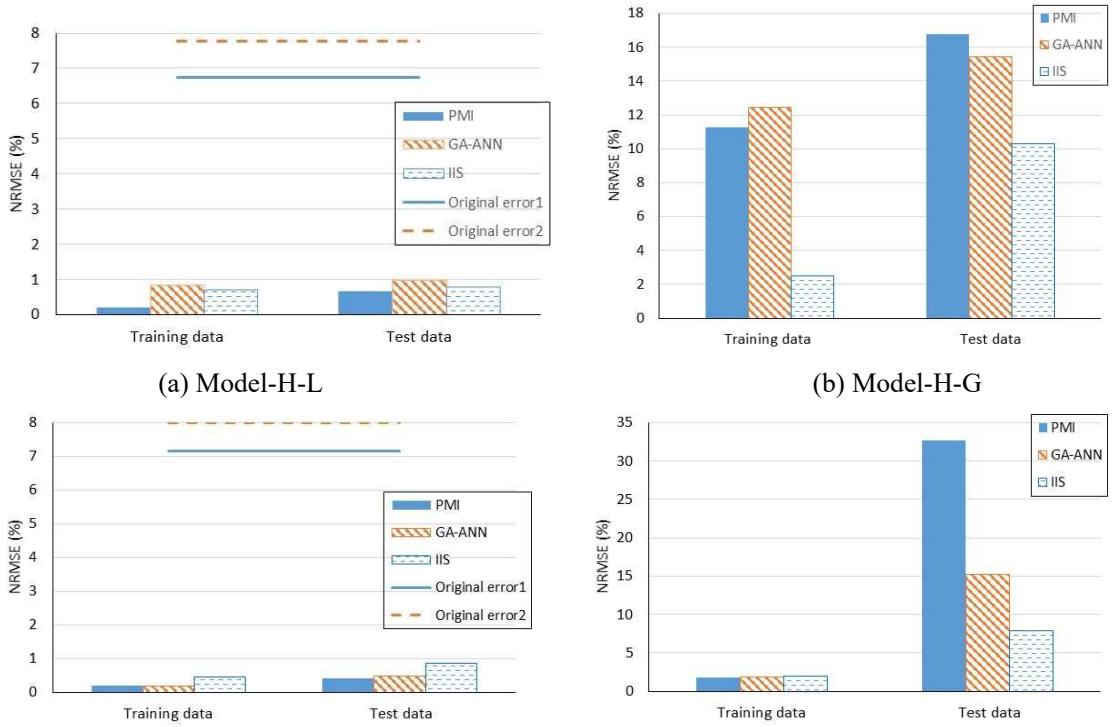
$$NRMSE = \frac{1}{\bar{y}}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2} \tag{8}$$

where y_i is the reference value of the model output, $\bar{y}$ is the mean of y_i, $\hat{y}_i$ is the model output, and n is the number of observations.

For each model, three SVM-based models were established based on the selected variables from PMI, GA-ANN and IIS, respectively. Each SVM-based model was tested on the training data and test data to evaluate the performance of the models which was mainly affected by the input variables. The solid line (original error1) and dash line (original error2) in Fig. 7 (a) and (c) present the original errors of mass flowrate on training data and test data. After correction by SVM-based models, the error of

liquid mass flowrate from Coriolis flowmeters on horizontal and vertical pipes are both dramatically decreased. Due to the inherent limitation of generalization ability of data-driven models, the errors on test data are larger than those on training data. Through comparing the NRMSE values of the three SVM-based models, the model with the variables selected by PMI has the lowest error than the other two. This means the selected variables by PMI include better and more completed information to explain the liquid mass flowrate.

As shown in Fig. 7 (b) and (d), the SVM-based models with PMI input variables does not perform well to predict gas volume fraction. Alternatively, IIS input variables make the model predict gas volume fraction with relatively lower error. This illustrates IIS provides the fundamental variables for the prediction of gas volume fraction.
The performance of SVM-based models with GA-ANN input variables is not as good as the models based on variables from PMI and IIS.



(a) Model-H-L

(b) Model-H-G

(c) Model-V-L

(d) Model-V-G

Fig. 7 Performance of RBF neural networks with input variables selected for PMI, GA-ANN and IIS

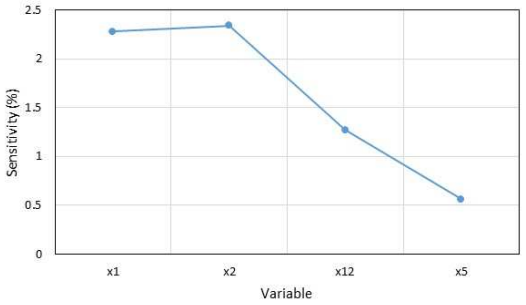
3.4.2 Sensitivity analysis

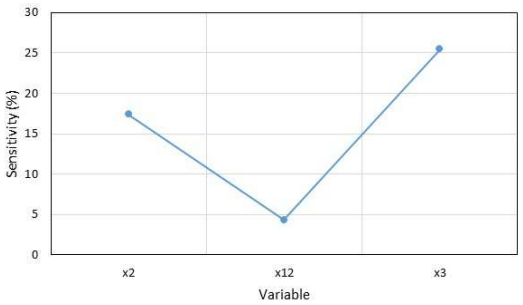
Sensitivity analysis is used to evaluate how sensitive the model output is to the changes in the value of input variables and also identify which input variables are important in contributing to the prediction of the output variables [28, 29]. In performing the sensitivity analysis, each variable of the model's inputs is increased by 5% in turn. The aim is to assess the effect of small changes in each input on the model output. The percentage change in the output as a result of the increase in each of the inputs is the sensitivity, which is defined as:

$$S_a = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\hat{z}_i - \hat{y}_i}{\hat{y}_i}\right)^2} \times 100\% \quad (9)$$

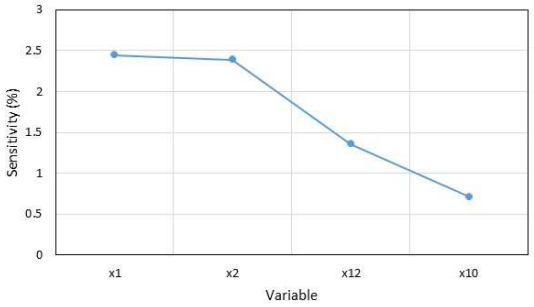
where, $\hat{z}_i$ is the model output after small changes added to the input variables, $\hat{y}_i$ is the model output with no changed input variables and n is the number of observations.

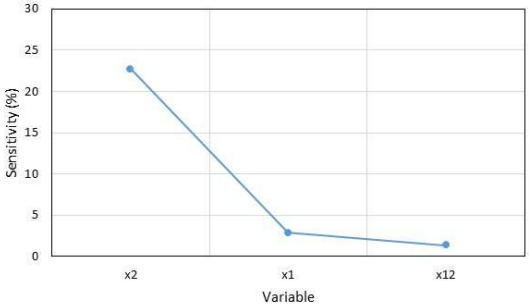
Sensitivity analysis was conducted on the SVM-based models with PMI input variables for liquid mass flowrate measurement, and the SVM-based models based on IIS input variables for gas volume fraction prediction. The sensitivity of each variable is plotted in Fig. 8. From Fig. 8 (a) and (c), it is clear that variables x_1 (apparent mass flowrate) and x_2 (observed density) have the higher sensitivity to the model output than the other variables. They have more relative influence on the measurement of liquid mass flowrate. Fig. 8 (b) and (d) show that small changes in variable x_2 (observed density) can result in more significant variations on the model output. Moreover, temperature fluctuation has more effect on the measurement of gas volume fraction than liquid mass flowrate.





(a) Sensitivity of input variables for Model-H-L (b) Sensitivity of input variables for Model-H-G





(c) Sensitivity of input variables for Model-V-L (d) Sensitivity of input variables for Model-V-G

Fig. 8 Sensitivity of input variables for the four models

4. Conclusions

Input variable selection as one of the most important steps in the development of data driven models, determines the quality and quantity of the information used in the modelling process. In order to establish a functional and optimal model for Coriolis flowmeters under two-phase flow conditions, input variable selection and validation have been conducted in this study. Moreover, parametric dependence, significance and sensitivity of the input variables to the desired outputs have been investigated. In view of the different installation conditions of the Coriolis flowmeters, four

independent models, H-L, H-G, V-L and V-G, have been established and evaluated to obtain the liquid mass flowrate and gas volume fraction. Three input variable selection approaches, including PMI, GA-ANN and IIS, have been applied to determine the most suitable variable subset for each model. The validity of the selected variables has been assessed on SVM-based models through a comparison of the NRMSE and sensitivity analysis. The following conclusions can be drawn from the results:

(1) In terms of time efficiency the PMI and IIS algorithms have outperformed the GA-ANN algorithm. The genetic operations and repeated model building and evaluation in the GA-ANN algorithm are achieved at the cost of long computation time. The IIS is semi-model based algorithm and hence takes more running time than the PMI.

(2) In terms of selection accuracy all the subsets selected from PMI, GA-ANN and IIS are able to explain the model output and have reduced the original errors of mass flowrate. Particularly, PMI provides more effective variables for the measurement of liquid mass flowrate than the other two approaches. For the prediction of gas volume fraction, the IIS has selected the most important information and thus generate a better performing model. The variables obtained from the GA-ANN contain some redundant information, while PMI and IIS can effectively avoid the redundancy to some extent and produce valid input variables for the models.

(3) With regard to the experimental data obtained in this study, the most important variable set for the measurement of liquid mass flowrate includes observed density, apparent mass flowrate, DP and damping while those for the prediction of gas volume fraction include observed density, apparent mass flowrate and DP.

The validation and analysis results suggest that the input variables selected from the PMI algorithm provide more effective information to measure the liquid mass flowrate while the IIS algorithm provides a fewer but more effective variables to predict the gas volume fraction. Although variable selection approaches can provide some valuable information to determine the input variables of a data-driven model, the accuracy of the methods also depend on the observational dataset, such as data size and their distributions. A dataset with less data or low-quality may result in underestimation or overestimation of the candidate variables for a data-driven model. Consequently, in order to ensure the selection accuracy with limited size of a dataset, it is necessary to determine the input subset using variable selection methods incorporating physical interpretation of the variables. The variable selection results in this paper have provided useful evidence for building suitable data-driven models of Coriolis flowmeters for two-phase flow measurement. It is envisioned that the results from the comparative studies of the three selection methods are applicable to other industrial measurement processes in relation to data-driven models. The performance of the developed data-driven models for flow conditions that are outside the experimental range will be studied and reported in the near future.

References

[1] T. Wang and R. Baker, Coriolis flowmeters: a review of developments over the past 20 years, and an assessment of the state of the art and likely future directions, Flow Measurement and Instrumentation, vol. 40, pp. 99-123, 2014.

[2] N. Keita, Behaviour of straight pipe Coriolis mass flowmeters in the metering of gas: Theoretical predictions with experimental verification, Flow Measurement and Instrumentation, vol. 5, no.4, pp. 289-294, 1994.

[3] M. Anklin, G. Eckert, S. Sorokin and A. Wenger, Effect of finite medium speed of sound on Coriolis mass flowmeter, in Proceedings of the 10th International Flow Measurement Conference, 2000. Available online: http://library.ceesi.com/FLOMEKO_Proceedings.aspx (accessed on 19 November 2016).

[4] J. Kutin and I. Bajsic, An analytical estimation of the Coriolis meter's characteristics based on modal superposition, Flow Measurement and Instrumentation, vol. 12, pp. 345-351, 2002.

[5] J. Kutin, J. Hemp, G. Bobvnik and I. Bajsic, Weight vector study of velocity profile effects in straight-tube Coriolis flowmeters employing different circumferential modes, Flow Measurement and Instrumentation, vol. 16, pp. 375-385, 2005.

[6] J. Kutin, G. Bobovnik, J. Hemp and I. Bajsic, Velocity profile effects in Coriolis mass flowmeters: Recent findings and open questions, Flow Measurement and Instrumentation, vol.17, pp. 349-358, 2006.

[7] S. Enz, J. Thomsen and S. Neumeyer, Experimental investigation of zero phase shift effects for Coriolis flowmeters due to pipe imperfections, Flow Measurement and Instrumentation, vol. 22, pp. 1-9, 2011.

[8] D. Gysling and T. Banach, Accurate liquid phase density measurement of aerated liquids using speed of sound augmented Coriolis meters. ISA EXPO, Houston, USA, 5-7 October 2004.

[9] A. Rieder, W. Drahm and H. Zhu, Coriolis mass flowmeter: On measurement errors in two-phase conditions, in Proceedings of the 13th International Flow Measurement Conference, 2005. Available online: http://library.ceesi.com/FLOMEKO_Proceedings.aspx (accessed on 19 November 2016).

[10] J. Hemp and J. Kutin, Theory of errors in Coriolis flowmeter readings due to compressibility of the fluid being metered, Flow Measurement and Instrumentation, vol. 17, pp. 359-369, 2006.

[11] N. Basse, A review of the theory of Coriolis flowmeter measurement errors due to entrained particles, Flow Measurement and Instrumentation, vol. 37, pp. 107-108, 2014.

[12] J. W. Kunze, R. Storm and T. Wang, Coriolis mass flow measurement with entrained gas, in Proceedings of Sensors and Measuring Systems, pp.1-6, Nuremberg, Germany, 3-4 June 2014.

[13] R. Liu, M. Fuent, M. Henry and M. Duta, A neural network to correct mass flow errors caused by two-phase flow in a digital Coriolis mass flowmeter, Flow Measurement and Instrumentation, vol. 12, no. 1, pp. 53-63, 2001.

[14] B. Safarinejadian, M. Tajeddini and L. Mahmoodi, A new fuzzy based method for error correction of Coriolis mass flow meter in presence of two-phase fluid, in Proceedings of International Conference on Artificial Intelligence and Image Processing, pp. 192-196, Dubai, UAE, 6-7 October 2012.

[15] R. May, G. Dandy and H. Maier, Review of input variable selection methods for artificial neural networks, Chapter 2 in book of "Artificial neural networks- Methodological advances and Biomedical Applications", InTech, ISBN: 978-953-307-243-2, 2011.

[16] G. Bowden, G. Dandy and H. Maier, Input determination for neural network models in water resources applications. Part 1- Background and methodology. Journal of Hydrology, vol. 301, pp. 75-92, 2005.

[17] G. Bowden, H. Maier and G. Dandy, Input determination for neural network models in water resources applications. Part 2. Case study: forecasting salinity in a river. Journal of Hydrology, vol. 301, pp. 93-107, 2005.

[18] R. May, H. Maier, G. Dandy and T. Fernando, Non-linear variable selection for artificial neural networks using partial mutual information, Environmental Modelling & Software, vol. 23, pp. 1312-1326, 2008.

[19] S. Galelli and A. Castelletti, Tree-based iterative input variable selection for hydrological modelling, Water Resources Research, vol. 49, pp. 4295-4310. 2013

[20] S. Galelli, G. Humphrey, H. Maier, A. Castelletti, G. Dandy and M. Gibbs, An evaluation framework for input variable selection algorithms for environmental data-driven models, Environmental Modelling & Software, vol. 62, pp. 33-51, 2014.

[21] X. Li, H. Marier and A. Zecchin, Improvement PMI-based input variable selection approach for artificial neural network and other data driven environmental and water resource models, Environmental Modelling & Software, vol. 65, pp. 15-29, 2015.

[22] A. Sharma, Seasonal to interannual rainfall probabilistic forecasts for improved water supply management: Part1-A strategy for system predictor identification, Journal of Hydrology, vol. 239, no.1-4, pp.232-239, 2000.

[23] S. Galelli and A. Castelletti, Assessing the predictive capability of randomized tree-based ensembles in streamflow modelling, Hydrology and Earch System Sciences, vol. 17, no.7, pp. 2669-2684, 2013.

[24] T. Wang and Y. Hussain, Method for installing and operating a mass flowmeter and mass flowmeter. US 20100326204 A1, 2010.

[25] T. Wang and Y. Hussain, Coriolis mass flow measurement at cryogenic temperatures, Flow Measurement and Instrumentation, vol. 20, pp. 110-115, 2009.

[26] R. Storm, Process for operating a Coriolis mass flow rate measurement device. US 20080011101 A1, 2008.

[27] H. Drucker, C. Burges, L. Kaufman, A. Smola and V. Vapnik, Support vector regression machines, Neural Information Processing System, vol.9, pp. 155-161, 1997.

[28] M. Gevrey, I. Dimopoulos and D. Lek, Review and comparison of methods to study the contribution of variables in artificial neural network models, Ecological Modelling, vol. 160, no.3, pp.249-264, 2003.

[29] M. Gevrey, I. Dimopoulos and D. Lek, Two-way interaction of input variables in the sensitivity analysis of neural network models, Ecological Modelling, vol. 195, no. 1-2, pp. 43-50, 2006.