Heliyon 5 (2019) e02718

Contents lists available at ScienceDirect

Heliyon

www.elsevier.com/locate/heliyon

Research article

Data driven methodology for model selection in flow pattern prediction

Juan Sebastian Hernandez [a,**], Carlos Valencia [a,*], Nicolas Ratkovich [b,*], Carlos F. Torres [c,**], Felipe Muñoz [b,**]

[a] University of los Andes, Department of Industrial Engineering, Cra 1 Este No 19A-40, Bogota, Colombia
[b] University of los Andes, Department of Chemical Engineering, Cra 1 Este No 19A-40, Bogota, Colombia
[c] University of los Andes, Thermal Science Department, Merida, 5101, Venezuela

ARTICLE INFO

Keywords:
Chemical engineering
Two phase flow
Flow pattern
Decision tree
Bagging
Unified flow model

ABSTRACT

The determination of multiphase flow parameters such as flow pattern, pressure drop and liquid holdup, is a very challenging and valuable problem in chemical, oil and gas industries, especially during transportation. There are two main approaches to solve this problem in literature: data based algorithms and mechanistic models. Although data based methods may achieve better prediction accuracy, they fail to explain the two-phase characteristics (i.e. pressure gradient, holdup, gas and liquid local velocities, etc.). Recently, many approaches have been made for establishing a unified mechanistic model for steady-state two-phase flow to predict accurately the mentioned properties. This paper proposes a novel data-driven methodology for selecting closure relationships from the models included in the unified model. A decision tree based model is built based on a data driven methodology developed from a 27670 points data set and later tested for flow pattern prediction in a set made of 9224 observations. The closure relationship selection model achieved high accuracy in classifying flow regimes for a wide range of two-phase flow conditions. Intermittent flow registering the highest accuracy (86.32%) and annular flow the lowest (49.11%). The results show that less than 10% of global accuracy is lost compared to direct data based algorithms, which is explained by the worse performance presented for atypical values and zones close to boundaries between flow patterns.

1. Introduction

Multiphase flows in pipes are complex physical processes which are very common in chemical industry (Picchi and Poesio, 2017). For example, during petroleum transportation, fluids are pushed upwards from oil wells using gas injection, water and steam to improve the production rate of the system. Once the product is extracted, it is taken to processing facilities through a pipeline system, where the complexity of the process depends on the ground conditions of the area that in hilly-terrain carries to a wide range of pipe inclination angles. Accordingly, for design and planning of fluid transportation systems it is very important to correctly estimate and predict multiphase flow parameters such as flow regime, pressure gradient, hold-up, gas and liquid velocities and shear stress.

One of the main properties in the study of two-phase flows is the flow regime, which makes reference to the spatial distribution of the gas and liquid phases during the flow in pipes. The correct estimation of the regimes is fundamental in two-phase analysis, taking into account that design variables such as pressure drop, phase holdup, rate of chemical reaction and others, are strongly related to the registered flow pattern (Pereyra et al., 2012). There are two main approaches to predict the flow regime in a particular configuration. Firstly, there are direct methods based on data analysis that, considering different sets of variables, can estimate the flow type. Taking into account that flow patterns depend on parameters such as pipe inclination, diameter and length, physical properties of the phases, and superficial velocities (Shippen and Bailey, 2012), many machine learning approaches have been developed in the last years to identify flow patterns (e.g. Xie et al. (2004); Al-Naser et al. (2016); Amaya-Gómez et al. (2019)). These methodologies can achieve high predictive performance (e.g. accuracy), however, they are difficult to interpret do not predict simultaneously more two-phase flow characteristics apart of the specific regime.

* Principal corresponding authors.
** Corresponding authors.
E-mail addresses: js.hernadez2249@uniandes.edu.co (J.S. Hernandez), cf.valencia@uniandes.edu.co (C. Valencia), n.rios262@uniandes.edu.co (N. Ratkovich), ctorres@ula.ve (C.F. Torres), fmunoz@uniandes.edu.co (F. Muñoz).

https://doi.org/10.1016/j.heliyon.2019.e02718
Received 30 March 2019; Received in revised form 19 June 2019; Accepted 21 October 2019

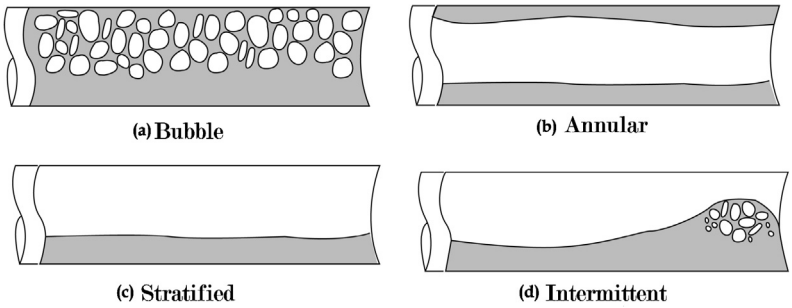

Fig. 1. Illustration of flow patterns. In the top: (a) bubble flow and (b) annular flow. At the bottom: (c) stratified and (d) intermittent flow types.

The second approach to predict flow patterns, is the use of mechanistic models that rely on theoretical equations to derive flow parameters. These methods can predict both, the flow pattern and the two-phase characteristics, but they require the definitions of a set of models named closure relationships. The closure relationships required are: liquid entrainment, gas-liquid interfacial friction, wall friction, mixture friction, slug length, slug holdup, slug drift velocity and slug translational velocity (Zhang et al. (2003a)). The main drawback of these models is the difficulty on the selection of the correct equations that explain the closure relationships.

The main aim of this work is to develop a data-driven methodology to select correctly the closure relationships of each submodel of the Unified Mechanistic Model for steady-state two-phase flow, using dimensionless numbers as input and a database of horizontal flows, and to calibrate the performance of the proposed approach on two-phase flow pattern prediction. The results of this analysis offer a starting point for the study of model selection based on machine learning in fluid dynamics, which in addition to flow pattern prediction, provides a great estimation of other properties such as pressure drop and liquid holdup, key aspects in the design process and real time control strategies. Ultimately, this novel approach would pretend to combine the good predictive performance of the pure data-driven models with the capacity of the mechanistic models to explain much more characteristics of the flow in pipes.

The methodology that is proposed, is built as an ensemble of tree based models (similar to random forest or boosted trees), which is trained based on the ability to explain the correct flow regime (considering a 4 class taxonomy: bubble, annular, stratified and intermittent). The difference with a direct prediction model is that the type of flow is predicted by the set of equations, and the model selects the equations with better prediction at each point of the input space (dimensionless numbers). After implementing the model, an accuracy of 74.84% was registered for flow pattern prediction using the combinations of closure relationships obtained for each point of the test set from the algorithm. As expected, most misclassifications were presented for observations located in the boundaries between similar flow regimes like annular-stratified and bubble-intermittent. For stratified (ST) and intermittent (IT) flow regimes, the method shows high accuracies, whereas for bubble (BF) and annular (A) flow types, the predictive performance was not as good.

The rest of the paper is organized as follows: Section 2 explains multiphase flow patterns, the origin of the data base, the dimensionless numbers included in the study, and the structure and origin of the Unified Mechanistic Model for steady-state two-phase flow. Section 3 describes the machine learning literature background and the proposed methodology. Section 4 shows the results of the implementation based on the metrics established previously. Section 5 develops a brief analysis of the results making special emphasis on misclassification problems. Conclusions and future work are presented in section 6.

2. Background

When gases and liquids flow simultaneously in a pipe, depending on a wide number of variables, the phases can distribute themselves in a variety of configurations. The configuration is determined by the interface distribution, which results in different flow characteristics (Pereyra et al., 2012). The overlapping between flow regimes, especially at the transition zones makes the identification a difficult work and introduces metering errors (Bratland, 2008).

There is not convention in the number of flow patterns in two-phase flow due to overlapping and characterization subjectivity, especially at the transition zones. Four multiphase flow patterns are considered in this work: bubble, intermittent, annular and stratified, which visual characteristics are shown in Fig. 1.

In bubble flow (BF), the gas phase is distributed as discrete bubbles in a continuous liquid phase, for the case of horizontal flow and inclined pipes, the presence of bubbles is higher in the zones which are closer to the top of the pipe (Taitel and Barnea, 2015).

Intermittent flow (IT) is registered when the inventory of liquid in the pipe is distributed in a non-uniform way in the axial direction, for horizontal flow, plugs or slugs of liquid separated by gas zones fill the whole cross-section of the pipe with a stratified liquid layer flowing along the bottom. For annular flow (A), the liquid flows as a continuous film in the border of the pipe due to high velocity of the gas. The last flow regime considered was stratified flow (ST), for which liquid flows along the bottom of the pipe and gas at the top (Taitel and Barnea, 2015).

2.1. Estimation of flow regime and the unified mechanistic model for steady-state two-phase flow

Several studies have used experimental data to estimate a statistical (or machine learning) model that can predict the flow regime given observable inputs. The direct methodologies, do not use theoretical equations that help to explain the phenomenon. Among the most representative studies, Xie et al. (2004) used a transportable artificial neural network for the classification of flow regimes in three phase gas/liquid/pulp fiber systems by using pressure signals as input. Tan et al. (2007) used features extracted from Electrical Resistance Tomography (ERT) data as input of a Support Vector Machine (SVM) algorithm to recognize the flow regime. Ozbayoglu and Yuksel (2012) implemented a back propagation neural network model for flow pattern identification and a regression tree for liquid holdup estimation. Al-Naser et al. (2016) used ANN for flow pattern identification with a preprocessing stage using natural logarithmic normalization, to reduce overlapping between flow patterns. Amaya-Gómez et al. (2019) proposed a Bayesian supervised algorithm with a novel visualization tool for flow pattern maps. Recently, deep learning has provided good results in predicting flow pattern (e.g. Ezzatabadipour et al. (2017)).

From the methods that rely on mechanistic models, the Unified Flow Model proposed in Zhang et al. (2003a), represents one of the most significant advances to better determine the multiphase flow properties in

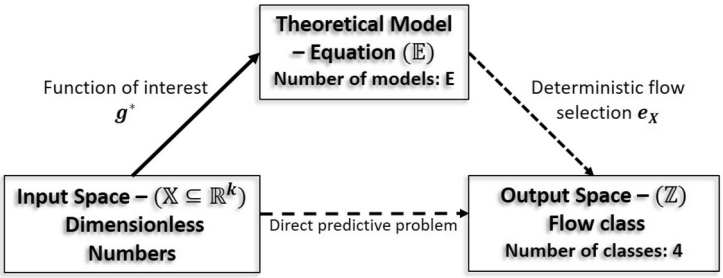

Fig. 2. Graphical description of the equation selection model.

pipes. Their model predicts flow pattern transitions, pressure gradient, liquid holdup and slug characteristics for all angles from -90° to 90° from horizontal. Considering slug flow (a subclass of intermittent flow) shares transition boundaries with all the other flow regimes, the unified model presented by Zhang bases its calculations on the dynamics of this flow regime. With this, equations of slug flow are used to calculate the slug characteristics and also predict transitions from slug flow to other regimes.

The closure relationships included in the Unified Mechanistic Model for steady-state two-phase flow consist basically of empirical (or semi-empirical) equations, established by many authors among different studies for different fluid combinations and property values (appendix A.2), which can be classified in 8 main models: Entrainment Model, Interfacial Friction Model, Wall Friction Model, Mixture Friction Model, Slug Model, Slug Body Holdup Model, Slug Drift Velocity Model and Slug Translational Velocity Model. As mentioned before, these models explain different factors that affect the flow behavior like the wall friction, interface friction and transition boundaries between slug flow and other flow regimes, which are essential for predicting the pattern.

Due to the huge number of possible combinations of equations that can be generated from the 8 subgroups of equations included in the model, 4 equations were selected from the 4 most important models taking into account expert criteria: entrainment models, interfacial friction models, slug body holdup models and slug drift velocity models, leaving the most popular closure relationship in the others.

2.2. Dimensionless numbers

Four dimensionless numbers were estimated from the original variables included in the data base: the modified Froude number (Fr), Weber number (We), Lockhart-Martinelli parameter (Xm) and Eotvos number (Eo). The inclusion of these numbers seeks to establish a global model and not one adjusted to the range of the variables included originally.

The modified Froude number and the Lockhart-Martinelli parameter (which expresses the superficial pressure gradient ratio), regulate the transition boundaries from segregated to non-segregated flow as well as annular to non-annular flow. These numbers have been used by authors like Graham et al. (1999) for estimating the void fraction, and Thome (2003) to deduce the transition from annular flow to intermittent flow.

On the other side, Weber and Eotvos number explain the bubble agglomeration that carries to the differentiation between bubble flow and non-bubble flow (Pereyra et al., 2012). In the case of Weber number, if the surface tension of the fluid decreases (we are greater) bubbles will tend to decrease because of higher momentum transfer between the phases. Authors like Zhao and Rezkallah (1993) used this number to determine the boundary between different regimes like bubble flow, annular flow and slug flow at microgravity conditions.

Eotvos number was used by Clift et al. (2005) to characterize shape regimes for bubbles and drops in unhindered gravitational motion liquids, additionally Ullmann and Brauner (2007) used this non-dimensional number to analyze the relation between pipe diameter and flow pattern transitions.

2.3. Experimental data

An experimental data base collected by Pereyra et al. (2012) , which consists of the most relevant studies on flow pattern prediction was used. The earliest set of data is in Shoham (1982), acquired in 50.8 and 25.4 mm pipe diameters utilizing air-water at atmospheric conditions, which was the first study covering systematically all the inclinations angles, from -90° to +90°. Lin (1985) developed horizontal flow experiments in 25.4 mm and 95.4 mm diameter pipes, varying the superficial gas velocity between 0.8 and 200 m/s. Four years later, Kouba (1986) using air-kerosene, studied slug-flow in a horizontal 76.2 mm diameter pipe. Afterward, Kokal (1987) carried out a study of two-phase flow patterns in horizontal and slightly inclined horizontal flow, using 25.8, 51.2 and 76.3 mm diameter pipes, with air and light oil as working fluid. On the next decade, Wilkens (1997) developed a study for two phase gas-liquid flows at 0°, 1° and 90° angles. Only his data for oil and CO2 were considered in the present study. Later, Manabe (2001) analyzed the relation between pressure and flow pattern for oil-natural gas systems using 0°, 1° and 90° inclinations. Mata et al. (2002) worked on a flow pattern map for high viscosity oil and air in a 50.8 mm horizontal pipe. Three years later, Gokcal (2005) developed a study in a 50.8 mm diameter horizontal pipe for two different liquid viscosities: 181 and 587 mPa.s. The complete origin of the data is explained in detail in the appendix A.1.

The data base consists of a total 37649 experimental data points with information related with fluid properties such as density, viscosity and surface tension, pipe configuration parameters like angle and diameter, and operational conditions like liquid and gas velocity. The previously mentioned variables were used to calculate values of fluid dimensionless numbers, which make the model results scalable to other variable values not included in the data base. 755 points that did not registered flow pattern were deleted.

3. Methodology

It is important to recall that the purpose of the method is not to predict directly a response variable (e.g. flow type) given some input values. That would be the solved by a standard machine learning predictive algorithm (e.g. boosting, support vector machine, artificial neural networks, etc.). Instead, our objective is to solve an inverse problem, that is, we aim to select the combination of theoretical model equations that better predicts the observed flow pattern. This creates a difficult problem given that the main purpose is to estimate a function from the dimensionless numbers space ($\mathbb{X} \in \mathbb{R}^k$) to the theoretical equation models space ($\mathbb{E}$) that contains E number of models resulted from all combinations. However, the observed data correspond to n independent duplets (x_i, z_i), where z_i is the type of flow in observation i. The equation model e_i is not directly observed. Fig. 2 presents a schematic view of the estimation problem.

The objective is to estimate the function $g^* : \mathbb{X} \to \mathbb{E}$ from the data. One option would be to first estimate the direct predictive function $f^* : \mathbb{X} \to \mathbb{Z}$ and then to select for each point $x \in \mathbb{X}$ the model in $\mathbb{E}$ that predicts the expected flow $f^*(x)$. However, this approach has two main

drawbacks: (i) in the majority of points x, there are several models ($e \in \mathbb{E}$) that predict the expected flow class (no unique solution), and (ii) the selection of the flow class that each model does is deterministic, and therefore, the partition generated in $\mathbb{X}$ may be very sensitive and difficult to interpret.

To obtain a smooth and unique estimator $\hat{g}$, we assume that for each x, the flow class selection made by equation e, which we call $e_x \in \mathbb{Z}$, is the result of a soft-thresholding process where $e_x = \arg\max_{z\in\mathbb{Z}} b_z^{e,x}$, and $b_z^{e,x}$ are relative belief weights that are larger when the equation e has a larger propensity to select flow type z for a particular x. For all considerations in this study, the weights $b_z^{e,x}$ for all $e \in \mathbb{E}$ may be re-scaled to a mass probability function P_{e_x}. Therefore, the target function g^* may be defined on each x as

$$g^*(x) = \arg\max_{e\in\mathbb{E}} P_{Z,e_x}\left(Z = e_x | X = x\right) = \arg\max_{e\in\mathbb{E}} \sum_{z\in\mathbb{Z}} P_Z\left(Z = z | X = x\right) b_z^{e,x}. \tag{1}$$

Consequently, a logical estimator $\hat{g}(x)$ corresponds to the equation model e such that a local calculation (on a neighborhood of x) of the proportion of points that e predicts correctly is the largest. With this logic, it is possible to adapt local machine learning methods for classification. We use a modification of the bagged classification trees ensemble algorithm. In the following sections we present a brief explanation of the statistical and machine learning models that inspired the methodology, followed by a detailed description of the developed algorithm. Afterwards, metrics to measure the performance of the method are described.

3.1. Decision tree

Decision trees are statistical models which are learnt from a given training data set to perform a classification or regression task (Aldrich and Auret, 2013). Training data set is made of a response vector $\mathbf{Z} \in \mathbb{R}^n$ and an input matrix $\mathbf{X} \in \mathbb{R}^{n\times k}$ with number of columns equal to number of variables, and both with n observations. When $\mathbf{Z}$ is a categorical variable (also known as factor), the model is called a classification tree, while in the case of a continuous response the model is known as regression tree.

The fundamental point behind the algorithm for building decision tree models is to recursively split the input data space ($\mathbb{X}$) to generate a particular number of regions with a higher purity index for the output (in the case of classification). Purity is a measure for how homogeneous is the classification of the class with the majority of votes in each region. High purity, means that the partition has a good fit of the data. The purity index is established based on the task the model is seeking to solve, establishing measures like the gini index for classification or the least-squares deviation for regression. When the method stops iterating and the data-driven subspaces have been found, simple models or values are fitted to each of the obtained regions. Therefore, a decision tree consists in a set of non-overlapping local models with regions determined by a recursively data driven partition of the training data space.

3.2. Purity index

As mentioned before, literature has established many purity indexes to measure the performance of classification trees, nevertheless, the task treated in this paper is a model selection problem, for which purity indexes have not been developed.

For this study development we used the flow pattern as the response variable for measuring the perfromance of the model selection method, and therefore, our problem can be compared with a classification rather than regression task. The main difference between a classification model and the one developed in this paper, is that for each observation we have as much estimated responses as combinations of model equations implemented, while in a regular classification problem there is only an estimated response per point. Taking into account this, we proposed a new purity index to help us selecting the best combination of equations of the Unified Flow Model based on the flow regime prediction. First, for each observation data point, we calculate $\hat{Y}$, the vector of binary variables that represents for each model combination if the observed flow pattern is predicted correctly or not, for example:

$$Y_i = (1, 0, 1, 1, 0, \ldots, 1) \in \mathbb{R}^E,$$

where E i the number of models considered (for the present case $E = 256$), and $i = 1, \ldots, n$. Each position, is filled with 1 if the selected model predicts correctly the flow pattern or 0 otherwise.

Once the response vector $\hat{Y}_i$ is obtained for each point, it is possible to calculate $\hat{y}_p \in \mathbb{R}^E$ for each region of the partition ($p = 1, \ldots, P$), which consists of the total points classified correctly by each combination of equations, like a multinomial distribution. P stands for the total number of partitions the decision tree has in the current iteration. Therefore,

$$Y_p = \sum_{i\in R_p} Y_i, \tag{2}$$

where the sum is over the points in the region R_p. After this vector is built up for each region, the values were taken and organized decreasingly, to get the combinations of equations ordered from the best to the worst. That is, $Y_p = (y_p^{[1]}, y_p^{[2]}, \cdots, y_p^{[E]})$, where $y_p^{[1]}$ is the largest value of Y_p. Taking the resulting vector of this process it was possible to establish a purity index for model selection

$$\text{Purity index} = \sum_{p\in P} 2y_p^{[1]} - y_p^{[2]}. \tag{3}$$

The purity index we defined seeks to find the best combination of equations for each partition (first term) and differentiate the accuracy of this model from the others (second term). If this term is maximized, we obtain a good estimation $\hat{g}$ according to the definition of the target function (1). That is, a high purity index means that the selected equation model predicts well on region p ($y_p^{[1]}$ is large), and also, the rest of equations do not predict as well (implying that $b_z^{e,x}$ is large for the best model).

3.3. Bagging

The main advantage of decision tree algorithms relates to their capability to fit almost any data distribution, nevertheless, this can be considered also as a problem if we analyze noisy distributions, which can be erroneously over-fitted. The over-fitting can be reduced significantly by restricting the complexity of the tree using a stopping criterion or by combining decision tree models in ensembles with methods such as random forest, boosting and bagging (Aldrich and Auret, 2013).

Bagging, also called bootstrap aggregation, is a general-purpose procedure for reducing the variance of machine learning methods that is frequently used for decision tree ensembles (James et al., 2013). This method starts by generating a group of new data sets with the same size of the original one by sampling randomly with replacement (bootstrapping). Once the new sets are created, the same machine learning algorithm is applied to each one to obtain different data driven models. For classification problems, the final answer for each observation is the most frequently class registered by the resulting models.

3.4. Proposed method

We propose a tree based method, using bagging to reduce overfitting (the variance of the method), for selecting accurately equations from the Unified Flow Model. The method starts by dividing the sample into a training set and a test set for a later validation. Once the data set is divided, bootstrapping is applied to the training data set to generate 100 new data sets by sampling randomly with replacement. For each

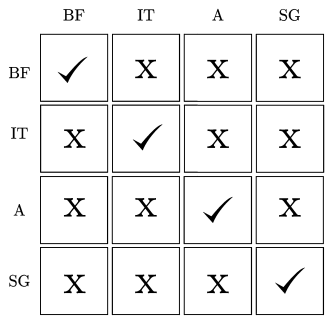

Fig. 3. Confusion matrix structure. The diagonal contains the number of points well classified for each flow type.

data set, a 15 partitions data tree is built up for model selection using the purity index introduced previously. The established tree-growing process is based in popular algorithms such as CART (Breiman et al., 1984) and C4.5 (Quinlan, 2014), which create subspaces by recursively searching for a partition on a single input variable that registers the greatest reduction in impurity for the output variable.

The tree construction starts by setting the vector $\hat{Y}_i$ to each observation and estimating the total correctly classified points by each combination of equations for the database without partitions in $\hat{Y}_p$. Once these vectors are estimated, the purity index is calculated based on the original $\hat{Y}_p$ vector, taking the best two combinations of equations for the non-divided database. In the first partition, the method searches recursively for a division in each single input, calculating the change in the purity index for each possibility, which means, nxm changes in the index are estimated, where n stands for the number of observations and m for the number of input variables, drawing the first partition over the input and observation that registers the greatest change. For the second partition the same procedure is repeated, but with a $\hat{Y}_p$ vector for each subset. This time, the method goes over all the observations for each input variable, searching initially in the first subset and then in the second one, which means that when the method finds the best partition in terms of the purity index, the division is drawn for the value of the input variable only in the subset the observation is located. The previously explained procedure is repeated until the tree reaches 15 partitions and 16 subsets are created. Once the subsets are generated, the best combination of equations of the Unified Flow Model is assigned to each one.

This tree growing process runs simultaneously for each data set generated by bootstrapping, until the 100th tree is generated. After this procedure is over, the method selects the best combination of equations for each observation by taking the most frequent selection over all the trees and estimates the flow pattern taking the values of the input variables.

3.5. Evaluation metrics

Taking into account that the proposed method is a model selection tool, which result is used for estimating the flow pattern, it is not possible to build a ROC curve or estimate an AUC considering that the result is a flow regime and not a probability, which makes it impossible to establish a threshold that can be changed in order to estimate different values of sensitivity and specificity. Based on this, the proportion of correct classifications (accuracy) was established as the most suitable metric for the problem.

For estimating this metric, it is necessary to build a confusion matrix first, which allows a more detailed analysis of the final result. The confusion matrix is a table, where each row represents the classes of the real values and each column represents the class of the predicted values (see Fig. 3).

Each $x_{i,j}$ is the number of observations that corresponds to the class i and are predicted as j. Based on this it is possible to establish the accuracy as shown in the next equation.

$$Accuracy = \frac{\sum_i x_{ii}}{\sum_i \sum_j x_{ij}}$$

A confusion matrix was built for each data set (train and test) for measuring the performance and estimating the accuracy in the model construction and validation. The information registered in the confusion matrices was also used to analyze which flow regime registered the worst classification rate and the causes of the misclassification.

Additionally, the variable importance was estimated for each of the inputs by calculating the change in the purity index generated by the partitions associated with each variable for each tree. This procedure was repeated for all the grown trees to establish the variable importance as the percentage of the total purity index change explained by the input.

4. Results

This section starts by showing the results obtained for one of the decision trees that were grown and its interpretation, followed by the confusion matrix obtained for the test data sets, ending with the results registered for the variable importance. After implementing all the proposed methodology, accuracies of 75.75% and 74.84% were registered for the training and test sets respectively. Also, significant values were obtained for variable importance by three inputs (Xm, Eo and Fr), while one input (We) presented a low importance on the equation selection.

4.1. Decision tree results

Fig. 4 shows the results for a tree grown based on these inputs, where left and right branches represent the positive and negative outcome of the condition respectively.

Taking into account that the proposed methodology is based in bagging (a non-deterministic algorithm), the trees generated can change for each iteration, considering that the data sets created by bootstrapping take observations randomly. The grown trees are made of 15 conditions which take to 16 leafs that return a number related to a combination of closure relationships. The final combination selected by the model, was the most frequent sequence returned by the total constructed trees.

4.2. Comparison of model prediction

As mentioned before, after running the tree based model proposed, over the test set, the accuracy obtained for the flow pattern prediction was 74.84%, with 6903 correctly classified points from the total 9224 experimental observations as shown in Table 1.

The left side of the table shows the total number of points, correctly classified observations and percentage of accuracy per flow pattern. For example, in the case of bubble flow (BF) from the total of 912 experimental points, 494 (54.17%) were successfully predicted. A deeper analysis of the table shows that the best accuracy was registered for intermittent flow (86.32%) while annular flow registered the worst (49.11%).

Right side of the table shows the confusion matrix for test set. The values registered for the annular regime row present additional causes of the poor performance of the classification for this pattern. Looking at the values related with intermittent flow and segregated flow for this row, it can be seen that 340 (23.25%) and 389 (26.60%) were wrongly classified into these groups respectively.

4.3. Variable importance

After building the whole tree based model described before and analyzing its performance on flow pattern classification, it was possible to estimate the importance of the variables as established in section 3.5.

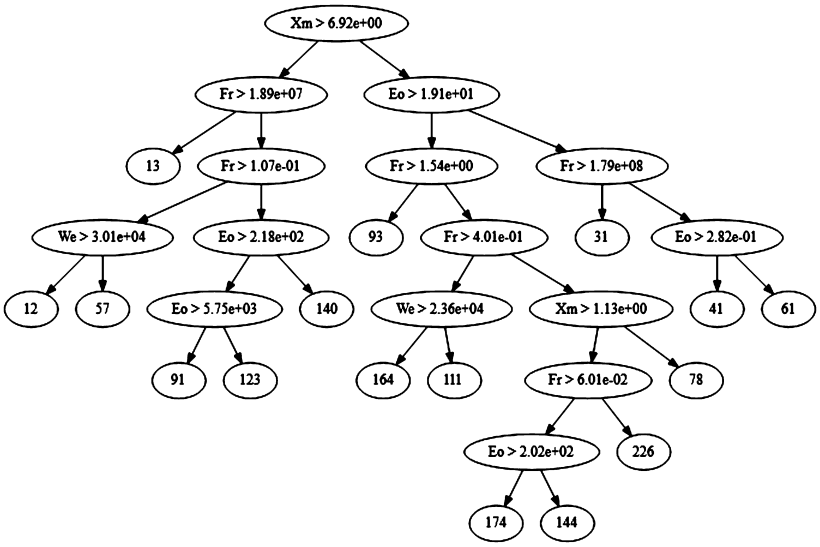

Fig. 4. Estimated decision tree. At each node on split is performed. The left branches represent that the condition is true and the right branches that is false. The circles at the final leafs contain the reference number of the predicted combination of closure relationships.

Table 1
Predictive performance of the model evaluated on the test set. The total number of points for testing was 9224, and the total accuracy was 0.7484. The right part of the table presents the confusion matrix.

Accuracy				Confusion matrix			
	Total	Correct	Accuracy [%]	BF	IT	A	ST
BF	912	494	54.17	494	363	5	50
IT	4489	3875	86.32	251	3875	118	245
A	1462	718	49.11	15	340	718	389
ST	2361	1816	76.91	53	357	135	1816
Total	9224	6903	74.84	813	4935	976	2500

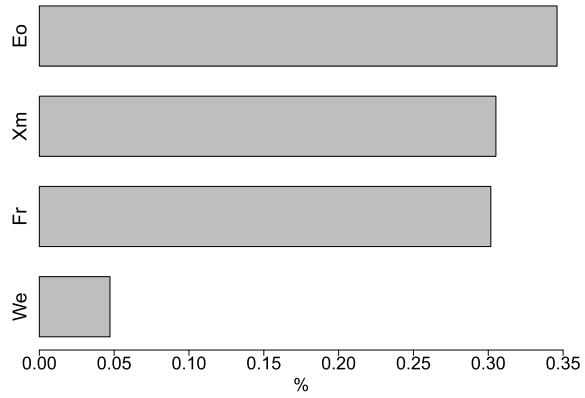

Fig. 5. Variable importance measured as percentage of contribution of each predictor to the objective function.

As shown in Fig. 5 the variables that registered the greatest values for variable importance were Ettvos number, Lockhart-Martinelli parameter and the modified Froud number.

Considering the connection between non-dimensional numbers and flow patterns explained in section 2.3, the values registered for Xm and Fr result consistent with the theoretical relations reported in literature, taking into account both dimensionless numbers establish boundaries between segregated and non-segregated flow, as well as annular and non-annular flow, cases that represent 41.59% of the train set. Following the previous analysis, it results expected to register a low value for Weber number importance considering it provides a boundary between non-bubble and bubble flow, which represents only 10.37% of the complete training set.

Table 2
Predictive performance on the test set of the direct problem with the same data using a random forest algorithm.

Accuracy				Confusion matrix			
	Total	Correct	Accuracy [%]	BF	IT	A	ST
BF	912	667	73.14	667	188	12	45
IT	4489	3945	87.88	155	3945	148	241
A	1462	1157	79.14	12	163	1157	130
ST	2361	1879	79.58	27	326	129	1879
Total	9224	7648	82.91	861	4622	1446	2295

The greatest value for variable importance was registered by Eotvos number. Apart of describing the boundary for bubble flow, this dimensionless number has been used in many fluid studies to characterize the shape of bubbles in liquid flows, which could have explained the transitions from slug flow (subclass of intermittent flow) to the other flow regimes as established in the Unified Flow Model.

4.4. Comparison with direct predictive problem

Considering that the flow regime is not predicted directly from the dimensionless numbers, but through the equations instead, it is expected that the performance (e.g. accuracy) is not as good as the direct problem. One of the questions that appears is how much predictive power is lost by using the indirect method that we propose (it is important to recall that we prefer to use the theoretical equations because from them, it is possible to obtain more relevant information).

Table 2 presents the flow regime predictions for the random forest predictive algorithm when used with the same dimensionless numbers as predictors. The annular (A) and bubble (BF) flows present an improvement with respect to the indirect problem, achieving accuracies of about 80%, proving that this two types of flow are not so well specified by the considered theoretical equations. On the other hand, for intermittent (IT) and stratified (ST) flows the good accuracies remain similar in both methods. The overall accuracy of the direct problem is 82.9%, and therefore, there is a 8.07% of total accuracy that is lost in the pretiction of the flow regime by using the indirect problem and predict with the set of equations.

Some recent models that implemented the direct problem with similar datasets found similar results. For example, Ezzatabadipour et al. (2017) implemented deep learning methods with global accuracies of 81.9%, with bubble flow being the most difficult to predict. Also, Du et al. (2018) used convolutional networks in a similar experiment but

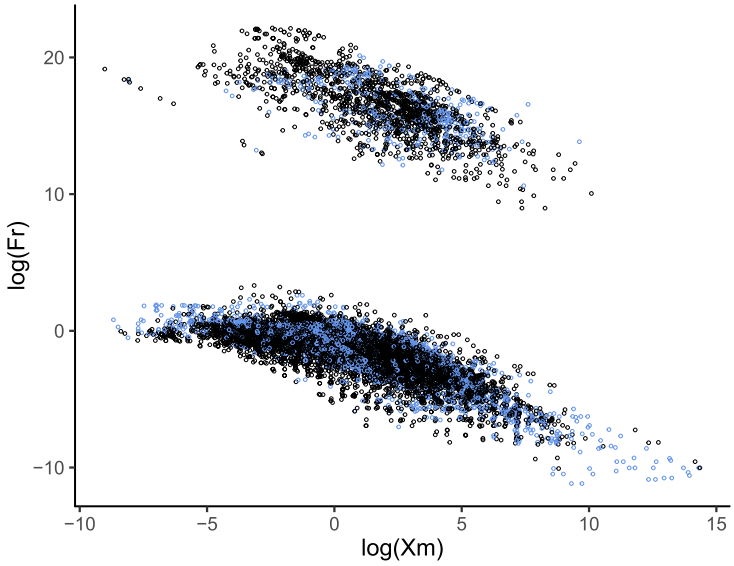

Fig. 6. Graphical analysis of the predictive performance. Modified Froude number vs Lockhart-Martinelli parameter in logarithmic scale. Darker points are well classified.

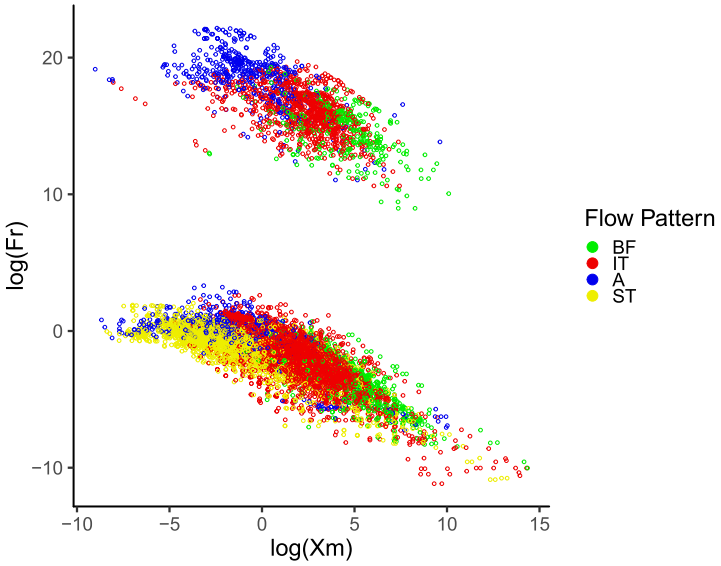

Fig. 7. Real flow pattern - Fr vs Xm in logarithmic scale.

using images instead of dimensionless number obtaining a global accuracy of above 90%.

5. Discussion

5.1. Graphical analysis of model performance

The results shown in Fig. 6 register the correct and misclassified points of the test set, in black and blue respectively, against their Modified Froude number and Lockhart-Martinelli parameter values in logarithmic scale.

As can be seen, most of the points in the test sample have Xm values (in logarithmic scale) from -5 and 10 and Fr mostly between 5 and -10, where correctly classified points were more frequent than misclassified points. However, in the inferior point cloud, is possible to notice some regions where the model selection algorithm did not predict flow pattern properly. Specifically, observations which have Xm values greater than 7.5 with Fr values lower than 0, registered a poor performance in flow pattern prediction with an accuracy of 44.63%. Additional graphs of classification performance against other inputs are presented in appendix B.

Fig. 7 allows a deeper analysis of the causes of misclassification in the lower-right corner of the previous graph. A brief look at the zone mentioned before, shows that most of the points there correspond to observations which registered intermittent flow as real flow regime. The analyzed points can be located also at the lower part of Fig. 8 which shows the predicted flow pattern for each observation against modified Froude number and Weber number values. After crossing the results of both graphs it was possible to conclude that the mentioned missclassified observations were incorrectly classified as bubble regimes, which makes sense taking into account the low values of Weber number presented by these observations.

Other big problem mentioned in section 4.2 was that many annular flow observations were missclassified as segregated regimes. As can be seen in Fig. 7, these flow patterns register similar values of Lockhart-Martinelli parameter and modified Froude number, which makes it difficult for the tree based model to select different combinations of equations in the zone where these flow patterns overlap and also adds

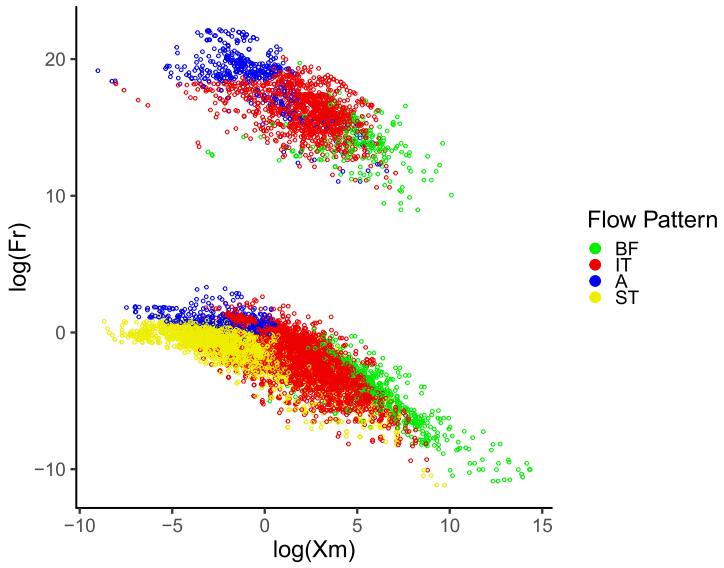

Fig. 8. Predicted flow pattern - Fr vs We in logarithmic scale.

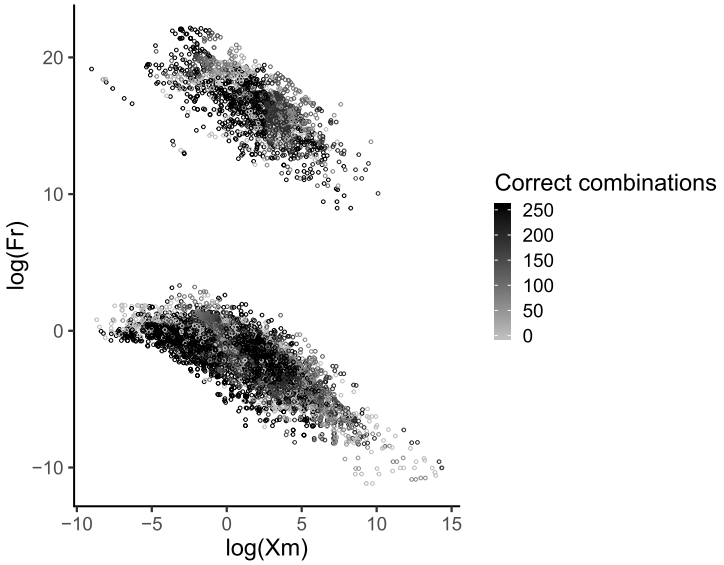

Fig. 9. Number of correct predictions per point.

some difficulty to the flow regime prediction made by the Unified Flow Model.

5.2. Non-covered zones

As was introduced in the previous section, additionally to the error registered by the closure relationship selection method, the 25.16% misclassification rate obtained for the test set can be explained partially by the lack of predictive capacity registered for the possible combinations of closure relationships included in the study. In detail, 1397 (14.95%) observations from the test set, which were mostly located in boundaries between patterns, did not get correct classifications with any of the included combinations, which makes it impossible for the tree based algorithm to classify them correctly.

Fig. 9 allows a deeper analysis of the performance of the selected combinations of closure relationships in the prediction of the flow pattern for the observations of the test set. The graph shows in darker colors the points that obtained a correct classification with a great amount of combinations, while lighter colors stand for those that were missclasified by most of them. A brief look at the zones mentioned in section 5.1, shows that none of the possible combinations of equations established for the study manage to predict correctly the flow pattern for points with very high values of Xm and low values of Fr.

Regarding to the problem of missclasification of annular flows, the cross analysis of Figs. 9 and 7 shows that the cause of this issue was not the combinations of closure relationships set by the tree based algorithm, but the lack of predictive capacity of the selected combinations for this zone. In detail, most of the annular flows of the low point cloud (low values of Fr) presented wrong classifications for almost all the combinations included in the study, which shows that the selected equations of the Unified Flow Model are prone to error in the boundary between annular and segregated flow.

6. Conclusions

A novel methodology is proposed for selecting closure relationships from the different models included in the Unified Flow Model. Considering the results obtained for flow pattern prediction, we conclude

that tree based methods provide an accurate tool for model selection in two-phase flow problems with the advantage of being able to predict also several characteristics such as pressure gradient, holdup, shear stress, etc. A total of 27670 points were used to build the model, which was tested later in a 9224 point set, registering the highest performance for intermittent flow classification (86.32%) and the lowest for annular flow (49.11%). The results show that less than 10% of global accuracy is lost when using the indirect method, which is explained by the worse performance presented for atypical values and zones close to boundaries between flow patterns. For example, several bubble type flow cases were predicted as intermittent. The misclassifications registered by the algorithm in these zones, were the result of the predictions of the UFM analyzed submodels and not by the selection of closure relationships developed by the decision tree algorithm. This means that even if the method is run with more partitions (more combinations of closure relationships are assigned) the performance of the model will not improve significantly.

Based on the lack of predictive capacity registered by the combinations of equations included in the study for the boundaries between similar flow patterns, we recommend for future work and extensions, including more proposed closure relationships for each of the 8 models that conform the UFM, making emphasis in the 4 models which the study focused on. In addition, include more dimensionless numbers after optimizing the UFM prediction capacity to get a better understanding of the relation between the assigned closure relationships and the studied variables. Furthermore, once these problems are solved, the model selection performance can be tested in the prediction of other properties such as pressure drop and liquid holdup.

For the present study, the studied flow was measured on horizontal pipes. For future research, vertical and pipes with inclinations are planned to be added for further evidence and analysis.

Declarations

Author contribution statement

Juan Sebastian Hernandez: Conceived and designed the experiments; Performed the experiments; Analyzed and interpreted the data; Wrote the paper. Carlos Felipe Valencia Arboleda, Nicolas Ratkovich: Conceived and designed the experiments; Analyzed and interpreted the data; Wrote the paper. Felipe Munoz Giraldo: Conceived and designed the experiments; Contributed reagents, materials, analysis tools or data. Carlos Torres: Analyzed and interpreted the data; Contributed reagents, materials, analysis tools or data.

Funding statement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Competing interest statement

The authors declare no conflict of interest.

Additional information

Supplementary content related to this article has been published online at https://doi.org/10.1016/j.heliyon.2019.e02718.

Acknowledgements

The authors thank to Tulsa University Fluid Flow Projects (TUFFP) for providing the Unified Model software used on this research.

References

Al-Naser, M., Elshafei, M., Al-Sarkhi, A., 2016. Artificial neural network application for multiphase flow patterns detection: a new approach. J. Pet. Sci. Eng. 145, 548–564.

Aldrich, C., Auret, L., 2013. Unsupervised Process Monitoring and Fault Diagnosis with Machine Learning Methods. Springer.

Amaya-Gómez, R., López, J., Pineda, H., Urbano-Caguasango, D., Pinilla, J., Ratkovich, N., Muñoz, F., 2019. Probabilistic approach of a flow pattern map for horizontal, vertical, and inclined pipes. Oil Gas Sci. Technol. (Revue d'IFP Energies nouvelles) 74, 67.

Bratland, O., 2008. Update on commercially available flow assurance software tools: what they can and cannot do? In: 4th Asian Pipeline Conference & Exhibition. Kuala Lumpur, Malaysia.

Breiman, L., Friedman, J., Stone, C.J., Olshen, R.A., 1984. Classification and Regression Trees. CRC Press.

Clift, R., Grace, J.R., Weber, M.E., 2005. Bubbles, Drops, and Particles. Courier Corporation.

Du, M., Yin, H., Chen, X., Wang, X., 2018. Oil-in-water two-phase flow pattern identification from experimental snapshots using convolutional neural network. IEEE Access 7, 6219–6225.

Ezzatabadipour, M., Singh, P., Robinson, M.D., Guillen-Rondon, P., Torres, C., 2017. Deep learning as a tool to predict flow patterns in two-phase flow. arXiv preprint, arXiv: 1705.07117.

Gokcal, B., 2005. Effects of High Oil Viscosity on Two-Phase Oil-Gass Flow Behavior in Horizontal Pipes. Ph.D. thesis, University of Tulsa.

Graham, D., Kopke, H., Wilson, M., Yashar, D., Chato, J., Newell, T., 1999. An Investigation of Void Fraction in the Stratified/Annular Flow Regions in Smooth, Horizontal Tubes. Technical Report, Air Conditioning and Refrigeration Center. College of Engineering. University of Illinois at Urbana-Champaign.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An Introduction to Statistical Learning, vol. 112. Springer.

S.L., Kokal, 1987. An Experimental Study of Two Phase Flow in Inclined Pipes. Chemical and Petroleum Engineering, University of Calgary.

Kouba, G.E., 1986. Horizontal Slug Flow Modeling and Metering. Technical Report, Tulsa Univ., OK, USA.

P.Y., Lin, 1985. Flow Regime Transitions in Horizontal Gas-Liquid Flow. University of Illinois at Urbana-Champaign.

Manabe, R., 2001. A Comprehensive Mechanistic Heat Transfer Model for Two-Phase Flow With High-Pressure Flow Pattern Validation. Ph.D. Dissertation U. Tulsa, Tulsa.

Mata, C., Pereyra, E., Trallero, J., Joseph, D., 2002. Stability of stratified gas-liquid flows. Int. J. Multiph. Flow 28, 1249–1268.

Ozbayoglu, A.M., Yuksel, H.E., 2012. Analysis of gas–liquid behavior in eccentric horizontal annuli with image processing and artificial intelligence techniques. J. Pet. Sci. Eng. 81, 31–40.

Pereyra, E., Torres, C., Mohan, R., Gomez, L., Kouba, G., Shoham, O., 2012. A methodology and database to quantify the confidence level of methods for gas–liquid two-phase flow pattern prediction. Chem. Eng. Res. Des. 90, 507–513.

Picchi, D., Poesio, P., 2017. Uncertainty quantification and global sensitivity analysis of mechanistic one-dimensional models and flow pattern transition boundaries predictions for two-phase pipe flows. Int. J. Multiph. Flow 90, 64–78.

Quinlan, J.R., 2014. C4.5: Programs for Machine Learning. Elsevier.

Shippen, M., Bailey, W.J., 2012. Steady-state multiphase flow: past, present, and future, with a perspective on flow assurance. Energy Fuels 26, 4145–4157.

Shoham, O., 1982. Flow Pattern Transition and Characterization in Gas-Liquid Two-Phase Flow in Inclined Pipes. Ph.D. Dissertation.

Taitel, Y., Barnea, D., 2015. Encyclopedia of Two-Phase Heat Transfer and Flow I: Fundamentals and Methods.

Tan, C., Dong, F., Wu, M., 2007. Identification of gas/liquid two-phase flow regime through ert-based measurement and feature extraction. Flow Meas. Instrum. 18, 255–261.

Thome, J.R., 2003. On recent advances in modeling of two-phase flow and heat transfer. Heat Transf. Eng. 24, 46–59.

Ullmann, A., Brauner, N., 2007. The prediction of flow pattern maps in minichannels. Multiph. Sci. Technol. 19.

Wilkens, R.J., 1997. Prediction of the Flow Regime Transitions in High Pressure, Large Diameter, Inclined Multiphase Pipelines. Ph.D. thesis, Ohio University.

Xie, T., Ghiaasiaan, S., Karrila, S., 2004. Artificial neural network approach for flow regime classification in gas–liquid–fiber flows based on frequency domain analysis of pressure signals. Chem. Eng. Sci. 59, 2241–2251.

Zhang, H.Q., Wang, Q., Sarica, C., Brill, J.P., 2003a. A unified mechanistic model for slug liquid holdup and transition between slug and dispersed bubble flows. Int. J. Multiph. Flow 29, 97–107.

Zhao, L., Rezkallah, K., 1993. Gas-liquid flow patterns at microgravity conditions. Int. J. Multiph. Flow 19, 751–763.