A machine learning approach on two-phase flow characterization and calculation based on a large experimental dataset

Faller, A. C., Paulo, P. H. C, Vieira, S. C., Fabro, A. T., Castro, M.S.

Email: faller@petrobras.com.br, pedro.cardosopaulo@petrobras.com.br, saonvieira@petrobras.com.br, fabro@unb.br, mscastro@unicamp.br

INTRODUCTION

Two-phase flow calculations for pressure gradient and void fraction and flow pattern characterization are performed using empirical correlations or mechanistic modelling. We offer a machine learning data-driven approach to obtain such correlations. Our work is built upon a proprietary experimental dataset containing approximately 22k experimental data points featuring the fluid velocities: liquid and gas superficial velocities (v_SL, v_SG); the fluid properties: liquid and gas densities (ρ_L, ρ_G), liquid and gas viscosities (μ_L, μ_G) and the nterfacial tension (σ_I); the pipe section geometry: roughness (ε), diameter (D) and inclination (θ); and the observed parameters: pressure gradient (dP/dL), void fraction (α) and flow pattern (dispersed, intermittent, annular or stratified). The goal is to obtain black box models that take the fluid properties and the geometry data as inputs and predict the pressure gradient, the void fraction, and the flow pattern.

METHOD

To obtain regressors for the continuous outputs (pressure gradient and void fraction) and classifiers for the discrete output (flow pattern), we used the following scikit-learn supervised learning algorithms: K-Nearest-Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), AdaBoost (AB), Gradient Boosting (GB) and Multi-Layer-Perceptron (MLP). The train/test split is 80%/20% and stratified. The regressors are evaluated with the fitness (R^2) score and the classifiers with the F1 score.

As input features to the models, besides the dimensional features contained in the dataset, we tested combinations with sets of dimensionless and augmented features, both inspired in dimensionless numbers used by empirical and mechanistic modeling techniques found in literature. Some of the dimensionless features used include: no-slip holdup (λ), relative roughness (ε/D), dimensionless pressure gradient (G), liquid and gas velocity numbers (N_{LV} and N_{GV}), liquid and gas viscosity numbers (N_L and N_G), diameter number (N_D), mixture Reynolds number Re_M), mixture, liquid and gas Froude number ($1/Fr_M$, $1/Fr_L$ and $1/Fr_G$) and several others. The augmented features combine the dimensionless ones and include $G \log \lambda$, $\lambda^2 \log Re_M$, $\sqrt{Fr_L} \log N_L$, $\cos^2 \lambda^2$, and so on. As part of the feature engineering process, all the features are ranked according to the impurity-based feature importance after fitting a Random Forest model.

RESULTS

Despite the dataset being large, it is somewhat biased towards the experiments' characteristics, which were performed mostly (more than 50%) on horizontal or near horizontal pipes, with fluids such as water, oil, kerosene, glycerol, air, among others. This bias is noticeable when a Random Forest model is fitted only with the dimensional input features (superficial velocities, fluid properties and pipe geometry) and the impurity-based feature importance reveals that the fluid properties have nearly zero importance, compared to the superficial velocities. This shows that a model trained only with these dimensional features won't be able to represent any physics related to the two-phase flow, which is in fact dependent of the fluid properties. Repeating the feature importance analysis, but with the dimensionless features set, though some other characteristics start to become relevant, the most important features are still highly dependent on the superficial velocities, which are: no-slip holdup (λ), dimensionless pressure gradient (G) and liquid velocity number (N_{LV}). With the augmented features set, the feature importance analysis also showed that the most important ones are also related to the superficial velocities: G log λ, mixture Reynolds number (Re_M) and N_{LV} log N_D. However, the addition of such dimensmodel and augmented features should increase the model's generalization capability, as they convey some physical meaning into the model.

Figure 1 shows the F1-score comparison between the beforementioned machine learning algorithms with each feature set as inputs, for the flow pattern classification. The best performers are the tree--based ones (DT, RF and GB). However, GB shows less degradation when evaluating the test dataset, because it is more robust and less susceptible to overfitting. Therefore, this model was selected for further analysis and hyperparameter optimization.

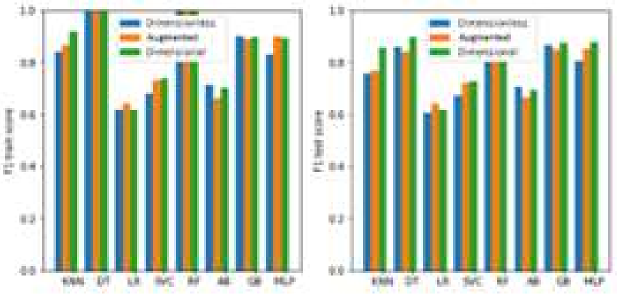

Figure 1 - Machine learning algorithms comparison

Evaluating the Gradient Boosting model with all the input sets combinations and tweaking hyperparameters (such as limiting the trees depth and increasing the minimum samples per leaf) to increase the generalization power, the best feature sets and results found for each output are: **Void fraction (α)**: only the dimensional inputs are enough to obtain a good fitness (R^2_{train}= R^2_{test}=0.99, and R^2_{xval} E[0.98,0.99]). **Pressure gradient** (dP/dL): the best fitness was found with the union of the dimensional and the dimensionless features sets (R^2_{train}=1.0, R^2_{test}=0.98 and R^2_{xval} E[0.85,0.99]). **Flow pattern:** the best F1-score is also obtained with the union of the dimensional and the dimensionless features sets ($F1_{train}$=1.0, $F1_{test}$=0.93 and $F1_{xval}$ E [0.93,0.94]). Figure 2 show the results for all the models. In (a), (b) and (c), the good fitness for the void fraction and pressure gradient can be visualized. In (d), the model's generated flow pattern map resembles to the one generated by Barnea's method in (e).

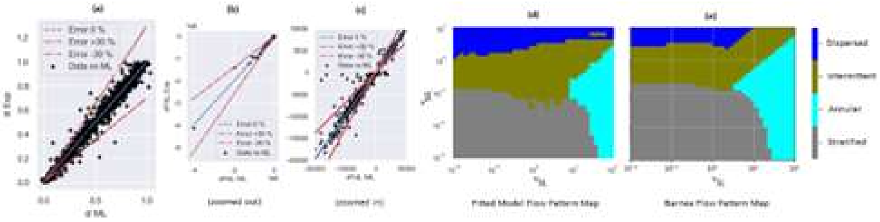

Figure 2 - Results: (a) Void Fraction; (b) Pressure Gradient; (c) Zoomed in Pressure Gradient;
(d) Model's Flow Pattern Map; (e) Barnea's Flow Pattern Map.

CONCLUSION

The presented machine learning approach yields robust models that can accurately predict the void fraction, the pressure gradient, and the flow pattern in two-phase flow, within the dataset parameters range. It is not expected that the models can extrapolate to different pipe diameters and fluid properties, due to the lack of physics knowledge embedded in the model, which is limited to the engineered features, that improve generalization and robustness, but won't grant extrapolation capabilities.