energies

MDPI

Article

Assessing Steady-State, Multivariate Experimental Data Using Gaussian Processes: The GPExp Open-Source Library

Sylvain Quoilin [1,2,*] and Jessica Schrouff [3,4]

1 Energy Systems Research Unit (B49), University of Liège, Sart-Tilman, Liège 4000, Belgium
2 Institute for Energy and Transport, European Commission DG Joint Research Centre, P.O. Box 2, Petten NL-1755 ZG, The Netherlands
3 Laboratory of Behavioral and Cognitive Neuroscience, Stanford University, 300 Pasteur Drive, Stanford, CA 94305, USA; jschrouf@stanford.edu
4 Department of Computer Science, University College London, Gower Street, London WC1E 6BT, UK
* Correspondence: squoilin@ulg.ac.be; Tel.: +32-4-366-4822

Academic Editor: Enrico Pontelli
Received: 25 March 2016; Accepted: 26 April 2016; Published: 30 May 2016

Abstract: Experimental data are subject to different sources of disturbance and errors, whose importance should be assessed. The level of noise, the presence of outliers or a measure of the "explainability" of the key variables with respect to the externally-imposed operating condition are important indicators, but are not straightforward to obtain, especially if the data are sparse and multivariate. This paper proposes a methodology and a suite of tools implementing Gaussian processes for quality assessment of steady-state experimental data. The aim of the proposed tool is to: (1) provide a smooth (de-noised) multivariate operating map of the measured variable with respect to the inputs; (2) determine which inputs are relevant to predict a selected output; (3) provide a sensitivity analysis of the measured variables with respect to the inputs; (4) provide a measure of the accuracy (confidence intervals) for the prediction of the data; (5) detect the observations that are likely to be outliers. We show that Gaussian processes regression provides insightful numerical indicators for these purposes and that the obtained performance is higher or comparable to alternative modeling techniques. Finally, the datasets and tools developed in this work are provided within the GPExp open-source package.

Keywords: Gaussian processes; experimental data; outlier; surface response; kriging; regression; feature selection

1. Introduction

Experimental data are widely used in different energy research fields, e.g., to assess and compare the performance of different processes, to point out the main sources of losses or to calibrate and validate models.

These experimental data, however, are subject to different sources of uncertainty, noise and errors, such as sensor malfunctions, transient phenomena, operator misuse of the test bench, noise in the data acquisition chain, unaccounted for external influences, *etc*. It is therefore of primary importance to assess their quality, by evaluating the level of noise, the presence of outliers or to compute the variability of the measured variables with respect to the externally-imposed operating conditions. This allows for example rejecting non-relevant (input) variables or outliers in the data. These tasks are not straightforward, especially if the data are sparse and multivariate.

This work therefore aims at answering the following questions:

- What is the most likely multidimensional de-noised response surface corresponding to the experimental data?
- What is the sensitivity of the dependent variable with respect to the measured inputs?
- Which inputs are relevant to predict a selected output?
- With what accuracy can the output be predicted with a given set of inputs?
- Which observations are likely to be outliers?

Gaussian processes, also known as kriging regression, provide an automatic and robust framework to perform multivariate regressions and, thereby, constitute an interesting tool to explore high-dimensional data [1]. Their Bayesian formulation allows predicting the variable of interest for new/unseen data points and provides coherent estimates of predictive uncertainty. The method reduces predictive confidence when extrapolating away from the data points: if the data density is locally high, the variance is small; on the opposite side, if the density is low, the variance is larger, leading to more distant confidence boundaries. Furthermore, the method is highly flexible and can accommodate a range of covariance structures, including non-linear relationships, and delivers state of the art prediction performance. It has been successfully applied to a variety of domains, including geostatistics, economics and cognitive neuroscience [2].

It can be shown that Gaussian processes perform better than traditional linear regression in multiple ways [1]. They are less subject to over-fitting and to the Runge phenomenon and exhibit a better behavior outside of the fitting range. They build an underlying smooth representation of the data, which can then be compared to the actual distribution of the experimental data and provide a measure of the noise.

This paper explores the suitability of Gaussian processes to answer the previously formulated questions for multivariate steady-state datasets. A brief overview of the main features of the method is first presented, followed by a description of the proposed tool. The derived methodology is then tested over two datasets from the open literature.

2. Gaussian Processes as a Data Analysis Tool

2.1. Gaussian Processes' Regressions

This section provides a brief explanation of Gaussian processes' regression. The interested reader can refer to [1] for a more detailed description.

When performing a regression, the goal is to find a function f that maps each input x to the variable of interest, also known as target y. The type of function f is usually set *a priori* by the user. Furthermore, hyperparameters, such as the order of a polynomial fit, also need to be fixed *a priori*. Increasing the complexity of f can in most cases lead to an excellent fit of the data. However, too complex models also fit the noise in the data (*i.e.*, they "over-fit"), which is not desirable [3].

Gaussian processes, on the contrary, are based on the Bayesian analysis of the standard linear model:

$$f(\boldsymbol{x}) = \boldsymbol{x}^{\mathrm{T}}\boldsymbol{w} \tag{1}$$

$$\boldsymbol{y} = f(\boldsymbol{x}) + \boldsymbol{\varepsilon} \tag{2}$$

with $\boldsymbol{w} \in R^m$ the vector of parameters (weights) of the model and $\boldsymbol{\varepsilon}$ an error term distributed according to a Gaussian distribution with zero mean and variance σ_n^2.

Let $X \in R^{n \times m}$ be the matrix concatenating all n data points and $\boldsymbol{y} \in R^n$ be their corresponding targets. Bayes' rule allows computing the probability density of the observations given the model parameters $p(\boldsymbol{y}|X, \boldsymbol{w})$, also known as the likelihood and inferred on the model parameters:

$$p(\boldsymbol{w}|\boldsymbol{y}, X) = \frac{p(\boldsymbol{y}|X, \boldsymbol{w})p(\boldsymbol{w})}{p(\boldsymbol{y}|X)} \tag{3}$$

Exact inference is possible when using Gaussian priors over the model weights: $p(\boldsymbol{w}) \sim N(0, \Sigma_p)$ [1]. Finally, the predictive distribution of a new/unseen sample x^* can be estimated from $p(f^*|x^*, X, \boldsymbol{y})$, which is also Gaussian, allowing one to make predictions.

A simple idea to extend the standard linear model is to project the inputs into a high dimensional space using a set of basis function, $\boldsymbol{\phi}(x)$. For example, a polynomial model can be obtained by projecting x onto $\boldsymbol{\phi}(x) = (1, x, x^2, ...)^{\mathrm{T}}$. As long as the basis functions are independent of the model parameters, the model is linear in the model weights and, hence, analytically tractable. $\boldsymbol{\phi}(x)$ can then replace any occurrence of x in the standard linear model. This leads to the definition of the covariance function, or kernel:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^{\mathrm{T}} \Sigma_{\mathrm{p}} \boldsymbol{\phi}(\boldsymbol{x}') \tag{4}$$

where $\boldsymbol{x}$ and $\boldsymbol{x}'$ are the input vectors of two different observations.

Replacing inner products in input space by $k(\boldsymbol{x}, \boldsymbol{x}')$ is referred to as the "kernel method" [4,5]. It shifts the importance of the feature space to the kernel, which is central in Gaussian Processes (GP) modeling (see further).

In the Gaussian processes approach, the priors are defined over the function f instead of on the model parameters $\boldsymbol{w}$ (this is referred to as the "function space"). f is assumed to follow a Gaussian process, *i.e.*, a multivariate Gaussian distribution:

$$f(\boldsymbol{x}) \sim N(\mu(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x}')) \tag{5}$$

with $\mu(\boldsymbol{x})$ the mean latent function and $k(\boldsymbol{x}, \boldsymbol{x}')$ the covariance function.

This formulation highlights the central role of the kernel, which defines the distribution over functions. The kernel is chosen based on *a priori* information, although this choice is qualitative only. Hereunder are some examples of commonly-used kernels:

The homogeneous linear kernel:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^{\mathrm{T}} \cdot \boldsymbol{x}'; \tag{6}$$

The squared exponential (SE) isotropic kernel:

$$k(\boldsymbol{x}, \boldsymbol{x}') = \sigma_{\mathrm{f}}^2 \cdot \exp(-\frac{(\boldsymbol{x} - \boldsymbol{x}')^2}{2 \cdot l^2}) \tag{7}$$

with σ_{f} and l the hyperparameters of the kernel.

In this work, it is particularly interesting to know which variables are the most relevant to build the model. This can be achieved using an automatic relevance determination (ARD) [6] kernel, which is therefore the one selected for this analysis. This kernel is similar to the SE kernel (Equation (7)), except that a distinct parameter l is defined for each variable $d = 1, ..., D$. This hyperparameter represents the "length-scale" of the kernel in each direction. In the ARD kernel, if a length-scale is large, a long distance needs to be traveled before seeing significant changes in the corresponding direction. Therefore, the corresponding variable does not have much influence on the model.

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sigma_{\mathrm{f}}^2 \cdot \exp(-\frac{(\boldsymbol{x}_i - \boldsymbol{x}_j)^2}{2 \cdot l_i^2}) \tag{8}$$

The hyperparameters σ_{f} and l_i are optimized based on the marginal likelihood $p(\boldsymbol{y}|X)$ (Equation (3)), which takes into account both the goodness-of-fit and the complexity of the model.

2.2. Model Performance

This section describes the computation of different numerical indicators, allowing one to assess the performance of the model. The complete flow chart is provided in Figure 1, and each aspect of the analysis is discussed in the next sub-sections.

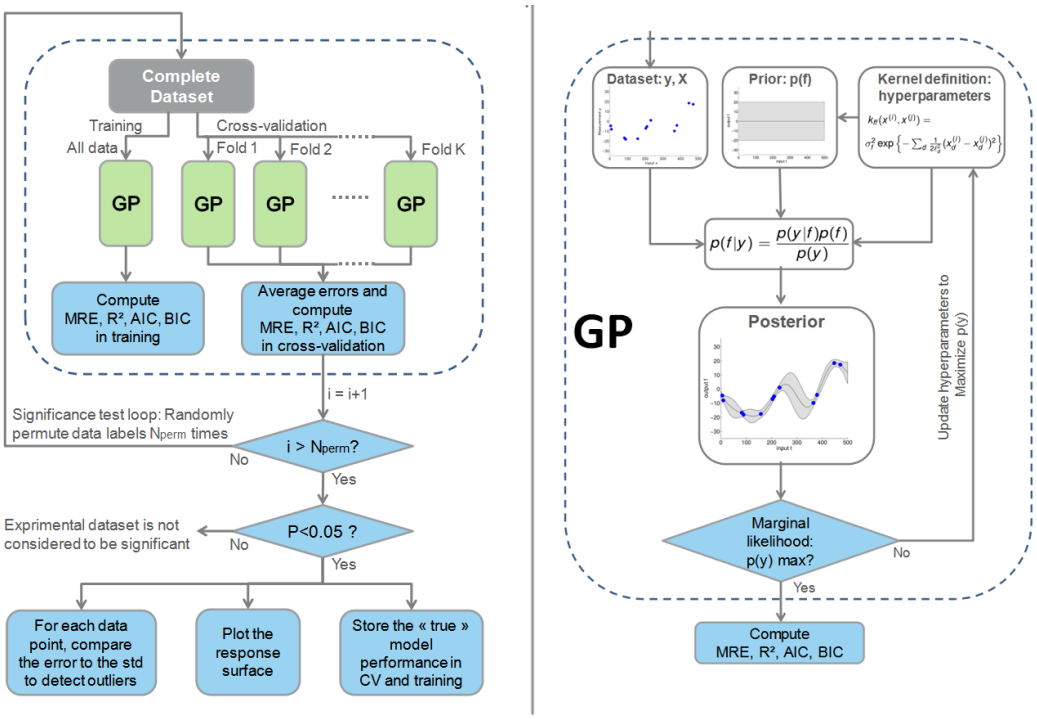


Figure 1. Schematic view of the analysis.

2.2.1. Goodness-of-Fit and Complexity

Once a regression over the data has been performed, its quality should be assessed with proper numerical indicators. The most common indicators are based on the difference (in absolute value) between the predicted target $f(\boldsymbol{x})$ and the "true" target $\boldsymbol{y}$ (*i.e.*, the measured value). Other indicators are based on the marginal likelihood.

Our goal being to compare the model performance across different datasets, all measured values (inputs and target) are normalized between zero and one. In the following, scaled features are indicated by X' for the inputs and $\boldsymbol{y}'$ for the output.

The first performance indicator selected for this work is the mean absolute error (MAE). Note that it is defined with the normalized values:

$$\text{MAE} = \frac{1}{N} \cdot \sum_{i=1}^{N} |y'_i - f(X'_i)| \tag{9}$$

where $f(X')$ is the GP model prediction for the inputs X'. The second performance indicator is the normalized root mean square error (RMSE), which gives a larger weight to large errors. It is calculated by:

$$\text{RMSE} = \sqrt{\frac{1}{N} \cdot \sum_{i=1}^{N} (y'_i - f(X'_i))^2} \tag{10}$$

The coefficient of determination (R^2) is not scale dependent and does not need to be normalized. It reflects the correlation between the predicted outputs $f(\boldsymbol{x})$ and the targets $\boldsymbol{y}$. It is calculated by:

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y'_i - f(X'_i))^2}{\sum_{i=1}^{N} (y'_i - \bar{y}')^2} \tag{11}$$

where $\bar{y}'$ is the average of the measured outputs.

Finally, the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) are defined by:

$$\text{AIC} = -2 \cdot ln(p(\boldsymbol{y}|X)) + 2 \cdot q \tag{12}$$

$$\text{BIC} = -2 \cdot ln(p(\boldsymbol{y}|X)) + q \cdot ln(n) \tag{13}$$

These two indicators depend on the natural logarithm of the marginal likelihood ($\ln(p)$), on the number of hyperparameters q (corresponding to the number of length-scales evaluated and sigma, *i.e.*, $m + 1$) and on the number of data points n. They account for both the goodness-of-fit and the complexity of the model [7].

2.2.2. Generalization Ability

In a machine learning setting, the model performance is computed as the ability of the model to predict the target y for a new/unseen sample $\boldsymbol{x}^*$, *i.e.*, in terms of its generalization ability. This is usually performed by dividing the dataset into two: the training set, on which the model is built, and the test set, on which the model performance is assessed.

Splitting the dataset into the training and test sets can be performed in multiple ways. However, due to the scarcity of data samples, a commonly=used approach is cross-validation: the data are partitioned into a fold (e.g., the first 80% for training and the last 20% for testing). This partition then circles across the dataset, defining as many folds as possible by putting each point in the test set once (e.g., fold two would consist of the last 80% for training and of the first 20% for testing, with a maximum of five folds). For each fold, a model is built and then tested, its model performance being assessed through MAE, RMSE and/or R^2. The final model performance is computed as the average across folds. It should be noted that the cross-validation significantly impacts the computational expenses, since the analysis has to be repeated according to the number of folds (e.g., five times in the previous example).

2.2.3. Significance of the Experimental Data

To assess the significance of model performance (*i.e.*, what is the probability that those values of MAE, RMSE and R^2 could be obtained by chance?), a non-parametric permutation test is performed, as illustrated in Figure 2. In this setting, the null hypothesis states that the labels (*i.e.*, the correspondence between each sample of inputs $\boldsymbol{x}$ and each output y) do not bring any information, such that any random permutation of the targets y could lead to the same model performance. For each random permutation of the targets, a model is built and its performance computed (using the same cross-validation as for the "true" target vector). These values are then compared to the "true" model performance, allowing one to associate a p-value to the "true" MAE, RMSE and R^2 values. The p-value is calculated using the MAE in cross-validation by:

$$p_{\text{MAE}} = \frac{N_{\text{perm | mae<mae}_{\text{ref}}}}{N_{\text{permutations}}} \tag{14}$$

where the numerator is the number of permutations whose computed MAE value is smaller than the MAE value of the "true" target. In this work, the model performance is considered significant if $p < 0.05$.

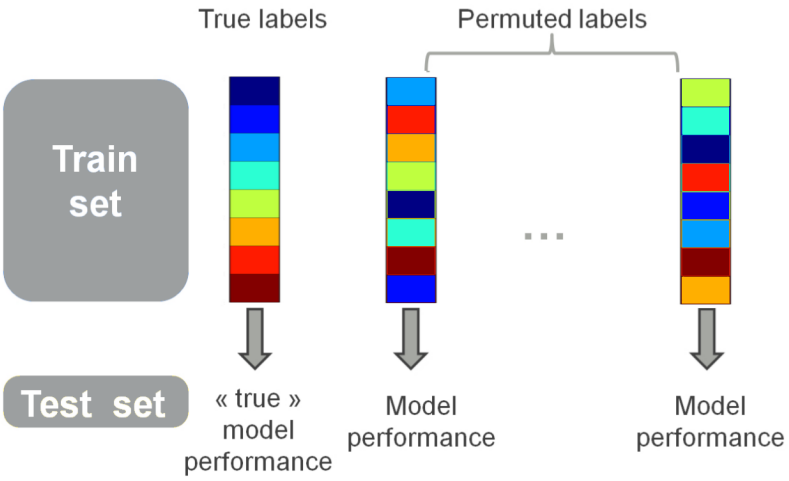


Figure 2. Illustration of the non-parametric permutation testing.

2.3. Illustration of the Method for Simulated Univariate Data

In this section, the different criteria and their behavior are illustrated in situations that can typically arise in test-bench experiments, such as noise in the measurements, outliers, low number of data points, non-uniform sampling of the input space, *etc*. The considered univariate data are generated from the second-order polynomial:

$$\boldsymbol{y} = -1.5 \cdot X^2 + 3 \cdot X + 4 \tag{15}$$

with $X \in R^{n \times 1}$, $n = 20$ points randomly drawn between zero and one.

2.3.1. Effect of Noise

Uncertainties in the measurements, as well as the influence of external perturbations not accounted for in the inputs can be simulated by adding random noise to Equation (15). Gaussian noise was added to the simulated data, with the standard deviation of the noise varied from 0.02 to 1.12 (fixed arbitrarily). The addition of noise should lead to increases in MAE (both for model fitting and in cross-validation), in AIC and in BIC. Furthermore, at high levels of noise, it is expected that the model cannot predict correctly, leading to non-significant values of the MAE in cross-validation.

Figure 3 shows the simulated noisy data and the fitted GP model. At high noise levels, the model still fits the data smoothly. However, the uncertainty of the model, represented by its confidence boundaries at 95%, increases with the level of noise. Figure 3 further shows that the models have similar shapes for all levels of noise. When fitting (in a least-squares sense) the same data with a polynomial, the results are similar if a quadratic fit is selected, which is expected, since the law used to generate the targets is quadratic (Equation (15)). However, this information is unknown *a priori* for real datasets, and changing the order of the polynomial model leads to completely different regressions, with a clear over-fitting pattern for the eighth order polynomial (Figure 3). This highlights a key advantage of GP *versus* polynomial fitting: only the main characteristic of the relationship needs to be fixed *a priori* (smooth, periodic, linear) in GP modeling. The hyperparameters are then obtained with a simple non-linear optimization and do not need to be imposed. This difference is sometimes referred to as the distinction between parametric and nonparametric models.

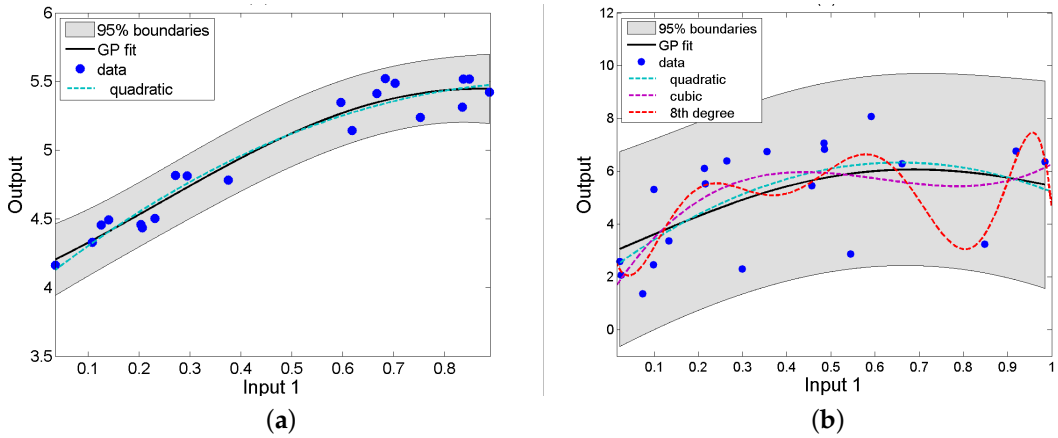


Figure 3. Gaussian processes regression and polynomial fits for two different levels of noise: (**a**) 0.02 and (**b**) 1.12.

2.3.2. Outliers

In addition to random noise, some data points can have target values that do not represent the underlying latent function. Such situations can arise, e.g., in the case of sensor malfunction or if the output was impacted by a phenomenon that is not accounted for in the inputs. These points can be considered as outliers and might have to be removed from any further analysis. Gaussian processes have proven to be a powerful framework to detect outliers (see, e.g., [8]), since the variance of the GP regression function varies with the data density and with the noise, as shown in the previous section. The effect of outliers is illustrated in Figure 4: two data samples are offset by 0.5 to simulate a possible error in the measurements. The following can be stated:

- When the data density is high (left outlier), the outlier clearly appears outside of the 95% boundaries and is easily detected.
- When the data density is low, the presence of an outlier cannot be clearly stated. The variance of the function increases, and its mean value (the black line in Figure 4) is significantly impacted. As a result, the outlier remains within the 95% boundaries. This effect is desirable since predictions made in regions of the input space with a low density or high spread of points will have appropriately low predictive precision (high variance) [8].

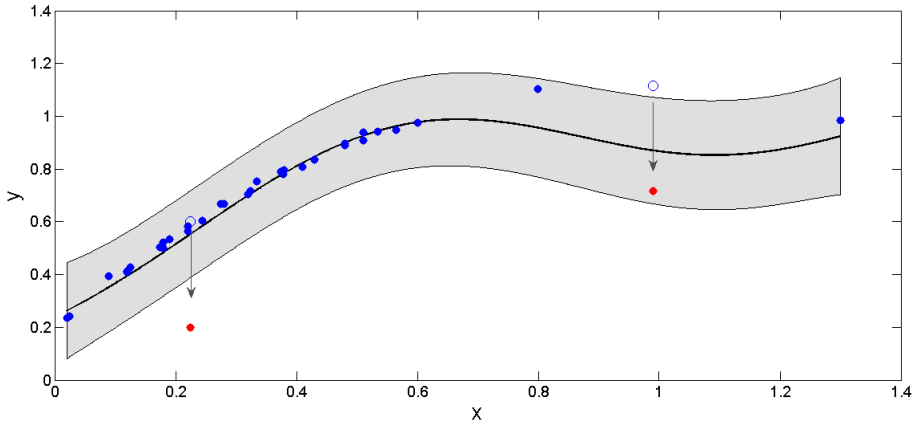


Figure 4. Effect of two outliers (red) on the Gaussian processes regression.

2.3.3. Number of Data Points and Overfitting

In the case of a low number of data points, optimizing the marginal likelihood can lead to overfitting, *i.e.*, selecting a very short length-scale to fit the noise in the data, as shown in Figure 5. In that particular case, the maximization of the marginal likelihood resulted in a (correct) length-scale value of 1.6. The value of 0.08 was imposed manually for the sake of illustration. When overfitting, the variance of the GP model is high in the zones without data, and the regression presents a more erratic behavior.

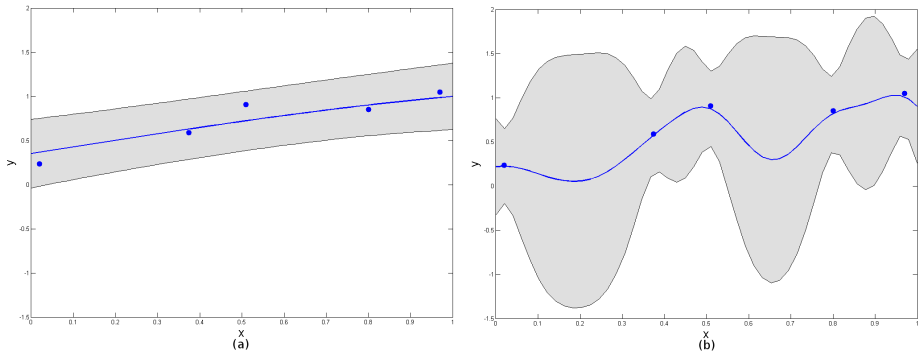


Figure 5. Gaussian processes fit with two length-scales: (**a**) 1.6 and (**b**) 0.08.

The effect is easy to apprehend in a univariate analysis. It is however more complex in a the multivariate case, in which the number of data points per dimension should increase exponentially with the number of variables. As an example, if the data samples are equally spaced on a four-dimensional hypercube (four input variables), the number of data points in each direction is given by $N^{1/4}$. If the number of data samples N is 81, this results in only three data points in each direction. The ratio between the number of data points and the dimensionality of the input space (here set to one for display purposes) hence needs to be taken into account when performing GP modeling.

The effect of the size of the dataset on the considered quality criteria is presented in Figure 6 for the univariate case. MAE is computed for the GP model with all data points (orange) and in cross-validation (blue) by varying the number of data samples. In cross-validation, MAE tends to be high for a low number of data samples, because the model is overfitting, while for a regression with all of the data samples, MAE is lower with fewer observations (overfitting cannot be detected). The difference between the two values however decreases with the the size of the dataset, which indicates that there is no longer overfitting.

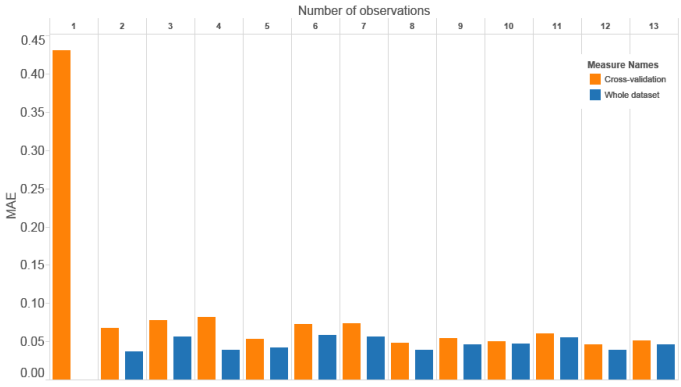


Figure 6. Impact of the number of data points on the quality criteria.

Therefore, in order to detect overfitting and to ensure a sufficient amount of data points with regard to the dimensionality of the input space, two indicators are considered:

- The ratio between the MAE values for the whole training set (all data points) and in cross-validation: a high value of this ratio is a good indicator of overfitting or of a too low number of data samples, as shown in Figure 6.
- Visual check of the shape of the GP surface response: this qualitative indicator allows checking that the shape of the regression is in agreement with the user's knowledge of the process being modeled. Experience shows that overfitting can easily be visually detected.

3. Multivariate Experimental Datasets

In the previous sections, the features of the Gaussian processes method were illustrated in the univariate case. However, the true potential of the approach lies in multivariate analyses, *i.e.*, when it is not possible to plot the output as a function of one (2D plot) or two (3D plot) inputs because of the variations of the remaining inputs. In that case, the numerical indicators provided by the method are useful for the detection of outliers, for de-noising and for feature selection.

To illustrate the GP method with test-bench experimental data, two datasets have been selected in the open literature with the following requirements:

- Steady-state data.
- Multivariate data (at least three inputs).
- The data have been exploited and a model has been proposed by the authors.
- Performance indicators for the model(s) are provided.

The proposed analysis is therefore an *ex post* analysis, in which the benefits of the proposed method are evaluated in regards to the results previously obtained with traditional methods.

It should be noted that the GP method has already been successfully tested by the authors on an experimental dataset [9] relative to an injection scroll compressor. It was demonstrated that the proposed tool could efficiently detect outliers, assess the level of noise in the data and determine the relevant inputs to predict a certain output. However, no comparison was proposed against alternative modeling or analysis approaches, which is the aim of this section.

3.1. Open-Drive Scroll Expander

The first selected experimental dataset originates from a test-rig dedicated to a scroll expander integrated into a low-capacity organic Rankine cycle. In such systems, the expansion machine is a key component for the overall performance, but no off-the-shelf solution is currently commercially available. Scroll expanders have proven to be an excellent alternative to micro-turbines in this power range, which has entailed a significant research effort towards the development and the experimental characterization of such machines (see [10] for a review).

In a scroll expander, the decrease of the pressure is caused by an increase of the volume of the expansion chambers. The ratio between the volume of the expansion chamber(s) at the end of the expansion and that at the beginning is called the "built-in volume ratio". This expansion is illustrated in Figure 7; fluid is admitted at the center and trapped in a pocket of fluid that is progressively expanded while traveling to the periphery, where the working fluid is discharged.

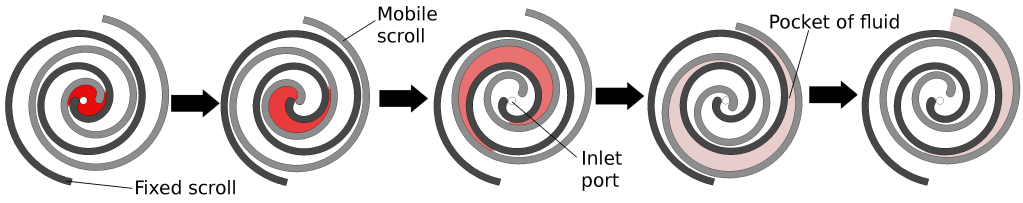


Figure 7. Working principle of a scroll expander.

The results of the experimental campaign were published in [11], together with a comprehensive description of the test-rig. In that setup (Figure 8), the input values "imposed" on the scroll expander were the working fluid flow rate $\dot{M}$, the rotational speed N_{rot}, the supply temperature T_{su}, the exhaust pressure p_{ex} and the ambient temperature T_{amb}. The measured outputs were the supply pressure p_{su}, the exhaust temperature T_{ex} and the output shaft power $\dot{W}_{sh}$. It should be noted that there is no firm causality relationship between these variables, except for T_{ex}, which is always an output, and T_{amb}, which is always an input. All of the other variables can independently be inputs or outputs of the model. As an example, if p_{su} is imposed on the test-rig instead of $\dot{M}$, the flow rate is imposed by the expander and becomes an output. In total, there are always five different inputs, the other variables being consequences of the scroll expander performance and, therefore, outputs of the model.

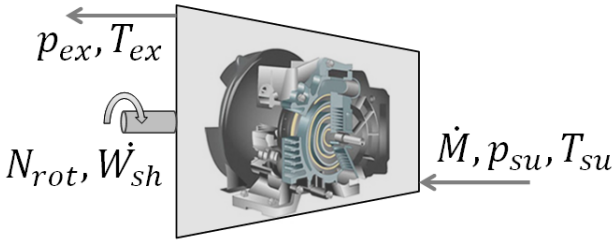
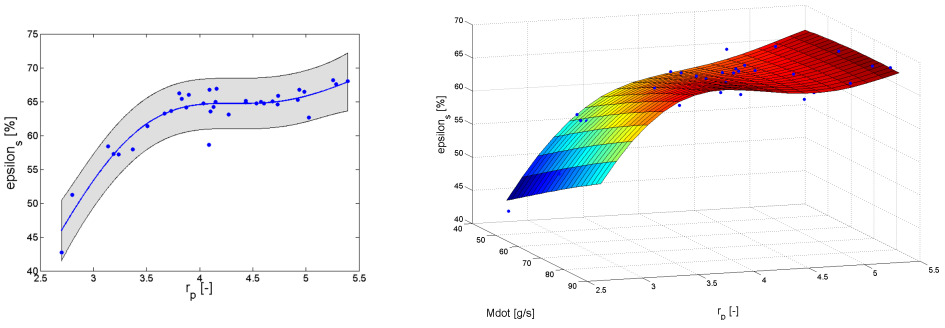


Figure 8. View of the tested scroll expander with the measured variables.

In [12], the authors of the study developed a semi-empirical model of the machine. This kind of model accounts for most of the physical phenomena and losses in the machine, but requires experimental data for their calibration. A total of 36 data points were used to tune the six empirical parameters of the model. A comparison between the model prediction and the measurement was performed, but without cross-validation. The obtained MAE was 1.94%, and the R^2 value was 98.81%.

In this *ex post* evaluation, the same dataset has been used and tested using the methodology described above. The most relevant output being the power generated by the expander, it is the one that has been selected for the analysis. Feature selection (*i.e.*, finding the most relevant set of inputs and disregarding the other ones) has been performed by iterative block addition, as proposed in [2]: if adding a new input decreases the MAE in cross-validation, the input is considered relevant. Otherwise, this additional variable only adds noise by increasing the complexity of the model without contributing to the prediction of new/unseen data samples. It can therefore be disregarded. This is illustrated in Figure 9, presenting the expander efficiency as a function of one or two inputs. In the univariate case, the error in cross-validation is 5.9%, and the standard deviation is relatively high, whereas in the bi-variate case, the average error is reduced to 5.3%. This new input is therefore considered as relevant.



(**a**) (**b**)

Figure 9. GP regression of the expander efficiency as a function of (**a**) one input or (**b**) two inputs.

The results of the analysis are presented in Figure 10 and Table 1 for a "leave-one-out" cross-validation and a number of permutations of 500. As expected, the MAE for the whole training set is higher than the MAE in cross-validation for all combinations of input, but the difference remains limited (lower than a factor of two in this case). This indicates that the model is most likely not overfitting. It can also be noted that the MAE in cross-validation keeps decreasing when adding inputs to the model, except for T_{amb}. The ambient temperature should therefore not be taken into account for the prediction of the output power, since it does not have a significant impact in this dataset and would only bring random variations to the model.

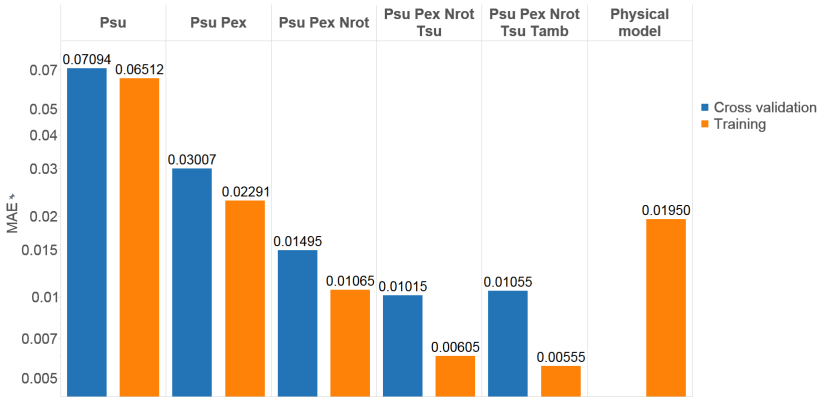


Figure 10. Mean absolute error (MAE) for different combinations of inputs and for the physical model.

Table 1. Main analysis results for different models. Akaike information criterion: AIC; Bayesian information criterion: BIC.

Model Inputs	Whole Training Set		Cross-Validation					
	MAE	R^2	MAE	R^2	p_{MAE}	AIC	BIC	Outliers
Physical model	0.0219	98.70%	-	-	-	-	-	-
P_{su}	0.0651	80.79%	0.0709	77.33%	0	−54.7	−51.5	26
P_{su}, P_{ex}	0.0229	98.08%	0.0300	96.59%	0	−107.8	−103.0	
P_{su}, P_{ex}, N_{rot}	0.0106	99.56%	0.0149	99.18%	0	−141.9	−135.5	3
$P_{su}, P_{ex}, N_{rot}, T_{su}$	0.0060	99.83%	0.0101	99.62%	0	−153.7	−145.7	3, 7
$P_{su}, P_{ex}, N_{rot}, T_{su}, T_{amb}$	0.0055	99.87%	0.0105	99.60%	0	−153.8	−144.3	3, 7

The p-value is zero for all simulations, *i.e.*, no randomly-permuted dataset performed better than the original dataset. The significance of the data is thereby confirmed.

Finally, the implementation of the ARD kernel [6] allows evaluating the sensitivity of the output to different inputs. Large length-scales prohibit fast variations of the GP function in the direction of the respective input. In the present analysis, the five inputs can be ranked in terms of increasing sensitivity (*i.e.*, decreasing length-scale): T_{su} $(l_i = 11.9)$, N_{rot} $(l_i = 5.0)$, p_{ex} $(l_i = 3.8)$, p_{su} $(l_i = 2.0)$, T_{amb} $(l_i = 7620)$.

The results of the GP analysis for this dataset provide insightful information:

- The relevant inputs to predict the output in this dataset are $P_{su}, T_{su}, \dot{M}, N_{rot}$. The ambient temperature T_{amb} should not be considered.
- Two data points are suspected to be outliers. They should be checked carefully and removed if justified.
- The minimum error that a physical model could reach to predict this data (*i.e.*, the MAE in cross-validation) is close to 1%. A model that predicts the output with a higher accuracy is most likely overfitting. In the present case, the physical model originally proposed by the authors presented an MAE of 1.94%, which indicates a remaining margin for improvement.

- A de-noised response surface of the process has been generated and allows visualizing the influence of all individual (or pairs of) inputs, maintaining the other inputs constant.
- The *p*-value of the MAE allows rejecting the null hypothesis and confirms the significance of the model performance.
- The sensitivity analysis indicates that inputs can be classified in terms of relevance as follows: p_{su}, p_{ex}, N_{rot}, T_{su}.

4. Second Example: Absorption Chiller

This second dataset is selected because of a high number of publicly-available data points and because the data have been extensively analyzed by the authors with different types of regressions or physical models [13]. The goal of the present analysis is to evaluate how the GP method can complement these approaches and how it compares in terms of the performance indicators.

The considered data comprise 138 steady-state measurements on a low-capacity (10 kW) absorption chiller (Model: *Chillii® PSC 12*) with NH3-H2O as the working fluid mixture [13]. Absorption machines provide opportunities for energy saving, because they can use heat to produce cooling (or heating, if necessary), instead of electricity used by conventional compression machines. They are based on an absorption process between a liquid absorbent and the refrigerant: the interrelations between absorption temperature, pressure and mixture concentration allow generating pressure differences from temperature differences and can therefore replace the compressor of a traditional vapor compression cycle.

A schematic view of the system is provided in Figure 11. The inputs are the values imposed by the operator, *i.e.*, the water inlet (supply) flow rates and temperature in the three circuits: $\dot{V}_{su,c}$, $\dot{V}_{su,m}$, $\dot{V}_{su,h}$, $T_{su,c}$, $T_{su,m}$, $T_{su,h}$, where $\dot{V}$ is the measured volumetric flow rate and T the measured temperature. The outputs are the variables resulting from the process: $\dot{V}_{ex,c}$, $\dot{V}_{ex,m}$, $\dot{V}_{ex,h}$, $T_{ex,c}$, $T_{ex,m}$, $T_{ex,h}$. The relevant performance indicators, however, are not the temperatures and flow rate, but the heat flows in the three different circuits. The selected heat flow for the present analysis is the evaporator heat flow (the chiller water circuit) since it corresponds to the useful effect and is the one that should be maximized:

$$\dot{Q}_{eva} = C_p \cdot \rho \cdot \dot{V}_{su,c} \cdot (T_{su,c} - T_{ex,c}) \quad (16)$$

where C_p is the specific heat capacity and ρ the density. This value is considered as the single output of the analysis and is expressed as a function of the whole set or of a subset of the input values.

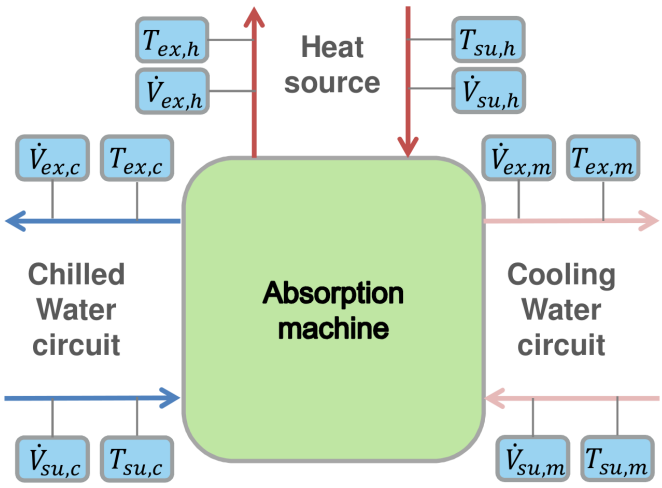


Figure 11. Schematic view of the inputs/outputs of an absorption machine.

The original study [13] includes five different modeling methods using both mechanistic and empirical approaches:

1. Thermodynamic model (TD): Simple physical model of the system built by applying mass and energy balances over all of the components and by formulating simplifying hypotheses. The model parameters are the experimentally-identified heat transfer coefficients and re-generator effectiveness.
2. Adapted Gordon–Ng model (GNA): This model is a combination of mechanistic and empirical approaches, accounting for the irreversibilities due to the finite-rate mass transfer.
3. Adapted characteristic equation ($\Delta\Delta t$): Approximate method to compute both the cooling capacity and the driving heat input by simple algebraic equations, expressed as a function of the so-called characteristic temperature function ($\Delta\Delta t$).
4. Multivariate polynomial regressions (MPR) belong to the black-box group of models, *i.e.*, they are exclusively empirical and do not account for the physical phenomena occurring within the modeled process. The order of the regression can be user-selected and is set to two in this particular case.
5. Artificial neural networks (ANN) are also black-box models, inspired by biological neural networks. They are used to estimate or approximate functions that can depend on a large number of inputs. In [13], the number of neurons is determined by a trial-and-error method, and the training is performed based on the error backpropagation method in conjunction with the Levenberg-Marquardt optimization algorithm.

For the sake of comparison, the same dataset is modeled using the GP method (Figure 2). In this particular example, feature selection is performed based on the optimal length-scales of the ARD kernel, since they are revealed to be good indicators of the relevance of a particular input in Section 3.1. The analysis is first run with the six inputs, leading to the results presented in Table 2. It clearly appears that the maximization of the marginal likelihood leads to extremely high length-scales for $\dot{V}_{\mathrm{su,m}}$ and $\dot{V}_{\mathrm{su,h}}$, which indicates that their variations are not significant to predict the considered output. The three temperature levels are significant, and the influence of $\dot{V}_{\mathrm{su,c}}$ cannot be clearly established. The analysis is therefore repeated for the cases with three and four input variables leading to MAE values in the cross-validation of 1.01% and 0.91%, respectively. It is therefore concluded that the influence of $\dot{V}_{\mathrm{su,c}}$ is significant in the present dataset and that it should be accounted for.

Table 2. Optimal length-scales for a six-input analysis.

Variable	l_{i}
$T_{\mathrm{su,m}}$	4.1
$T_{\mathrm{su,c}}$	2.8
$T_{\mathrm{su,h}}$	1.1
$\dot{V}_{\mathrm{su,m}}$	1324
$\dot{V}_{\mathrm{su,c}}$	15.5
$\dot{V}_{\mathrm{su,h}}$	681

The comparison between the GP regression and the five models proposed in [13] is displayed in Table 3. It should be noted that the model performance indicators are not computed in cross-validation, but with all of the data points, because this is how they were reported in the original work. It should also be noted that the five models take only the three temperature levels as inputs and neglect $\dot{V}_{\mathrm{su,c}}$, which, according to the previous analysis, has a non-null contribution to the best model fit. Therefore, for the sake of comparison, the results of the GP regression are displayed for both the three- and four-input cases.

Table 3. Model performance for the absorption machine. TD, thermodynamic model; GNA, adapted Gordon–Ng; MPR, multivariate polynomial regression; ANN, Artificial neural networks.

Model	$N_{parameters}$	R^2	Outliers
TD	4	0.932	-
GNA	2	0.9014	-
DT	4	0.9871	-
MPR	10	0.9937	-
ANN	36	0.9982	-
GP 3 inputs	4 *	0.9982	44, 97, 123
GP 4 inputs	5 *	0.9986	88, 97, 123

* In GP, the number of hyperparameters (and not parameters) is reported.

As indicated in Table 3, the GP method outperforms the first four models and seems to perform as well as ANN. If the forth relevant variable is added to the model, the accuracy of the prediction increases slightly, and interestingly, the data point of 44 leaves the set of detected outliers and 88 enters.

5. Implementation

The methodology presented in Figure 2 is implemented into GPExp, an open-source data analysis framework written in MATLAB. The suite uses the GPML library [1] and comprises a graphical user interface (Figure 12). It can perform the different analyses (permutations, cross-validation, minimization of the marginal likelihood, outlier detection, prevention of over-fitting, *etc.*) in an automated way.

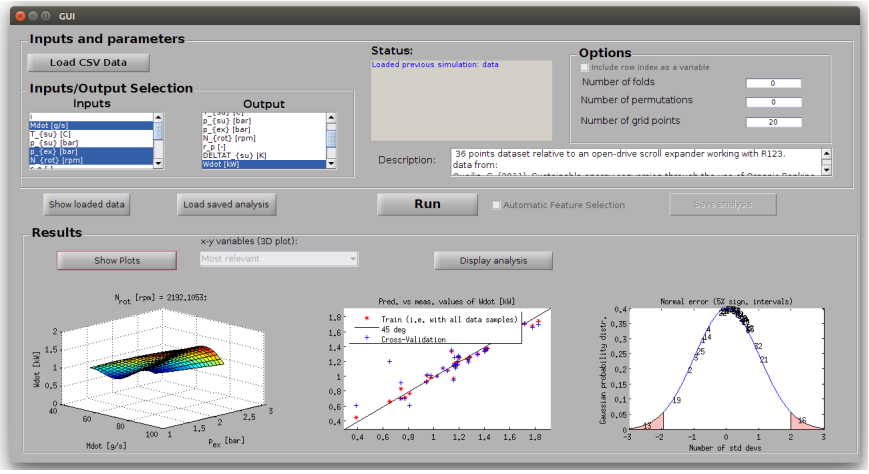


Figure 12. Main window of the GPExp graphical user interface (GUI).

The input data are provided in the form of a .csv file, a .mat file or an Excel file. This comprises a matrix of the experimental inputs and outputs. The user is required to select the considered inputs and the output before running the analysis. User-defined parameters also include the number of folds, the number of permutations and different plotting options.

The algorithm is run according to Figure 2 and generates a data analysis report that provides a qualitative interpretation of the main simulation results. These interpretations are generated based on heuristic rules and are mainly dedicated to users who are not specialists in machine learning. They are however provided with warnings, since the rules (e.g., the threshold to detect overfitting) were defined empirically and cannot be guaranteed for all datasets.

The results are finally stored in a structure and can be saved in a .mat file. They can further be reloaded in GPExp or used for the prediction of new/unseen data points using a dedicated GPExp function.

6. Conclusions

This paper presents a methodology to analyze experimental steady-state data. The method relies on Gaussian processes' regression, which is a well-known technique, but had, to our knowledge, never been applied to the critical analysis of monitoring data.

Data quality is evaluated using numerical model performance indicators by comparing it to the GP regression latent function. These indicators are useful, e.g., to assess the quality of the correlation between some measured operating conditions (inputs) and some measured performance data (outputs). They also set a benchmarking standard to compare different sets of experimental data. Furthermore, the probabilistic formulation of Gaussian processes provides confidence intervals to predict the output with a given set of inputs, which are a function of the noise and of the local data density.

In addition to the evaluation of the data quality, the method also helps with evaluating which variables are relevant to the selected model. The feature selection capability allows determining the relevant inputs for the prediction of one output variable. In this paper, this is achieved by means of two different, but converging techniques: the comparison of the cross-validation errors with recursive feature addition and the comparison of the length-scales relative to each input.

It is further demonstrated, through examples, how the proposed tool can efficiently be used to detect the main dependencies, shortcomings and outliers in experimental data. Examples are first described for the univariate case, whose quality can be assessed visually, and then extended to processes with multiple input variables.

In a first test case relative to an expander dataset, it is shown that the GP model provides a better prediction of the output power than semi-empirical models and can be used as an indicator of the maximum achievable prediction accuracy.

In a test case relative to an absorption machine, it is shown that GP regression performs as well as ANN in terms of prediction accuracy, but does not require any kind of tuning or parametric identification. It also outperforms other black-box or physical methods, such as polynomial regressions or thermodynamic models.

The method is implemented within the open-source tool GPExp, which is developed in such a way that a qualitative interpretation of the results is provided to users without machine learning expertise. It comprises a graphical user interface (GUI) and can be freely downloaded and tested.

Acknowledgments: The results presented in this paper have been obtained within the frame of three different programs: The Project SBO-110006 "The Next Generation Organic Rankine Cycles", funded by the Institute for the Promotion and Innovation by Science and Technology in Flanders (IWT); the Postdoctoral fellowship of the Belgian Fonds National de la Recherche Scientifique (FNRS); the "Marie Sklodowska-Curie Actions (MSCA) fellowship" No. 654038. This financial support is gratefully acknowledged. The authors also want to thank Arnaud Legros for the development of the graphical user interface (GUI) of GPExp.

Author Contributions: Jessica Schrouff and Sylvain Quoilin participated in the development and implementation of the mathematical method. Testing and reporting were performed by Sylvain Quoilin.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Rasmussen, C.E. *Gaussian Processes for Machine Learning*; MIT Press: Cambridge, MA, USA, 2006.
2. Schrouff, J.; Kusse, C.; Wehenkel, L.; Maquet, P.; Phillips, C. Decoding Semi-Constrained Brain Activity from fMRI Using Support Vector Machines and Gaussian Processes. *PLoS ONE* **2012**, *7*, doi:10.1371/journal.pone.0035860.
3. Friedman, J.; Hastie, T.; Tibshirani, R. *The Elements of Statistical Learning*; Springer: Berlin, Germany, 2001; Volume 1.
4. LaConte, S.; Strother, S.; Cherkassky, V.; Anderson, J.; Hu, X. Support vector machines for temporal classification of block design fMRI data. *NeuroImage* **2005**, *26*, 317–329.
5. Shawe-Taylor, J.; Cristianini, N. *Kernel Methods for Pattern Analysis*; Cambridge University Press: Cambridge, UK, 2004.

6. Neal, R.M. Assessing relevance determination methods using DELVE. In *Nato ASI Series F: Computer and Systems Sciences*; Springer: Berlin, Germany, 1998; Volume 168, pp. 97–132.
7. Schrouff, J. Pattern Recognition in NeuroImaging: What Can Machine Learning Classifiers Bring to the Analysis of Functional Brain Imaging? Ph.D. Thesis, University of Liege, Liege, Belgium, 2013.
8. Marquand, A.F.; Rezek, I.; Buitelaar, J.; Beckmann, C.F. Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biol. Psychiatry* **2016**, doi:10.1016/j.biopsych.2015.12.023.
9. Quoilin, S.; Schrouff, J. Assessing the quality of Experimental Data with Gaussian Processes: Example with an Injection Scroll Compressor. In Proceedings of the 2014 Purdue Conferences: Compressor Engineering, Refrigeration and Air-Conditioning, West Lafayette, IN, USA, 14–17 July 2014.
10. Quoilin, S.; Van Den Broek, M.; Declaye, S.; Dewallef, P.; Lemort, V. Techno-economic survey of Organic Rankine Cycle (ORC) systems. *Renew. Sustain. Energy Rev.* **2013**, *22*, 168–186.
11. Quoilin, S.; Lemort, V.; Lebrun, J. Experimental study and modeling of an Organic Rankine Cycle using scroll expander. *Appl. Energy* **2010**, *87*, 1260–1268.
12. Quoilin, S. Sustainable Energy Conversion Through the Use of Organic Rankine Cycles for Waste Heat Recovery and Solar Applications. Ph.D. Thesis, University of Liege, Liege, Belgium, 2011.
13. Labus, J. Modelling of Small Capacity Absorption Chillers Driven by Solar Thermal Energy or Waste Heat: Tesi Doctoral. Ph.D. Thesis, Universitat Rovira i Virgili, Tarragona, Spain, 2011.

© 2016 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC-BY) license (http://creativecommons.org/licenses/by/4.0/).