



**HAL**  
open science

# Data-driven Model Generation Process for Thermal Monitoring of Wind Farm Main Components through Residual Indicators Analysis

Théodore Raymond, Alexis Lebranchu, Christophe Bérenguer, Sylvie Charbonnier

## ► To cite this version:

Théodore Raymond, Alexis Lebranchu, Christophe Bérenguer, Sylvie Charbonnier. Data-driven Model Generation Process for Thermal Monitoring of Wind Farm Main Components through Residual Indicators Analysis. ESREL 2022 - 32nd European Safety and Reliability Conference (ESREL 2022), TU Dublin; ESRA, Aug 2022, Dublin, Ireland. pp.1393-1400, 10.3850/978-981-18-5183-4\_R23-01-601-cd . hal-03877225

**HAL Id: hal-03877225**

<https://hal.science/hal-03877225v1>

Submitted on 29 Nov 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

# Data-driven Model Generation Process for Thermal Monitoring of Wind Farm Main Components through Residual Indicators Analysis

Théodore Raymond

*Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, 38000, Grenoble, France., Data Engineering, Valemo, France. E-mail: theodore.raymond@valemo.fr*

Alexis Lebranchu

*Data Engineering, Valemo, France. E-mail: alexis.lebranchu@valemo.fr*

Christophe Bérenguer

*Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, 38000, Grenoble, France. E-mail: christophe.berenguer@grenoble-inp.fr*

Sylvie Charbonnier

*Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-Lab, 38000, Grenoble, France. E-mail: sylvie.charbonnier@grenoble-inp.fr*

Most of the Supervisory Control and Data Acquisition (SCADA) fault indicators proposed in the literature to detect a fault that induces a temperature increase of the physical components of a wind turbine are temperature residuals. Temperature residuals measure the difference between the current value of the temperature of a component and its prediction by a normal behavior model. In the literature, normal behavior models built from variable selection algorithms are ad-hoc models, designed to correctly predict the temperature of a specific component of a specific turbine of a specific wind farm. In practice, these models cannot be used to predict the temperature of a component of another turbine, let alone a turbine in a different wind farm, because the sensors used by wind turbine manufacturers are not the same. It is therefore impossible for an industrial wind farm manager to deploy a residual-based fault detection system on a wind farm scale. In order to make it possible to deploy these methods in an industrial context, we propose in this paper a methodology to automatically build linear models capable of predicting all temperatures of any component of any turbine of a given wind farm. The method is designed to be easy to implement, interpretable by the operator, and fast to execute to meet industrial constraints. The set of models obtained allows to build a network of thermal state indicators, which can be used for fault isolation. The method is applied to the monitoring of the thermal condition of a real French wind farm for illustration.

*Keywords:* Model validation, linear multi-variable systems, health monitoring, data-driven decision making, sensor data fusion, estimation and fault detection.

## 1. Introduction

A wind turbine is a complicated system which is exposed to many disturbances that can damage the machine components and generate faults. These faults, if left unattended, can lead to the failure of major components and result in high production losses. To address this concern, new maintenance strategies are being developed and implemented, based on condition monitoring. These strategies aim at monitoring the health status of components by analyzing synthetic condition indicators. Residuals built from the supervisory control and

data acquisition (SCADA) data are often used as fault indicators. They measure the difference between the on-line value of a variable recorded by the SCADA and its prediction by a model learned during a period when the machine operates in normal conditions (normal behavior modeling (NBM) strategy).

Developing models for wind turbine condition monitoring is a blooming field of research and some of the solutions proposed in the literature have already proved their potential on concrete scenarios like Schlechtingen and Santos (2014)

using normal behavior modeling strategy or the intelligent system presented in Garcia et al. (2006) (see Isermann (2005) for an introduction to fault-diagnosis theory). Yet, the latest researches in wind turbine condition monitoring are usually focused on one major component - e.g., Yuan et al. (2019) focus on the gearbox while Lebranchu et al. (2019) studied the bearings of the generator -, or a small subset of them. No study focuses on the monitoring of the whole wind turbine system.

In this paper, a generic wind farm health monitoring method is proposed, which can be applied to all the main components of the wind turbine nacelle, whatever the technology used by the manufacturers. To do so, a method to generate in an automatic way normal behavior models for any component of a wind turbine nacelle of any technology is developed.

A literature review shows that statistical modeling of the most at-risk components can be done according to two main approaches: (i) by mixing general and expert knowledge to build physics-based models, (ii) by using algorithms that allow automated selection of all potential explanatory variables to build data-driven models.

The physic-based models require in-depth knowledge of the internal dynamic relationships between variables measured on the component, which is not always available. As a consequence, the latest researches have focused on the modeling of some major components only, such as the gearbox, as can be seen on Bie et al. (2021). Yet, all the components of a wind turbine can be exposed to more or less critical failure. This is why many authors use a data driven modeling approach, by implementing an automatic selection of variables for a given variable to be estimated. This enables a generally large number of variables available in the SCADA system to be exploited.

Many algorithms can perform variable selection. Let us mention those that select the variables of the model by iteratively ranking the variables according to a given performance criterion (Pohjankukka et al. (2018)), and those that allow the model to be modified during its construction in order to avoid any risk of sub optimality (Nguyen et al. (2021)). However, in the literature, variable

selection algorithms are generally made to model specific components of a wind turbine from a given manufacturer with the aim of generating efficient health indicators for these components. Consequently, the generalization of the algorithms to other components or to turbines of other farms with different technologies is not assured, which strongly limits their industrial interest. Indeed, each wind turbine manufacturer provides its own network of sensors, and consequently its own database. This is why the variable selection algorithms must be as insensitive as possible to the data.

This paper presents a method to automatically generate normal behavior linear models using all the operational data at disposal. The method is made to be easy to implement, interpretable for the operator, and fast to execute for industrial constraints. These models can be used to predict the temperature of any component of all the turbines of the same farm. The models generated can then be used to build residuals, with the aim to monitor the thermal state of the main wind turbine nacelle components (rotor, slow shaft, gearbox, generator, converter, transformer) and isolate the fault. The residuals are built in a multi turbine approach, to be less sensitive to operational and environmental variations. While this paper focuses on turbines from a single manufacturer, the next step of the study will be to increase the generalization of the process by extending it and applying it to the entire fleet of wind farms to validate its industrial interest.

The outline of this paper is as follows. The method proposed to generate the models and to build the residuals is detailed in Section 2. Section 3 presents the database used in this study and Section 4 presents the results obtained on a case of converter cooling system failure.

## 2. Proposed methodology

### 2.1. Framework and assumptions

The study is divided into two parts: (i) the automatic selection of the variables of the model, (ii) the computation of a thermal state indicator (TSI), from the model generated in the first part.

Let  $y$  be a variable to model (output) and  $V$  the

set of eligible variables to be used in the model (inputs). The model generation process aims at building linear models by selecting variables  $\nu$  belonging to  $V$  to predict the variable  $y$ , based on a defined criterion. The generated model is a linear model of size  $L$  of the following form:

$$\hat{y} = a_1 * \nu_1 + \dots + a_L * \nu_L, a_1 \dots a_L \in \mathbb{R} \quad (1)$$

In this paper, only thermal failures such as component overheating caused by bearing wear or by ventilation defect are considered. Thus, the variables  $y$  to be predicted are temperatures. The variables  $\nu$  may be any of the  $V$  variables recorded by the SCADA system of the farm. The choice of the linear model structure is made to improve the interpretability of the generated models, in addition to strongly improving the execution speed of the algorithm, at the expense of the estimation performance.

The goal is to generate linear models that can be used on any temperature recorded on any turbine of a wind farm. We suppose that such models can be found because (i) wind turbines of the same farm are geographically close and are thus submitted to the same environmental conditions, and (ii) they are of the same manufacturer.

It is important to emphasize that the goal of this process is to generate relevant linear models for any temperature measured on the wind turbine. Indeed, in this study, all the components of the turbine are considered equally important to monitor and not just one major component. A normal behavior modeling strategy is used (Tautz-Weinert and Watson (2016)): only data from a period where the wind turbine is assumed to be working in normal conditions is used to train the models.

At the end of the model generation process, one model is assigned to one component to estimate its temperature. The same model is used to predict the component temperature of all the turbines of a wind farm: the selected variables  $\nu$  are the same but the coefficients of the model  $a$  are different for each turbine (Eq.(1)). Each model can thus provide a residual that forms a thermal state indicator (TSI). Let  $c$  be a component of the nacelle of the turbine.  $\text{TSI}_c$  is then the thermal state indicator for

the component  $c$ , and  $\Omega_{\text{TSI}}$  the bank of TSI to be used for the detection and isolation of faults.

## 2.2. Model generation process

### 2.2.1. Presentation of the algorithm

The proposed automatic selection of variables follows a Greedy Forward Selection (GFS) approach: from an empty model and given an output  $y$ , the eligible variables (inputs) are sorted according to a ranking function  $f_r$  which will be developed in the following section. The variable  $\nu_*$  which obtains the best  $f_r$  value is then selected as the first element of the model under construction. The process is then repeated for the model containing  $\nu_*$  until the stopping criteria is achieved. The GFS algorithm is chosen for its simplicity of implementation, for the strong controllability offered by the choice of the ranking function, and because it can provide equivalent performances to more complex feature selection algorithm such as genetic algorithm, as showed in Pohjankukka et al. (2018). The ranking function  $f_r$ , which selects the best variable, uses the Mean Absolute Error, referred as MAE. It measures the average deviation between the estimated value  $\hat{y}$  and the measured one  $y$  on a given period of  $N$  samples (3).

The variable selection process proposed in this paper aims at generating relevant and efficient models whatever the output is. The MAE is a good choice to achieve that goal because of its easy implementation, its high interpretability and because it is less sensitive to outliers that may remain on the training data and have impact on the feature selection result than other classical metrics such as the Root Mean Squared Error or RMSE. Chai and Draxler (2014). Moreover, performance stabilization can be rapidly achieved with very simple models of less than five variables using this performance criterion.

### 2.2.2. Complete GFS algorithm development

For a given temperature variable of a component to be estimated  $y_c$ , the GFS algorithm aims to find a set of  $L$  variables out of the  $V$  eligible one to predict  $y_c^{T_j}$  where  $j \in \llbracket 1; N_T \rrbracket$ . The estimated variable is then called  $\hat{y}_c^{T_j}$ , and is described by the

following equation.

$$\hat{y}_c^{T_i} = f(\nu_{*1}^{T_i}, \dots, \nu_{*L}^{T_i}) \quad (2)$$

Let  $N_T$  be the number of wind turbines in the wind farm considered,  $T_i$  a wind turbine of the farm,  $s$  the step of the GFS algorithm such that  $s > 1$  and  $\nu_{s-1}^*$  the variable selected at the previous step.

For a given variable to be ranked  $\nu_{ks}$  ( $k \in \llbracket 1; N_{V_s} \rrbracket$ ) in  $V_s$ , the set of eligible variables at step  $s$ , we first train  $N_T$  linear models (one per turbine) composed of the  $s - 1$  last selected variables and  $\nu_{ks}$  ( $\hat{y}_{\nu_{ks},c}$ ), using the  $N$  values of the variables on the training period of the corresponding wind turbine. The models are trained using a linear regression following an  $L_1$  regularization (or Lasso). This regularization is chosen for its high selectivity, which allows to exclude the non relevant variables for the estimation of the output Greenwood et al. (2020). Their corresponding MAE is then computed following Eq.(3).

$$\text{MAE}_{\nu_{ks}}^{T_i} = \sum_{n=1}^N \frac{|y_c^{T_i}(n) - \hat{y}_{\nu_{ks},c}^{T_i}(n)|}{N} \quad (3)$$

The goal is to build models that can be used on every wind turbine of the farm. Thus, we further compute the median of these  $N_T$  MAE to evaluate the performance of the given model on the whole farm. The median is preferred to the mean because it is less sensitive to possible outliers affecting the data of a machine in the farm.

The difference between the resulting  $\text{MAE}_{\nu_{ks}}$  and the best performance  $\text{MAE}_{\nu_{s-1}^*}$  obtained at the previous step  $s - 1$  forms the ranking function  $f_r(\nu_{ks})$  (see Lines 6 to 10 of Algorithm 1).

The process is repeated for every variable of  $V_s$ . Then, the variable  $\nu_s^*$  that gets the highest ranking function  $f_r$  is added to the model (for  $s = 1$  the variable that gives the minimum MAE is added). The corresponding  $\text{MAE}_{\nu_s^*}$  is stored for the next step and  $\nu_s^*$  is removed from the eligible variable set  $V_s$ . Models composed of three variables are generated ( $L = 3$ ). The complete variable selection process is detailed in Algorithm 1.

---

### Algorithm 1 Variable selection process

---

```

1: for  $\nu_{ks} \in V_s$  do
2:   for  $j \in \llbracket 1; N_T \rrbracket$  do
3:      $\text{MAE}_{\nu_{ks}}^{T_j} \leftarrow (3)$ 
4:   end for
5:    $\text{MAE}_{\nu_{ks}} \leftarrow \text{median}_{j \in \llbracket 1; N_T \rrbracket}(\text{MAE}_{\nu_{ks}}^{T_j})$ 
6:   if  $s = 1$  then
7:      $f_r(\nu_{ks}) \leftarrow \text{MAE}_{\nu_{ks}}$ 
8:   else if  $s > 1$  then
9:      $f_r(\nu_{ks}) \leftarrow \text{MAE}_{\nu_{s-1}^*} - \text{MAE}_{\nu_{ks}}$ 
10:  end if
11: end for
12: if  $s = 1$  then
13:    $\nu_s^* \leftarrow \text{argmin}_{\nu_{ks} \in V_s}(f_r(\nu_{ks}))$ 
14: else if  $s > 1$  then
15:    $\nu_s^* \leftarrow \text{argmax}_{\nu_{ks} \in V_s}(f_r(\nu_{ks}))$ 
16: end if

```

---

## 2.3. Thermal State Indicator (TSI)

### 2.3.1. Step by step building

Once the model is generated by the GFS algorithm to predict a variable  $y_c$ , the corresponding thermal state indicator, named TSI, is computed as follows.

For a given component  $c$ , a residual is computed for each of the  $N_T$  turbines of the farm. A farm reference is then calculated as the median of the residual indicators of the  $N_T$  wind turbines of the farm. Finally, the TSI is calculated for each turbine as the difference between the residual of the turbine considered and the farm reference.

This indicator evaluates the thermal state of a component according to the data from the considered machine and according to the median thermal state of the wind farm. This multi-turbine approach was introduced in Lebranchu et al. (2019) to make the TSI both robust to production mode and to cyclic phenomena such as seasonality.

Fig. 1 illustrates the TSI computation process for a given wind turbine  $T_i$ , where  $\hat{y}$  is the output estimated by the GFS model.

To handle possible residual outliers, a daily sliding average filter is applied on the TSI values to smooth the signal. The filtered TSI will be referred as  $\text{TSI}_f$ . A  $\text{TSI}_f$  is computed for each component of a wind turbine of the wind farm.

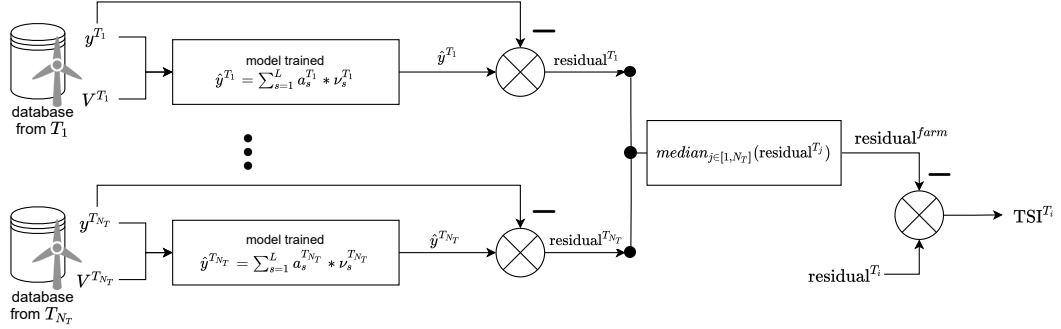


Fig. 1. TSI computation process

All the  $\text{TSI}_f$  obtained form a network of health indicators which can be used to detect and isolate a fault.

### 3. Presentation of the data set

#### 3.1. Wind farm test case

The method proposed is implemented on a wind farm located in the south of France. It is a recent one: only data from the second semester of 2020 are available. It contains 10 wind turbines ( $N_T = 10$ ) from the same manufacturer and with the same technologies. It has suffered numerous converter and transformer failures since its commissioning.

The bank of TSI is used to detect a converter failure which impacted all the turbines of the farm in the summer period of 2020. The cooling system of the converters was disturbed mainly because particles from a harvest in a field near the farm clogged the fan grills. The cleaning of the fan grills on  $T_2$  took place on August 12 for a duration of two days. Section 4 will focus on two turbines of the farm:  $T_2$ , the machine the most impacted by the cooling failure, and  $T_9$ , the least impacted one.

#### 3.2. SCADA data

##### 3.2.1. Wind turbine manufacturer database

The SCADA variables available depends on the wind turbine manufacturer. Indeed, the set of sensors installed and the technology they use is different from one manufacturer to the other. Moreover, only a subset of the variables provided by the manufacturer is transmitted to the company.

The fleet of wind farms operated by the company is composed of wind turbines from six different manufacturer. This reduction of the eligible variable set can limit the modeling potential of the GFS algorithm, as well as the efficiency of the generated models. One of the next objectives of the study will be to take into account all the data provided by the manufacturer.

For the purpose of the study, a wind farm with a set  $V$  of 42 eligible variables is considered. These variables are provided by sensors located on the different components of the wind turbine's nacelle (rotor, gearbox, generator, converter, transformer). They are of different types: temperature, shaft rotation speed, active power, angle. . . . The sensor values are acquired every 40ms by the wind turbine control system. The average of these values over the last 10 minutes is then stored in the SCADA. These average values are the ones used for the study.

##### 3.2.2. Output variables

One temperature for each of the six components is selected as output variable to generate the bank of linear models.

The oil temperature of the gearbox is known to be a good output variable for the thermal state estimation of the component, because it is the first element to be impacted by an overheat (Bie et al. (2021)). Concerning the generator, less research has been done, but the temperature of the front and rear bearings are generally used (Lebranchu et al. (2019)). As the thermal modeling of the four re-

maining components (rotor, slow shaft, converter, transformer) has not been the object of concrete research, and as the study is limited by the set of available variables, the variables physically closest to the studied components have been chosen as output.

### 3.2.3. Experimental data

Due to the lack of available data because of the age of the studied wind farm, four months of data from September 2020 to the end of December 2020 are used for model learning. The TSIs derived from these models are applied on the data of the validation period, which starts on June 1 and ends on August 15, 2020. **test** During this period, the converter failure described in Section 3.1 impacted some turbines of the farm. All the data are standardized, using the Z-score normalization, to ensure convergence of the regression algorithm.

## 4. Results and discussions

### 4.1. Generation of the bank of TSI

Table 1 shows the variables selected by the GFS algorithm for each of the six temperatures  $y_c$ . To evaluate the performances of the residuals built from these models, the root mean squared values of each filtered  $\text{TSI}_{f,c}^{T_i}$  is calculated on the training period for each of the turbine  $T_i$  (4). This value is named PM (Performance Measurement). Table 1 shows the average of these  $N_T$  PM calculated for each model. This metric should be considered as a score assigned to each model generated, and it should be as close to zero as possible.

$$\text{PM}(\text{TSI}_{f,c}^{T_i}) = \sqrt{\sum_{n=1}^N \frac{\text{TSI}_{f,c}^{T_i}(n)^2}{N}} \quad (4)$$

Table 1. Description and score of the models.

Output	Variables	Average PM
$t_{\text{hub}}$	$t_{\text{internal}}, P_{\text{apparent}}, t_{\text{automatism}}$	0.046
$t_{\text{slowshaftbearing}}$	$t_{\text{automatism}}, t_{\text{gearboxoil}}, t_{\text{hub}}$	0.138
$t_{\text{gearboxoil}}$	$t_{\text{gearboxbearing1}}, P_{\text{apparent}}, t_{\text{converter}}$	0.064
$t_{\text{stator}}$	$t_{\text{gearboxbearing2}}, i_{\text{generated}}, t_{\text{converter2}}$	0.374
$t_{\text{converter}}$	$t_{\text{gearboxoil}}, P_{\text{apparent}}, t_{\text{mastbase}}$	0.141
$t_{\text{transfo}}$	$S_{\text{wind2}}, i_{\text{generated}}, P_{\text{activesetpoint}}$	0.102

The models generated are simple ones. Most of the variables  $\nu$  selected are temperatures, which means that the GFS algorithm favors variables of the same type as the output  $y_c$ . One can also notice that the variables selected are not restricted to the component studied, which shows the strong interaction between the major components of the nacelle.

We can see on Table 1 that the average PM value, computed from  $\text{TSI}_f$ , is different for each component. These values express the estimation performance of the  $\text{TSI}_f$  during a period when the turbine is assumed to be operating in normal condition.

## 4.2. Application on the selected failure

### 4.2.1. Detection threshold tuning

The most common way to conduct health monitoring using residuals is to set a threshold value to separate low residual values, illustrating a normal operating period, and higher values, representing a potential defect in progress. This threshold is called the detection threshold.

In our case, the health indicators used to detect a fault are the processed TSIs. Tuning the detection threshold is usually done according to a fault detection performance criterion such as the false alarm rate (defined as the percentage of samples over the detection threshold in a normal operating period). In the case of this study, the detection thresholds were set as the 99th percentile of the  $\text{TSI}_f$  computed during the learning period, which sets the false alarm rate to 1%.

### 4.2.2. Visual analysis

For a given wind turbine  $T_i$ , the weekly alarm rate is defined as the number of  $\text{TSI}_f$  samples above the detection threshold over the number of  $\text{TSI}_f$  samples during the week considered. These values are computed for every  $\text{TSI}_f$ .

The comparative analysis of the residuals is performed using a heatmap. In this heatmap, each row correspond to a  $\text{TSI}_f$ . Each column is a week during the validation period (see Section 3.2.3). The weekly alarm rate is then displayed as a colored indicator graduated from zero (white) to one (purple). This visual representation allows

an overview of the detection performances of the bank of  $\text{TSI}_f$ .

Fig. 2 show the resulting heatmap for  $T_2$  and  $T_9$  for the fault period from June 1 to August 15, 2020. As a reminder,  $T_2$  is the most impacted machine of the farm while no intervention on the converter has been done on  $T_9$  during the year 2020. On the heatmap from  $T_2$  values, the fifth row,

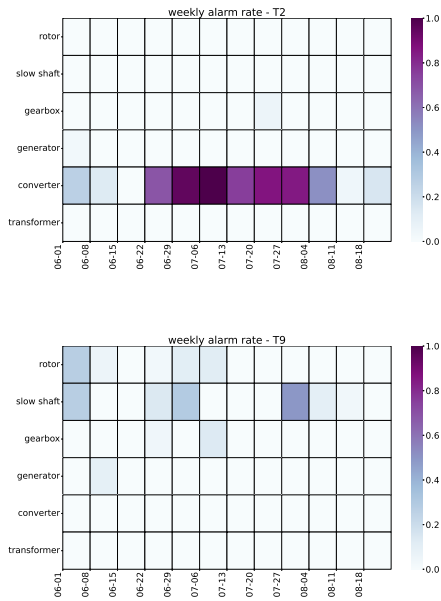


Fig. 2. Weekly alarm rate heatmap of the  $\text{TSI}_f$  for  $T_2$  (top) and  $T_9$  (bottom)

representing the converter  $\text{TSI}_f$ , is highlighted during a period of six weeks, which corresponds to the time when the field near the wind farm was harvested until the time when ventilation fan were cleaned. The weekly alarm rate remains at a level close to zero for the other  $\text{TSI}_f$  from  $T_2$  and  $T_9$  as can be seen in Fig. 2. This means that on all the TSI computed except  $\text{TSI}_{\text{converter}}^{T_2}$ , no significant crossing of the detection threshold has been observed on the validation period, and so that the associated components are in a good thermal condition. These results show that the models built using the strategy presented in Section 2 are able to predict correctly the variations of the temperatures when the turbine operates in normal conditions. Thus, the resulting  $\text{TSI}_f$  networks allow

the failure on the converter of turbine  $T_2$  to be detected and localized.

However, for a deeper analysis of the results, we will take a look at the temporal profiles of the TSI, which are used to build the weekly alarm rate heatmap.

Fig. 3 shows the six  $\text{TSI}_f$  profiles in function of time during the validation period for the wind turbines referenced  $T_2$  (top plot) and  $T_9$  (bottom plot). The  $\text{TSI}_f$  values above the detection threshold are highlighted by a coloring of the area under the curve. The purple signal shows the profile of the  $\text{TSI}_f$  monitoring the converter thermal state. This figure confirms the observation done using the heatmap in Section 4.2.2.

## 5. Conclusion

In this paper, a method to build a network of health indicators, to monitor wind turbine main components and isolate possible faults is presented. The indicators are residuals computed from linear normal behavior models. The models are generated in an automatic way, using a data driven approach, which selects iteratively the best variables to add to the model. The models obtained are simple, and easy to interpret. They are able to predict efficiently the temperature of any component, for all the turbines of a wind farm. The residuals obtained for each turbine are compared to a farm reference, to make them less sensitive to variations due to environmental and operational changes. They form a network of thermal state indicators which allows a global view of the thermal state of the whole turbine. It gives good detection and localization performances on a real failure case. The method proposed can be easily applied to any new wind farm, which is its main advantage. Thus, in the future, an evaluation of the TSIs network performance detection will be conducted on other wind farms from Valemo's fleet.

## References

- Bie, Y., X. Liu, T. Xu, Z. Zhu, and Z. Li (2021, May). A review of the application of oil analysis in condition monitoring and life prediction of wind turbine gearboxes. *insight* 63(5), 289–301.



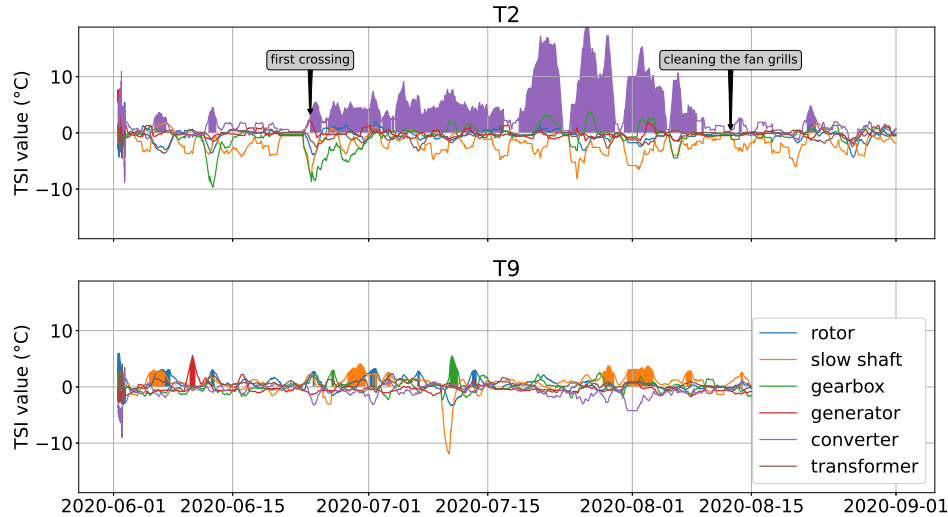


Fig. 3.  $TSI_f$  profiles for  $T_2$  (top) and  $T_9$  (bottom)

- Chai, T. and R. R. Draxler (2014, June). Root mean square error (RMSE) or mean absolute error (MAE)? – Arguments against avoiding RMSE in the literature. *Geosci. Model Dev.* 7(3), 1247–1250.
- Garcia, M. C., M. A. Sanz-Bobi, and J. del Pico (2006, August). SIMAP: Intelligent System for Predictive Maintenance. *Computers in Industry* 57(6), 552–568.
- Greenwood, C. J., G. J. Youssef, P. Letcher, J. A. Macdonald, L. J. Hagg, A. Sanson, J. McIntosh, D. M. Hutchinson, J. W. Toumbourou, M. Fuller-Tyszkiewicz, and C. A. Olsson (2020, November). A comparison of penalised regression methods for informing the selection of predictive markers. *PLoS ONE* 15(11), e0242730.
- Isermann, R. (2005). *Fault-Diagnosis Systems: An Introduction from Fault Detection to Fault Tolerance*. Springer Science & Business Media. Google-Books-ID: 6yUfoZhGMY0C.
- Lebranchu, A., S. Charbonnier, C. Bérenguer, and F. Prevost (2019). A combined mono- and multi-turbine approach for fault indicator synthesis and wind turbine monitoring using SCADA data. *ISA Transactions* 87, 272–281.
- Nguyen, M. H., P. Le Nguyen, and Nguyen (2021). PM2.5 Prediction Using Genetic Algorithm-Based Feature Selection and Encoder-Decoder Model. *IEEE Access* 9, 57338–57350. Conference Name: IEEE Access.
- Pohjankukka, J., S. Tuominen, and Pitkänen (2018, October). Comparison of estimators and feature selection procedures based on ALS and digital aerial imagery. *Scandinavian Journal of Forest Research* 33(7), 681–694.
- Schlechtingen, M. and I. F. Santos (2014, January). Wind turbine CBM using NBM. Part 2: Application. *Applied Soft Computing* 14, 447–460.
- Tautz-Weinert, J. and S. J. Watson (2016). Revue - SCADA data uses for WT CM. *IET Renewable Power Generation* 11(4), 382–394. Publisher: IET.
- Yuan, T., Z. Sun, and S. Ma (2019, January). Gearbox Fault Prediction of Wind Turbines Based on a Stacking Model and Change-Point Detection. *Energies* 12(22), 4224. Number: 22 Institute.