Layered Integration Approach for Multi-view Analysis of Temporal Data*

Michiel Dhont[1,2][0000−0002−5679−0991], Elena Tsiporkova[1], and Veselka Boeva[3][0000−0003−3128−191X]

[1] Sirris, Bd A. Reyerslaan 80, 1030 Brussels, Belgium
michiel.dhont@sirris.be, elena.tsiporkova@sirris.be
[2] Department of Electronics and Information Processing (ETRO), VUB, Belgium
[3] Blekinge Institute of Technoly, Sweden
veselka.boeva@bth.se

Abstract. In this study, we propose a novel data analysis approach that can be used for multi-view analysis and integration of heterogeneous temporal data originating from multiple sources. The proposed approach consists of several distinctive layers: (i) select a suitable set (view) of parameters in order to identify characteristic behaviour within each individual source (ii) exploit an alternative set (view) of raw parameters (or high-level features) to derive some complementary representations (e.g. related to source performance) of the results obtained in the first layer with the aim to facilitate comparison and mediation across the different sources (iii) integrate those representations in an appropriate way, allowing to trace back similar cross-source performance to certain characteristic behaviour of the individual sources.
The validity and the potential of the proposed approach has been demonstrated on a real-world dataset of a fleet of wind turbines.

Keywords: Data Integration · Data Mining · Temporal Data Clustering· Multi-view Learning.

1 Introduction

Mining data collected from continuous monitoring of industrial assets in the field allows to derive relevant insights about their operations and performance. Such complex real-world datasets are usually composed of heterogeneous subsets (or multi-views) of parameters, which should be considered explicitly during analysis in order to exploit fully the richness of the data. For instance, the performance of an industrial asset is impacted by a diverse set of factors e.g. operating modes concerned with the internal working of the asset and exogeneous factors such as weather conditions. However, it is not trivial to directly link or trace back certain performance to distinct operating modes due to the multitude of influencing factors, which are often also highly interdependent.

* This research was supported by the Brussels-Capital Region - Innoviris, and received funding from the Flemish Government (AI Research Program) and BitWind project.

In addition, real-world datasets often originate from different sources, which may differ in period coverage, resolution, data quality, technical configuration, etc. Pooling multi-source datasets together, which is often done to increase statistical representativeness, requires standardization and normalization, which often leads to information loss and may mask source-specific features. For instance, mining for distinct operating modes is more appropriate to be pursued per asset, rather than pooling everything together, since not all assets might go through all operating modes. This implies that one might need to approach multi-source analysis in an incremental fashion rather than aiming for brute force integration of all the available data.

Classical data mining and analysis approaches have still some shortcomings in this aspect aiming at delivering a total integration solution at once. An alternative approach is to **exploit the multi-view nature of the data**. Some rewarding techniques of multi-view mining have been already proposed in the literature [1,14]. However, they all were concerned with single-source datasets and dedicated to one specific mining approach (e.g. clustering, deep learning or classification). This research provides a general analysis methodology, which is agnostic to the specific mining techniques used and focuses on the following key aspects: initial *individual analysis* per source in order to preserve the richness and the authenticity of each source; individual *mediation analysis* per source aiming at bringing the sources closer together; cross-source *integration analysis* aiming at leveraging analysis results across the sources without compromising their individual characteristics.

More concretely, the proposed approach consists of several distinctive layers: (i) select a suitable set (view) of parameters in order to identify characteristic behaviour within each individual source (ii) exploit an alternative set (view) of raw parameters (or high-level features) to derive some complementary representations (e.g. related to source performance) of the results obtained in the first layer with the aim to facilitate comparison and mediation across the different sources (iii) integrate those representations in an appropriate way, allowing to trace back similar cross-source performance to certain characteristic behaviour of the individual sources.

The validity and the potential of the proposed approach have been demonstrated on a real-world dataset of a fleet of wind turbines. We have been able to identify distinctive profiles of production performance and subsequently, have been able to establish an explicit link between those performance profiles and well characterised operating modes.

The rest of the paper is organised as follows. Section 2 reviews related work and discusses the rationale motivating the proposed approach. Section 4 introduces the used methods and formally describes the proposed layered integration approach. Data and experimental setting used for the evaluation purposes are explained in Section 5. Section 6 presents the evaluation of the proposed approach and discusses the obtained results. Section 7 is devoted to conclusions and future work.

2 Related Work and Rationale

Multi-view datasets consist of multiple data representations or views, where each one may contain several features [5]. There are many scenarios where data can be described from multiple views [14]. In such multi-view scenarios it is more interesting to consider the diversity of different views rather than simply concatenating them. Furthermore, remote sensor technologies are very accessible these days, resulting in the appearance of high frequency sensor data collected for all kinds of environments and assets. Despite the accelerated development of mining techniques for multi-source data, managing and interpreting multi-source data is still very challenging. [15]

One way to exploit multi-source data is by data integration. Data integration is the combination of data from distinct data sources into a meaningful and useful format. It can either aim to bring data together for the purpose of visualization or fuse them together in one integrated dataset. Three main approaches have been developed [6]: (i) *Schema mapping*: a global mediating schema is used, e.g. by defining mappings between the distinct schemas of each data source; (ii) *Record linkage*: records that refer to the same entry across distinct data sources are matched together; (iii) *Data fusion*: data from distinct data sources are combined by probabilistic algorithms. One major risk in constructing an integrated dataset is the risk on losing source-specific characteristics.

2.1 Challenges Related to Real-world Datasets

In this research, we consider real-world datasets originating from multiple data sources, e.g. fleet data. An asset within the fleet captures data from multiple sensors and each sensor can moreover have a different accuracy and reliability. Two main issues arise when one wants to mine such complex real-world datasets.

First of all, exploiting fully all the properties of the captured data is not trivial since it is composed out of several **heterogeneous subsets of parameters**. Consider data generated by wind turbines, consisting of sensor data of operational parameters, such as oil temperatureand rotor speed, on the one hand, and data about power production in function of different exogeneous factors such as wind speedand outside temperature, on the other hand. Mining such data considering all the parameters at once is often not the best thing to do since the operational parameters are typically analysed in time, while the power production is better monitored as a function of the weather conditions.

In addition, taking into account and combining the information from the **different sources**, such as the fleet of turbines, is far from trivial. Each source may differ in period coverage, resolution, data quality, technical configuration, etc. To optimally use all information one could pool all multi-source datasets together. However, this requires suitable standardisation and normalisation, which could lead to information loss and may mask source-specific parameters. As example one may want to cluster timestamps according to their behaviour in case of wind turbines. However, rather than pooling everything together it is more appropriate to do that per turbine since not all turbines might go through all operating modes and pooling data would lead to noise/sub-optimal clusters.

2.2 Multi-view Learning

Multi-view learning is a semi-supervised approach that aims to obtain better performance by using the relationship between different views [14]. Multi-view unsupervised learning and specifically multi-view clustering has attracted great attention recently due to availability of inexpensive unlabelled data in many application domains [5]. The goal of multi-view clustering is to find groups of similar objects based on multiple data representations. In the past, multi-view clustering approaches have shown to outperform the single-view clustering approach in case of true single-source multi-view datasets. A multi-view clustering approach uses a conditional independence assumption of the different views [1]. However, a perfect conditional independence of different views is almost impossible in real-world datasets. Fortunately, in [7] one illustrates that in a more realistic case where each group (layer) of parameters is not perfectly independent, a similar approach can also be applied to outperform single-view clustering. The latter is called multi-layer clustering. However, a point of attention in those hierarchical clustering approaches is the tendency to construct too small clusters [1].

Hierarchical approaches are not only advantageous in cluster tasks, but can be used in all kinds of data mining strategies. In [14], a comparison is made concerning multiple multi-view learning techniques. The authors' main conclusion is that multi-view learning is effective and promising in practice, but there is still a lot of work to be done to make them useful in a wide variety of applications.

In this paper we propose a multi-layer data analysis methodology which cleverly benefits from the multi-view approach and demonstrates its potential to deal with multi-source data when applied in a well designed incremental fashion.

3 Use Case Context and Ambition

The proposed layered integration methodology is demonstrated on public sensor data originating from a fleet of wind turbines. The initial ambition of the studied experimental scenario is to identify and characterise potentially different operating modes across the fleet. Notice that wind turbines can have several different operating modes, e.g. working at full speed, reduced speed in order to limit the noise burden on the surroundings, tailored production due to oversupply on the net and others. Subsequently, the ultimate goal is to derive distinctive profiles of production performance and establish an explicit link between those performance profiles and the characterised operating modes.

Two main types of input data sources are used to capture the operation of a wind turbine: operational (endogeneous) and environmental (exogeneous) parameters. The former are referring to sensors measuring the internal working of the turbine, such as oil temperature and rotor speed, while the latter are considering different exogeneous factors impacting the production, such as wind speed and temperature. The performance of a wind turbine is typically expressed in terms of the produced active power as a function of the wind speed, called power curve and visualised as depicted in Figure 2b. A power curve typically has an S shape. Based on this curve, one can derive roughly the expected active power based on a certain wind speed. It is not trivial/possible to determine whether a particular production performance is as expected or there is some deviation since the impact of the internal working of the turbine is not explicitly considered. The same active power output may be induced by different operating modes of the turbine given the same exogeneous context.

The ultimate goal of our analysis is to derive an explicit link between the internal working modes (different compositions of the endogeneous parameters) and the expected output (active power) at fleet level. This will enable for quantitative labelling of the turbine operation with respect to the whole fleet, e.g. "as the rest of the fleet", "under-performing", "better than the fleet".

4 Methods and Proposed Approach

4.1 Clustering Analysis

Three partitioning algorithms are commonly used for data analysis to divide the data objects into k disjoint clusters [10]: k-means, k-medians, and k-medoids clustering. The three partitioning methods differ in how the cluster center is defined. In k-means clustering, the cluster center is defined as the mean data vector averaged over all objects in the cluster. In k-medians, the median is calculated for each dimension in the data vector to create the centroid. Finally, in k-medoids clustering, the cluster center is defined as the object with the smallest sum of distances to all other objects in the cluster.

The partitioning algorithms contain the number of clusters (k) as a parameter and their major drawback is the lack of prior knowledge for that number to construct. Unfortunately, determining a correct, or suitable, k is a difficult problem in a real-world dataset. For such cases, researchers usually generate clustering results for different numbers of clusters, and subsequently assess the quality of the obtained clustering solutions.

In the context of the presented study, we have no prior knowledge about the underlying structure of the data. Thus, we use four internal validation measures for analyzing the data and select the optimal clustering scheme. We have selected two validation measure for assessing compactness and separation properties - *Silhouette Index* [11] and *Davis-Bouldin Index* [4], one for assessing connectedness - *Connectivity* [8], and one for assessing the ratio of the within-cluster variance with the overall-between cluster variance - *Calinski Harabasz Index* [3].

4.2 Kernel Density Estimation (KDE)

As the name suggests KDE is a non-parametric method to estimate the probability density function of a random variable density by use of a kernel. Practically, the KDE f_b' is constructed by averaging the sum of a density estimation for each sample $X_1, X_2 ... X_n$, as shown in Equation (1). In this formula, K is a kernel function of choice, which needs to be symmetric around zero. Often one uses a Gaussian kernel (see Equation (2)) [12].

$$f_b'(x) = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{x - X_i}{b}\right) \qquad (1) \qquad\qquad K(y) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) \qquad (2)$$

In Equation (1), the hyperparameter bandwidth b acts as a smoothing factor. A large b will spread the kernel function, resulting in a very smooth KDE. However, if b gets too large, a lot of information is smoothed out. Since the ground truth is often unknown, some rules of thumb have been developed in the past. Amongst others, Silverman's rule is often used. This rule is defined as $b = (n(d+2)/4)^{-1/(d+4)}$, with d the number of dimensions [13].

4.3 Hypercube Binning Approach

The hypercube approach is a method to characterise (discretize) data by a multi dimensional binning approach. A hypercube is defined as a cube of N dimensions. Hypercube binning can be very useful when analysing multi-dimensional data since by dividing the parameter space into cubes, one can derive properties of interest for each cube. These properties might be for example the median, standard deviation or even the KDE of a (not yet used) parameter. The assumption is that the data points characterized with similar parameter values (so they end up in the same hypercube), exhibit similar properties.

4.4 Layered Multi-view Analysis: General Approach

In this study, we propose a novel approach for analysing complex real-world time series data. It is inspired by some previous study of Boeva et al. [2] dealing with the analysis of high-dimensional multivariate data generated in several different experiments. We have conceived a more generic approach, based on the idea that different in nature data parameters form distinctive views of the data and should be considered for separate analysis in a multilayered fashion.

Suppose that a particular phenomenon (e.g. biological/chemical process, physical asset, etc.) is monitored in time via multiple data capturing measurements of different nature (e.g. experimental setup, machine configuration, high-throughput measurements, operational parameters, exogeneous factors, etc.). This will result in collecting measurements of several parameters that each contains part of the relevant information. Furthermore, data analysis can often benefit from considering (pooling together) data from multiple observations/sources of the phenomenon under study, e.g. in case of biological or chemical processes multiple datasets generated in different experimental conditions are frequently explored together, while in industrial contexts datasets originating from a portfolio or a fleet of industrial assets are often consolidated for analysis.

Subsequently, let us assume that we have access to data of N different sources (e.g. a fleet of wind turbines) of the phenomenon under study monitored via n different types of parameters, which are the same across the different observations/sources, while the time periods covered, the data quality and the capturing resolution are not necessary the same and may vary across the sources.

Formally, the main steps (layers) of the proposed multi-view data analysis approach are explained in the subsections below. The overall data corpus is composed of N different datasets (multi-variate time series) $D_1, D_2, \ldots, D_N$, one per source i $(i = 1, 2, \ldots, N)$. Each individual dataset is composed of n time series $D_i = \{D_{i1}, D_{i2}, \ldots, D_{in}\}$, one per monitored parameter.

Individual Analysis Layer (View 1) This layer is concerned with individual per source data analysis, focusing on a subset of relevant parameters allowing to drill down for insights without the necessity to compromise across all sources.

(a) Select a subset of p common in nature parameters across the different sources based on the following criteria:
   - the selected subset of parameters provides *comprehensive view* about a particular aspect(s) (e.g. behavioural, operational or other characteristics) of the studied phenomenon
   - it is feasible to pool together per individual source the corresponding time series for analysis (e.g. cover the same time window and have the same resolution per observation).

(b) For each source i, the corresponding time series $D_{ij_1}, D_{ij_2}, \ldots, D_{ij_p}$, one per monitored parameter j, $(j = 1, 2, \ldots, p)$, are subsequently integrated into a dataset D_{i_p} of dimensions p by t_i (the size of the covered time window per source i), $(i = 1, 2, \ldots, N)$.

(c) Subsequently, each matrix D_{i_p} per source i, $(i = 1, 2, \ldots, N)$ is individually subjected to a suitable analysis (e.g. clustering, regression or classification).

(d) Thus, for each source i, $(i = 1, 2, \ldots, N)$, the foregoing data analysis step has generated a set of results or data models (e.g. clusters or regression functions) $R_{i1}, R_{i2}, \ldots, R_{ik_i}$, where k_i is a source specific parameter.

Mediation Analysis Layer (View 2) This analysis layer is building upon the results from the previous layer by considering an alternative subset of parameters (view) allowing to derive comparative insights across the sources.

(a) Select a subset of q parameters across all sources based on the criteria:
   - the parameters offer an alternative *complementary view* (representation) of the results obtained per source in the individual analysis layer
   - the obtained complementary representations allow for follow up comparative analysis across the different sources.

(b) For each source i, the corresponding time series $D_{ij_1}, D_{ij_2}, \ldots, D_{ij_q}$, one per selected parameter j, $(j = 1, 2, \ldots, q)$, are subsequently joined together to construct a complementary dataset CD_{il_i} for each result R_{il_i}, $(l_i = 1, 2, \ldots, k_i, i = 1, 2, \ldots, N)$.

(c) Subsequently, each complementary dataset CD_{il_i} per source i, is subjected to a suitable further analyse (e.g. profiling or clustering) leading to complementary results CR_{il_i}, ($l_i = 1, 2, \ldots, k_i$, $i = 1, 2, \ldots, N$). The latter can be easily interpreted and compared across the different sources and are uniquely associated with the corresponding results obtained from the previous layer.

Integration Analysis Layer (Linking the Views) This analysis layer is concerned with leveraging the results obtained in the previous analysis layers across the different sources. The ultimate goal is to derive an explicit link between the results generated in the different views.

(a) The results, obtained for each source in the mediation layer, are pooled together, i.e., the following dataset is composed CR_{il_i}, ($i = 1, 2, \ldots, N$, $l_i = 1, 2, \ldots, k_i$) and subjected to consolidation, e.g. grouping similar results. In this way a cross-source integration is achieved delivering a smaller number of representative, across the different sources, results S_r ($r = 1, \ldots, m$) where $m \leq k_1 + \ldots + k_i$ since each S_r is derived from a subset of CR_{il_i}.
(b) Subsequently, for each S_r ($r = 1, \ldots, m$) a unique link can be established with different subsets of the initial results obtained in the very first individual analysis layer i.e. R_{il_i}, ($i = 1, 2, \ldots, N$, $l_i = 1, 2, \ldots, k_i$). For instance, S_r can potentially define some unique representations or labels of distinctive classes formed by the corresponding R_{il_i} subsets.

4.5 Layered Multi-view Analysis: Instantiated in the Use Case

The layered multi-view analysis approach, introduced in Section 4.4, is instantiated for the considered fleet of wind turbines use case described in Section 3. The overall approach is visualised in Figure 1.

Recall that, two main types of input data sources are used to capture the operation of a wind turbine: operational (endogeneous) and environmental (exogeneous) parameters. The former are referring to sensors measuring the internal working of the turbine, such as oil temperature and rotor speed, while the latter are considering different exogeneous factors impacting the production, such as wind speed, wind direction and temperature.

Individual Analysis Layer: Operating Mode Characterisation (Internal View) This layer is concerned with data analysis only from the perspective of the internal working of each turbine detached from the other influencing factors i.e. based solely on the operational parameters. The aim is to derive clusters of timestamps with characteristic operating behaviour (operating modes) per turbine. Rather than pooling everything together, it is more appropriate to do that per turbine since it may occur that not all turbines go through all operating modes for the considered time period and pooling data would lead to noise/suboptimal clusters. Moreover, the datasets constructed per turbine may differ in period coverage since considering only the common period coverage may lead to a substantial reduction of the data and also mask some source-specific features.

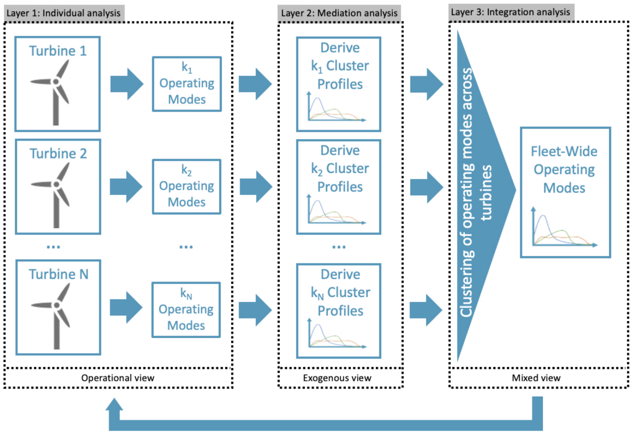

Layer 1: Individual analysis
Layer 2: Mediation analysis
Layer 3: Integration analysis
Turbine 1
k_1 Operating Modes
Derive k_1 Cluster Profiles
Turbine 2
k_2 Operating Modes
Derive k_2 Cluster Profiles
...
Turbine N
k_N Operating Modes
Derive k_N Cluster Profiles
Clustering of operating modes across turbines
Fleet-Wide Operating Modes
Operational view
Exogenous view
Mixed view

Fig. 1. Layered integration approach for the use case of fleet of wind turbines.

Subsequently for each turbine, a number of clusters are derived grouping together timestamps for which the values of the operational parameters relate to each other in a similar way. It is not necessarily expected that the same number of clusters will be derived for each turbine since as already mentioned above not all turbines go through all operating modes for the considered time periods. The assumption is that each cluster is representing a distinctive operating mode of the turbine. Each cluster will define a range of allowable values for each operational parameter and thus generates parametric characterisation of the mode. In this way, the pool of clusters produced for the fleet leads to the construction of a repository of operating modes as depicted in the left panel of Figure 1.

Mediation Layer: Performance Profiling (Exogeneous View) In this layer, we pursue a way to derive an alternative representation of each operating mode in terms of expected performance. The richness of our multivariate data allows to consider an alternative view for each cluster of timestamps generated in the previous layer. For instance, it can be useful for monitoring purposes to have an estimation of how likely is to observe certain production output for a given exogeneous context (i.e., wind speed, wind direction and temperature).

Thus for each cluster of timestamps, from the previous layer, a dedicated dataset can be constructed, composed of the corresponding values for wind speed, wind direction, temperature and active power. Such a dataset can be used to derive some performance profile per cluster estimating the expected production of active power. However, the active power behaviour might vary substantially for different exogeneous contexts or in other words for different combinations of the values of the 3 parameters wind speed, wind direction and temperature. Therefore, we will be pursuing the construction of performance profile per cluster in an incremental fashion by using the hypercube binning approach in order to limit as much as possible the impact of the exogeneous factors. The approach is described in more details below:

1. *Hypercube binning* In order to split the active power points into subsets produced in similar context, i.e. exogeneous parameters with similar values, the hypercube binning approach as explained in Section 4.3 is applied on each cluster dataset. The number of generated hypercubes depends on the granularity of the binning step. The higher the granularity, the more hypercubes will be constructed per cluster, the less points will be contained at average in each hypercube.
2. *Individual probability distributions per hypercube* As described in the previous step, each hypercube represents a group of similar points from perspective of the exogeneous context. Subsequently, the probability density of the active power can be estimated using the KDE approach from Section 4.2
3. *Mixture probability distributions per cluster* The individual distributions derived in the previous step per hypercube in a given cluster are subsequently combined to form mixture distributions for this cluster.

The derived mixture distributions per cluster (see the middle panel of Figure 1) can be interpreted as distinctive probabilistic profiles of the expected performance in terms of active power produced. It is also important to note that the actual operating mode (the ranges of allowable values for each operational parameter) generating this performance profile can be traced back through the cluster characterisation in the previous layer.

Integration Layer: Fleet-wide Performance Labeling (Mixed View) As result of the previous two layers, a repository of operating modes can be constructed, where each operating mode is: 1) characterised in terms of allowable ranges of the operational parameters; 2) associated with a probabilistic profile of expected production. However, the different operating modes have been derived by treating the data of each turbine separately, which does not allow for knowledge transfer and model leverage across the fleet. For instance, considering each set of characterised operating modes per turbine separately is much too limiting since some operating modes might not be observed for some turbines for the considered time window. The latter does not exclude that they might occur in the future. Subsequently, not sharing the operating mode characterisations across the fleet might result into too high rate of unseen operating modes per turbine or in other words high rate of false detection of anomalous operation. Moreover, it is also expected that several different operating modes might be exhibiting very similar production performance.

It is interesting to investigate how many distinct classes/profiles of production performance are detectable at fleet level. The associated with each operating mode probabilistic profiles of expected production can be compared directly with each other since they all are probability density functions of the active power. Subsequently, all the profiles are pooled together and subjected to clustering. In this way, several distinctive profiles of production performance are derived across the fleet (see the right panel of Figure 1) and subsequently, an explicit link between those performance profiles and the characterised operating modes can be established.

5 Dataset and Implementation

The proposed approach has been validated using public SCADA[4] data originating from a wind turbine fleet of Engie, located in La Haute Borne. The dataset contains measurements of a fleet of 4 wind turbines collected with a 10-minute interval for 31 parameters, listed on GitHub. The data is collected between January 2009 and March 2017.

5.1 Data Preprocessing

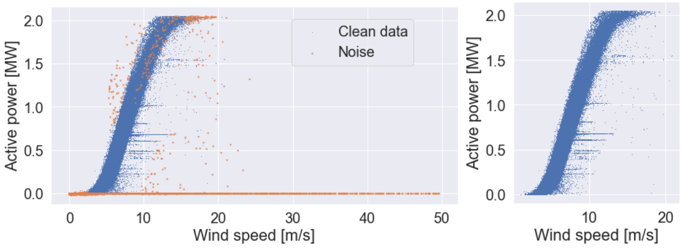
Eliminating Correlated Parameters Some of the monitored parameters in the Engie dataset produce values which are highly correlated due to several reasons 1) monitoring the same phenomenon with multiple sensors, e.g. the nacelle of each turbine is equipped with 2 different anemometers both measuring the wind speed; 2) derived parameters, e.g. the measured wind speed by the two nacelle anemometers is used to calculate the average wind speed; 3) internal dependencies between some parameters, e.g. generator speed and generator converter speed. Therefore in order to avoid over-fitting, only one parameter of the correlated parameters is kept in the experimental dataset, e.g. only the average wind speed is retained, while the values captured by each of the two nacelle anemometers are removed.

Removing Noise Considering that we are dealing with a real-world dataset, it is expected that the data will contain a substantial amount of noise, e.g. outliers, extreme values, etc., which will impact negatively the outcome of the mining if they are not removed. Several different filters based on the most important output parameter active power are applied in order to remove points with an unlikely active power based on their input parameters, by considering each wind turbine separately.

In Figure 2 one can see the effect of this cleaning approach on the power curve based on data from one of the wind turbines in the fleet.

Standardisation The different parameters monitored have values with very different ranges (e.g. the generator bearing temperature varies between -5 and 80 degrees of Celsius, while the generator speed has values between 0 and 1800 rpm), and are of different nature (angular versus non-angular). This makes it very difficult to compare and estimate similarity between parameter values (feature vectors) in time since most of the distance metrics will not perform well.

[4] Supervisory control and data acquisition (SCADA) is an architecture to control industrial systems by use of both external and internal sensors (sources).



(a) Original (raw) data (b) Cleaned data

Fig. 2. Power curve of the remained and cleaned points in one of the wind turbines.

Therefore, angular parameters are transformed into two non-angular values by there sine and cosine value. In the case of the wind direction parameter, we multiply the sine and cosine values with their wind direction. By doing this the information of both wind speed and wind direction are captured into the two new variables. Additionally, min-max normalisation [9] is applied on the parameters across the time window selected for analysis, per wind turbine. In this way, all parameter values are scaled relatively within the same turbine between 0 and 1, which is resulting in much more homogeneous feature vectors per timestamp.

5.2 Implementation and Availability

The proposed Layered Multi-view Analysis methodology has been implemented in Python version 3.6. In our experiments we have used four different cluster validation measures: Silhouette Index, Calinski-Harabasz Index, Davies-Bouldin Index and Connectivity. The first three indices and k-means clustering are used from the Python library Scikit-learn. Connectivity Index, min-max normalization and hypercube binning algorithm have been implemented in Python according to their original descriptions (see Section 5.1). Methods from Python Matplotlib and Seaborn libraries are used for visualisation. We have also used the implementation of KDE and Silverman rule provided by Python SciPy. Finally, Python Pandas library is used for its DataFrame implementation and NumPy library for a couple of mathematical manipulations.

The executable of the Layered Multi-view Analysis algorithm and the experimental results are available on GitHub[5]. The datasets can be found in the website of Engie[6].

6 Results and Discussion

The original public SCADA data of a fleet of 4 turbines have been downloaded and pre-processed individually per turbine applying the different steps described

[5] https://github.com/dataInnovationScientist/LIAMVARWD

[6] https://opendata-renewables.engie.com/explore/

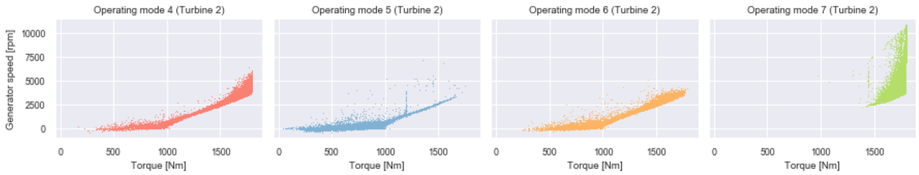


Fig. 3. Torque curves of the operating modes for one of the wind turbines.

in Section 5.1. The binning method used for the removal of outliers utilised bin widths of 0.33 m/s for wind speed and 9 degrees for wind direction. The extreme active power filter was set in such a way that all points with active power more than 500 kW higher than the expected, or more than 1250 kW lower than expected have been removed. A smaller upper threshold is used since the expected active power is quite close to the theoretical maximum which a wind turbine can produce, so there is a certainty that those points are noise.

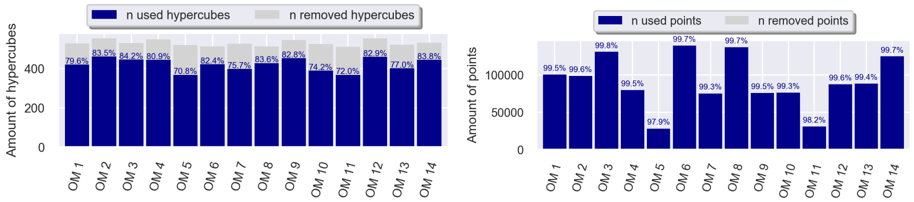
In summary, our pre-processed experimental dataset is covering a period of about 8 years and is split in 4 different datasets, one per turbine in the fleet. In this section we represent and discuss the results obtained by applying the proposed multi-view data analysis approach as outlined in Section 4.5.

6.1 Individual Analysis Layer: Operating Mode Characterisation

This layer is concerned with the internal working of wind turbine. The following selection of 12 endogeneous parameters, which are the ones retained after eliminating correlated parameters (Section 5.1), are considered: sine and cosine of the pitch angel, generator speed, generator bearing temperature 1 and 2, generator stator temperature, gearbox bearing 1 and 2 temperature, gearbox inlet temperature, gearbox oil sump temperature, rotor bearing temperature and torque.

In what follows, we will refer to the 12 parameters as $p_1, p_2, \ldots, p_{12}$ following the order in which they are listed. Subsequently, the k-means clustering algorithm has been applied on the 4 datasets, one per turbine, composed of the 12 parameters. The optimal amount of clusters (k) per turbine was determined by applying a majority voting (Section 4.1), resulting in $k = 3$ for two of the turbines and $k = 4$ for the other two. The difference between the obtained clusters is illustrated in terms of the behaviour of the torque curve (torque as a function of the generator speed), as depicted for one of the four turbines in Figure 3. The torque curve being derived from an endogeneous parameter is better suited to illustrate difference in operational behaviour rather than the most frequently used power curve.

In total, 14 clusters (operating modes) have been derived. The assumption is that each cluster is representing a distinctive operating mode of the turbine. Each operating mode is characterised in terms of the allowable ranges of each of the 12 internal parameters. Those can be consulted on our GitHub repository.



(a) Hypercubes (b) Data points

Fig. 4. Percentage of retained data per cluster after removal of sparse hypercubes.

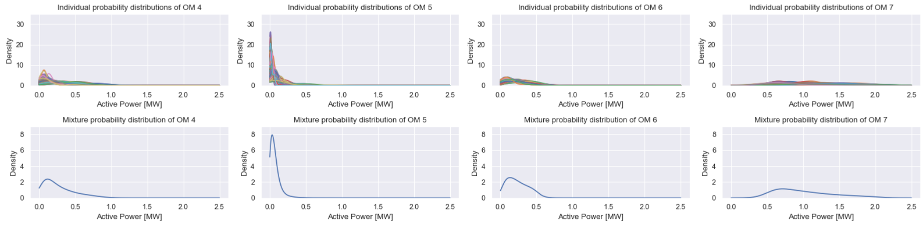


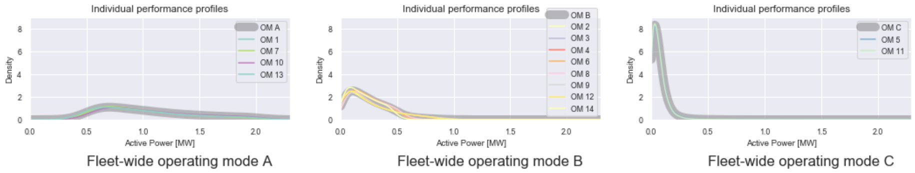
Fig. 5. Performance profiles per operating mode for wind turbine 2.

6.2 Mediation Layer: Performance Profiling

In this layer, performance profiles are derived for each operating mode following the steps outlined in Section 4.5. The corresponding values for temperature, wind speed and wind direction (after their non-angular transformation as stated in Section 5.1) per cluster are binned together using the hypercube approach (see Section 4.3) and the corresponding active power values per hypercube are used to compute a KDE using Gaussian kernel with Silverman's rule (see KDE Section 4.2).

Although it was expected that the KDE computation might be influenced by the number of points in each hypercube, or indirectly by the binning granularity, experiments with different sizes of the hypercubes demonstrated very robust KDE computation w.r.t. varying bin sizes. The results presented in the study have been obtained by splitting the solution space into 2250 equal size hypercubes, where each operating mode has around 500 hypercubes containing data points. Subsequently, sparse hypercubes (with less than 10 points) have been removed for the sake of statistical representativeness. The latter did not lead to substantial information loss since as it can be witnessed in Figure 4, the retained around 5% of the hypercubes for each cluster contain more than 97% of the original data points.

Subsequently, the mixture probability distribution for each cluster (operating mode) is derived as outlined in Section 4.5. Figure 5 depicts for one of the turbines the individual probability distributions derived for the different hypercubes and the corresponding mixture probability distributions per cluster.



Fleet-wide operating mode A

Fleet-wide operating mode B

Fleet-wide operating mode C

Fig. 6. Fleet-wide performance profiles and their corresponding individual components.

6.3 Integration Layer: Fleet-wide Performance Labelling

In this layer, the obtained performance profiles (mixture distributions) per individual operating mode have been pooled together and subjected to k-means clustering. The optimal number of clusters $k = 3$ has been derived as previously described by applying a majority voting (Section 4.1). Subsequently, 3 fleet-wide performance profiles (higher level mixture distributions) have been computed for the three clusters by combining the corresponding performance profiles (mixture distributions). The mixture weights have been computed as the number of points in the corresponding cluster from layer 1, normalised by the total number of points in the given fleet-wide cluster. The resulting very distinctive fleet-wide performance profiles (A, B and C) are depicted in Figure 6.

Note that each of the fleet-wide performance profiles can be traced back to a subset of individual operating modes (by use of the table constructed in Section 6.1 and available for consultation on our GitHub repository), resulting in fleet-wide (composite) operating modes, which we also denote with A, B and C: $A = \{1, 3, 5, 6, 8, 9, 12, 13\}$; $B = \{2, 4, 10, 14\}$; $C = \{7, 11\}$. It is interesting to observe that the composite operating mode linked to profile C can be traced back to only two of the four wind turbines.

The derived fleet-wide (composite) operating modes, each associated with a very distinctive performance profile (see Figure 6), can now be used to label the fleet data as follows: 1) for each timestamp, consider the values of the 12 operational parameters; 2) determine to which operating mode they can be assigned (based on the table constructed in Section 6.1); 3) identify the composite operating mode to which the identified mode belongs; 4) subsequently, assign the corresponding letter A, B, C or D (not seen) to the timestamp. In this way, each dataset per turbine can be converted into A, B, C or D code (as a DNA sequence), which can be very insightful for monitoring purposes (e.g. long periods of B would signify optimal performance), but is also a powerful representation enabling more advanced applications, e.g.: mining the fleet data for interesting patterns such as transitions between operating modes; zooming in periods with too many Ds; training a predictor of expected production on historical data to be used to detect deviations during real-time operations.

7 Conclusion and Future Work

We have proposed a novel data analysis approach that can be used for multi-view analysis and integration of heterogeneous real-world datasets originating from multiple sources. The validity and the potential of the proposed approach has been demonstrated on a real-world dataset of a fleet of wind turbines. The obtained results are very encouraging. The method is very efficient and robust in detecting characteristic operating modes across the fleet. Subsequently, distinctive performance profiles are derived and associated with each operating mode, which enable converting the fleet data into powerful letter code suitable for more advanced mining.

For future work, we are interested to extend our research in the following directions: 1) fine-tune the method by using e.g. an adaptive hypercube binning; 2) testing different experimental scenarios e.g. comparing different time periods from the same wind turbine; 3) consider additional validation use cases dealing with multi-source datasets e.g. mobility or manufacturing data; 4) extend further the method by exploiting the possibility to covert the fleet data into letter code.

References

1. Steffen Bickel and Tobias Scheffer. Multi-view clustering. In *ICDM*, volume 4, pages 19–26, 2004.
2. Veselka Boeva, Elena Tsiporkova, and Elena Kostadinova. Analysis of multiple dna microarray datasets. In *Springer Handbook of Bio-/Neuroinformatics*. 2014.
3. Tadeusz Caliński and Jerzy Harabasz. A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, 3(1):1–27, 1974.
4. David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, 2:224–227, 1979.
5. P Deepak and Anna Jurek-Loughrey. *Linking and Mining Heterogeneous and Multi-view Data*. Springer, 2018.
6. Xin Luna Dong and Divesh Srivastava. Big data integration. In *2013 IEEE 29th international conference on data engineering (ICDE)*, pages 1245–1248. IEEE, 2013.
7. Dragan Gamberger, Matej Mihelčić, and Nada Lavrač. Multilayer clustering. In *International Conference on Discovery Science*, pages 87–98. Springer, 2014.
8. Julia Handl, Joshua Knowles, and Douglas B Kell. Computational cluster validation in post-genomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005.
9. Zhenyu Liu et al. A method of svm with normalization in intrusion detection. *Procedia Environmental Sciences*, 11:256–262, 2011.
10. James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proc. of 5th Berkeley symposium on stat. and prob.*, 1967.
11. Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20, 1987.
12. Simon J Sheather. Density estimation. *Statistical science*, pages 588–597, 2004.
13. Bernard W Silverman. *Density estimation for statistics and data analysis*, volume 26. CRC press, 1986.
14. Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.
15. Jixian Zhang. Multi-source remote sensing data fusion: status and trends. *International Journal of Image and Data Fusion*, 1(1):5–24, 2010.