Data-driven virtual sensing and dynamic strain estimation for fatigue analysis of offshore wind turbine using principal component analysis

Tarpø, Marius; Amador, Sandro; Katsanos, Evangelos; Skog, Mattias; Gjødvad, Johan; Brincker, Rune

Published in:
Wind Energy

Link to article, DOI:
10.1002/we.2683

Publication date:
2022

Document Version
Publisher's PDF, also known as Version of record

Link back to DTU Orbit

Citation (APA):
Data-driven virtual sensing and dynamic strain estimation for fatigue analysis of offshore wind turbine using principal component analysis

Marius Tarpø1 | Sandro Amador1 | Evangelos Katsanos1 | Mattias Skog2 | Johan Gjødvad2 | Rune Brincker1

1Department of Civil Engineering, Technical University of Denmark, Kongens Lyngby, Denmark
2Development, Sigicom, Älvsjö, Sweden

Correspondence
Marius Tarpø, Department of Civil Engineering, Technical University of Denmark, Brevej B.118, Kongens Lyngby, Denmark. Email: mariotarpoe@gmail.com

Abstract
Virtual sensing enables estimation of stress in unmeasured locations of a system using a system model, physical sensors and a process model. The system model holds the relationship between the physical sensors and the desired stress response. A process model processes both the physical sensors and the system model to synthesise virtual sensors that ‘measure’ the desired stress response. Thus, virtual sensing enables mapping between the physical sensors (input) and the desired stress response (output). The system model is a mathematical model of the system based on knowledge or data of the system. Here, the data-driven system model is constructed directly on data analyses for the specific system. In this paper, supervised learning and data-driven system models are applied to strain estimation of an offshore wind turbine in the dynamic range through a novel use of principal component analysis (PCA); 40 min of training data is used to establish the data-driven system model that can estimate the dynamic strain response with high precision for 2 months, while the estimated fatigue damage averaged out to \( -1.76\% \) of the measured strain response.

KEYWORDS
data-driven model, principal component analysis, stress estimation, structural health monitoring, supervised learning, virtual sensing

1 INTRODUCTION

Virtual sensing is a technique that expands measured data to unmeasured locations and/or transforms it into other quantities by synthesising virtual sensors. The stress/strain estimation is a subcategory within virtual sensing that estimates the full-field stress/strain response of a system. Stress or strain estimation is known by many names and terms: stress/strain prediction, reconstruction of unmeasured stress/strain, fatigue prediction/estimation, hybrid modal analysis, full-strain fields, full-field stress/strain estimation, full-field stress/strain distribution, virtual sensing, soft sensing, full-state estimation and so forth. Unfortunately, there is a lack of consensus and common terminology in the field for virtual sensing. In this paper, we will use the terminology proposed by Marius Tarpø. In this terminology, we need three components to synthesise virtual sensors: system model, physical sensors and process model, as outlined in Figure 1. The system model could have any form and
In this paper, we will study supervised learning and data-driven strain estimation on an operating offshore wind turbine for 2 months. Data-driven system models consist of relationships between the states (input and output) of a system, and these develop purely from data without any explicit knowledge of the system. Thus, data-driven strain estimation enables a transformation into strain if the system model is based on both other sensors and strain gauges. The reader should note that the data-driven system model is limited by the data upon which it is based. Therefore, virtual sensors are restricted to the instrumented locations from the training dataset. The reasoning behind data-driven strain estimation is that strain gauges may lose their reliable performance over long-term monitoring—especially in an offshore environment. Thus, we can...
utilise temporary strain gauges along with geophones or accelerometers to establish a data-driven system model. This enables us to estimate the strain history in the same location as the temporary strain gauges using only the geophones or accelerometers. In this paper, we utilise mere 40 min of training data to establish a data-driven system model that can estimate the strain history over 2 months with high precision.

The scope of this paper is to propose a new data-driven virtual sensing technique for stress estimation that contributes to a more sustainable and stable operation of wind turbines. The proposed technique is applied to the data collected in a long-term structural health monitoring (SHM) campaign to improve the robustness of the strain measurements over time. It is well known that vibration-based SHM is a reliable and efficient structural assessment tool in detecting early stage structural damages, extending the lifetime of the monitored structures, as well as in reducing structural maintenance costs by avoiding unnecessary on-site inspections. In this paper, the proposed technique is verified for stress estimation by an implementation in the dynamic and broad frequency range on an offshore wind turbine. We organise the remainder of the paper as follows: We introduce the theory of data-driven strain estimation using principal component analysis (PCA) in Section 2. Section 3 presents the study of data-driven strain estimation on an operating offshore wind turbine during a 2-month monitoring campaign.

2 | STRAIN ESTIMATION

In this paper, we follow the flow chart layout from Tarpø² (see Figure 1). Following this terminology, we require three components to synthesise a virtual sensor: the system model, the physical sensors and the process model. Generally, the performance of the virtual sensors depends on the combination of physical sensors, system model and process model, which complicates any evaluation of a virtual sensor. Any potential error, including measurement errors and signal-to-noise ratio of the physical sensors, modelling errors of the system model, the sensitivity of the reduced system model, processing errors in the process model, and violations of the assumptions for the system and process model, can propagate and transfer into the virtual sensor. Hence, this complex network of potential errors determines the performance and quality of the virtual sensor.

In the mathematical modelling of the system model, we can model the system directly on the available data from our physical sensors where we have either fully data-driven (empirical) system models or partial data-assisted system models.¹⁹,²⁰,²² This process requires temporary reference sensors at the desired locations to train the virtual sensors. After the creation of the system model, we can remove and replace the temporary reference sensor with the virtual sensors. Data-driven system models build relationships within the data, while the data can calibrate and assist the mathematical modelling for the data-assisted system model. These system models hypothesise that they contain less modelling errors since they are based on the data of the system. The reader should note that the fully data-driven system model is limited by the data upon which it is based. Thus, we are unable to extend the system model beyond the data without adding information to the system model.

2.1 | Data-driven system model

There are many different ways to create a data-driven or partial data-assisted system model like machine learning, PCA and transformation matrix created directly from data. Generally, these applications require training datasets to set up the system model, and the quality of the system model depends on the training dataset representation of the actual conditions. Examples of data-driven system models can be found in Lu et al. and Deng et al.¹⁹,²⁰ In this paper, we will apply PCA to create a transformation matrix that transforms our displacement directly into the strain response in the same locations as we had strain gauges in the training dataset.

The idea behind PCA is essentially a dimensionality reduction of data to set of uncorrelated principal components that are ordered after their contribution to the original data.²³ Furthermore, we can regard PCA as an unsupervised version of linear regression.²¹ In this paper, however, we use the PCA in a supervised learning approach to uncover a mapping between the displacement of a system (the input) and the strain response of that system (the output) based on training data.

First, we stack the displacement vector, \( \mathbf{y}(t) \in \mathbb{R}^{n_y} \), and the strain response, \( \mathbf{\varepsilon}(t) \in \mathbb{R}^{n_\varepsilon} \).

\[
\mathbf{y}(t) = \begin{bmatrix} \mathbf{y}(t) \\ \mathbf{\varepsilon}(t) \end{bmatrix}
\]

where \( \mathbf{y}(t) \in \mathbb{R}^{N} \) is the stacked response vector and \( N = n_y + n_\varepsilon \).

This stacked vector, \( \mathbf{y}(t) \), has a subspace, \( V \), which is spanned by the column vectors, \( \mathbf{v} \). Furthermore, we assume that these vectors are unknown.

\[
V = \text{span}\{ \mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_p \}
\]
These column vectors form a basis for the subspace \( V \), and we form a transformation matrix with these vectors.

\[
V = [v_1 \ v_2 \ ... \ v_n]
\]  

(3)

We can express the stacked response vector as a linear combination of these column vectors since these vectors span the subspace of the stacked vector, \( \mathbf{y}_c(t) \).

\[
\mathbf{y}_c(t) = \mathbf{V} \tilde{\mathbf{q}}(t)
\]  

(4)

We want to estimate these column vectors by applying PCA. First, we calculate the covariance matrix of the stacked response.

\[
\mathbf{C} = \mathbb{E}[\mathbf{y}_c(t)\mathbf{y}_c(t)^T] = \mathbf{V} \mathbb{E}[\tilde{\mathbf{q}}(t)\tilde{\mathbf{q}}(t)^T]\mathbf{V}^T
\]  

(5)

where \( \mathbb{E}[\cdot] \) denotes the expectation operator and \( (\cdot)^T \) denotes the transpose of a vector or matrix.

Then we apply singular value decomposition on the covariance matrix, which is symmetric.

\[
\mathbf{C} = \mathbf{USU}^T
\]  

(6)

where \( \mathbf{S} \in \mathbb{R}^{N \times N} \) is a diagonal matrix holding the singular values in descending order and \( \mathbf{U} \in \mathbb{R}^{N \times N} \) holds the singular vectors as column vectors, and they correspond to the singular values.

The singular vectors are approximately equivalent to the transformation matrix, \( \mathbf{V} \approx \mathbf{U} \), and they are estimates of the column vectors, \( \mathbf{v}_i \). Furthermore, the singular vectors have components related to displacement and strain.

\[
\mathbf{U} = \begin{bmatrix} \mathbf{U}_y \\ \mathbf{U}_\varepsilon \end{bmatrix}
\]  

(7)

where \( \mathbf{U}_y \in \mathbb{R}^{n_y \times N} \) is the part of the singular vectors related to displacement and \( \mathbf{U}_\varepsilon \in \mathbb{R}^{n_\varepsilon \times N} \) is the part related to the strain response.

We can use the singular values, \( \mathbf{S} \), to detect the number of principal components needed to represent the subspace of the stacked response vector since the singular values show the contribution of each column vector, \( \mathbf{u}_n \). Singular values below a certain threshold indicate that the corresponding vectors have insignificant contributions to the response and they are removable. Furthermore, an additional indicator is the condition number of singular vectors related to the displacement, \( \mathbf{U}_y \), where this metric indicates the stability of the model. There is a trade-off between the precision and stability in the form of the number of principal components—indicated by the singular values—and the condition number. We can truncate the singular vectors to the first \( n \) components.

\[
\hat{\mathbf{U}} = [u_1 \ u_2 \ ... \ u_n]
\]  

(8)

In this paper, we propose a displacement-to-strain matrix based on the singular vectors and linear regression.

\[
\mathbf{H} = \hat{\mathbf{U}} \hat{\mathbf{U}}^\dagger_y
\]  

(9)

where \( \mathbf{H} \in \mathbb{R}^{n_\varepsilon \times n_y} \) is the displacement-to-strain matrix and \( (\cdot)^\dagger \) denotes the Moore–Penrose inverse.

This transformation matrix, \( \mathbf{H} \), holds the relationships between the displacement and the strain response of the system; therefore, it enables a mapping between these quantities. Thus, we can estimate the strain response in any other dataset or continuously in real time as

\[
\mathbf{\varepsilon}(t) = \mathbf{H}\mathbf{y}(t)
\]  

(10)

Here, the singular vectors form the system model, while the pseudo-inverse is the process model that transforms displacement into strain. Since the proposed technique is based on regression, the reader should note that it is similar to the modal expansion technique, which utilises mode shapes instead of singular vectors. The transformation matrix works as a subspace reduction that removes any noise or response perpendicular to the new subspace of the singular vectors. Moreover, we must avoid an ill-posed inverse problem; hence, the
transformation matrix requires a redundant sensor network with more physical sensors than the included singular vectors. Fortunately, we can somewhat bypass this limitation by separating the data into frequency bands using complementary filters and apply a new subset of the singular vectors for each band. In that case, the data are divided into frequency bands whenever the condition number of all relevant singular vectors related to displacement is too large. Furthermore, frequency ranges with low levels of response should be separated into individual frequency bands since the principal components from higher response levels will otherwise dominate.

3 | CASE STUDY

3.1 | Quality measures

There exist no rules or guidelines for evaluating virtual sensors. We can evaluate a virtual sensor like a physical sensor in terms of range, repeatability, sensitivity and accuracy. Furthermore, we should also evaluate the robustness and reliability of the virtual sensors. The most common approach is, however, to compare the output of virtual sensors with a set of reference sensors through different quality measurements with different strengths and weaknesses.

3.1.1 | Time domain quality measures

The coefficient of determination ($R^2$)\(^{25,26}\) is a popular quality measure in statistics and modal validation to check the correlation between a reference and an estimated quantity. The metric is equal to the mean square error (MSE) of the two quantities normalised with the variance of the reference quantity. Whereas the value of the MSE is relative to a specific dataset, coefficient of determination ($R^2$) is comparable between any datasets. Therefore, the coefficient of determination accounts for both amplitude differences and the general correlation between the quantities. A coefficient of determination with a value of 1 indicates perfect correlation with the same amplitudes. This metric should not be confused with the Pearson correlation coefficient.

\[
R^2 = 1 - \frac{E[(\epsilon(t) - \hat{\epsilon}(t))^2]}{\text{Var}[\epsilon(t)]}
\]

where $\epsilon_i(t)$ is the measured strain response for the $i$th strain gauge, $\hat{\epsilon}_i(t)$ is the estimated strain response at the same location and $\text{Var}[\cdot]$ denotes the variance operator.

3.1.2 | Fatigue damage quality measures

It is equally important to evaluate the stress estimation in fatigue damage when the estimation is intended for a fatigue analysis. Using the SN curve, the fatigue life is expressed as the following function:

\[
N_i\sigma_m^m = C
\]

where $N_i$ is the fatigue life for the given stress amplitude, $\sigma_m$ is the ‘slope’ of the SN curve and $C$ is the fatigue capacity (or intercept on the $N$ axis at a stress amplitude of 1).

In the case the amplitude of the stress amplitudes varies, we must apply cycle counting\(^{27}\) such as the rainflow counting technique,\(^{28}\) to count all stress cycles. According to the Palmgren–Miner rule, the accumulated fatigue damage is a summation of all partial damage caused by each stress cycle.\(^{27}\)

\[
D = \sum_{i=1}^{n_{\text{cycles}}} D_i \approx \sum_{i=1}^{n_{\text{cycles}}} \frac{1}{N_i}
\]

where $n_{\text{cycles}}$ is the number of stress cycles.

For this propose, we apply the normalised error of fatigue damage (NEFD),\(^{16}\) which is based on the SN curve (excluding the effect of a bilinear SN curve) and the Palmgren–Miner rule.
\[ \eta_i = \frac{\bar{D}_n - \hat{D}_n}{\bar{D}_n} = \frac{\sum_{j=1}^{n_{\text{cycles}}} \frac{\Delta \sigma^m_j}{C}}{\sum_{j=1}^{n_{\text{cycles}}} \frac{\Delta \sigma^m_j}{C}} - 1 = \frac{\sum_{j=1}^{\hat{n}_{\text{cycles}}} \Delta \hat{\sigma}^m_j}{\sum_{j=1}^{\hat{n}_{\text{cycles}}} \Delta \hat{\sigma}^m_j} - 1 \]  

where \( D_n \) and \( \bar{D}_n \) are the accumulated fatigue damage for the measured and estimated signal, respectively, in the \( n \)th fatigue location; \( n_{\text{cycles}} \) and \( \hat{n}_{\text{cycles}} \) are the total number of counted cycles for the measured and estimated signal, respectively; \( \Delta \sigma \) and \( \Delta \hat{\sigma} \) denote the stress range from cycle counting; and \( \Delta \epsilon \) and \( \Delta \hat{\epsilon} \) are the strain range from cycle counting. Here, \( \eta_i = 0 \) indicates a correct strain estimation in terms of fatigue damage, while a negative value suggests an underestimation and a positive value suggests an overestimation of fatigue damage. In this paper, we will apply \( m = 3 \), corresponding to welded steel structures without corrosion protection.\(^{29}\)

### 3.1.3 Normalised error in equivalent stress range

It is often preferable to access fatigue damage for variable amplitude loading in the form of a damage equivalent stress range. These stress ranges lead to the same fatigue damage as all stress ranges.\(^{27}\)

\[
\Delta \sigma_{eq} = \left( \frac{1}{n_{\text{cycles}}} \sum_{j=1}^{n_{\text{cycles}}} \Delta \sigma^m_j \right)^{1/m}, \Delta \hat{\sigma}_{eq} = \left( \frac{1}{\hat{n}_{\text{cycles}}} \sum_{j=1}^{\hat{n}_{\text{cycles}}} \Delta \hat{\sigma}^m_j \right)^{1/m}
\]

The normalised error of the equivalent stress range is given as

\[
\xi_i = \frac{\Delta \hat{\sigma}_{eq} - \Delta \sigma_{eq}}{\Delta \sigma_{eq}}
\]

### 3.2 Wind turbine

The studied structure is an offshore wind turbine—Vestas V90, 3 MW—positioned in the Great Belt near Sprogø in Denmark with a gravitation foundation (see Figure 2). The wind turbine is equipped with four Sigicom V12 triaxial geophones attached approximately 17 m above the foundation, while four strain gauges are placed at 0.75 m above the foundation (see Figure 3A for the elevation of the sensors). The strain gauges measure along the longitudinal direction of the tower, and they measure, in pairs, in opposite sides of the cross section (see Figure 3B). The reader should note that the position of sensors limits the measured response to the tower modes. Furthermore, the geophones have a frequency range...

---

FIGURE 2 Wind turbine at Sprogø near the Great Belt Bridge in Denmark [Colour figure can be viewed at wileyonlinelibrary.com]
with the low cut-off frequency of approximately 0.2 Hz, and measurement noise will dominate at lower frequencies. Therefore, the physical sensors restrict this case study to dynamic strain estimation—above 0.2 Hz.

The monitoring campaign was during a 2-month period (see Figure 4), where the sampling frequency was 20 Hz. In this study, we apply a high-pass filter to all data with a cut-off frequency of 0.2 Hz to exclude the quasi-static response of the wind turbine. The geophones are calibrated using digital correlation, and we use the vertical geophones to reduce the tilt effects on the geophones. We use the integration theorem for the Fourier transformation to integrate signals from the geophones into displacement. Afterwards, we divide the data into three frequency bands: 0.2–0.7, 0.7–4.5 and 4.5–10 Hz using complementary filters. We apply these filters to avoid an ill-posed problem for the data-driven system model, and the use of multiple frequency bands increases stability in the system model.

In this monitoring campaign, we found intermittent spikes in the measured strain response when wind turbine produced an average power near zero or negative (see Figure 5A). In this case, strain gauges 1 & 3 and 2 & 4 should measure the same quantity but the opposite operational sign. The intermittent spikes, however, have the same operational sign. Thus, concluding, the noise spikes are noise, which could be caused by electrical interference on the strain gauges that only occurs at or around negative power production. We reduce the noise on the strain gauge by subtracting the pair of strain gauge that measures the same quantity with opposite operational sign and divide with two (see Figure 5B) for the calibrated strain gauges.

3.3 Virtual sensing and strain estimation

We use 40 min of training data to create the data-driven system model. Figure 4 marks the training data with a wind direction of approximately 180° and a low-power production. We apply PCA as explained in Section 2.1 to generate the system model in the form of a transformation matrix in each frequency band. For the first frequency band, we utilise six principal component vectors, six principal vectors in the second band and two principal vectors in third frequency band to generate the three transformation matrices. We base the number of singular vectors on the singular values of each frequency band corresponding to the expected number of dominating modes. Figure 6 shows the results of the trained estimation of strain along with the measured strain for the training dataset.
For validation of the data-driven system model, we use a total of 6100 datasets with a time length of 10 min equally distributed over the entire monitoring campaign of 2 months, corresponding to 100 datasets each day. We present the results of the strain estimation in box plots for three quality measurement: Figure 7A shows the coefficient of determination ($R^2$), Figure 7B shows the error of equivalent strain range and Figure 7C shows the NEFD.
We look at the dataset with the worst strain estimation to analyse the result (see Figure 8). For this specific dataset, the wind turbine does not produce power (parked conditions) and the standard deviation of the measured strain is very low, and therefore, the signal-to-noise level decreases. Figure 8A shows that the measured strain response is dominated by the noise floor at −90 dB and the structural response only
exceeds the noise floor at 0.35 Hz, whereas the estimated strain response is mostly located below this noise floor. Therefore, the virtual sensors result in an underestimation of fatigue damage in comparison to the strain gauges. Concluding, the virtual sensor has a higher signal-to-noise ratio than the strain gauges. When the wind turbine does not produce power, the strain gauges act as a poor reference sensors since measurement

**FIGURE 8** The dataset corresponding to the worst estimation of strain with measured strain (black line) and estimated strain (red line): (A) frequency domain with the first three singular values of the spectral density matrix calculated using Welch averaging method with segments of 2048 data points and 50% overlap and (B) time domain with a zoom on the strain response [Colour figure can be viewed at wileyonlinelibrary.com]

**FIGURE 9** Without datasets with negative power production, Tukey box plot with the 2nd, 25th, 50th, 75th and 98th percentile for the estimated and measured strain for the two strain gauges
noise dominates the measured strain response. In that case, a comparison between the virtual sensors and the strain gauges is unfair since the virtual sensors could better represent the actual strain response than the strain gauges.

To make a fairer assessment of virtual sensors, we excluded the datasets for which the wind turbine does not produce power since the strain gauges, in these cases, mainly measure noise and therefore provide a poor frame of reference for the strain estimation. We could alternatively have applied low-pass filters to exclude the noise floor of both virtual sensors and strain gauges but that would enable a narrow frequency band for the comparison, which is outside the scope of this paper. Figure 9A–C shows a box plot of the coefficient of determination ($R^2$), error in equivalent stress range and NEFD, respectively, without datasets with negative power production. The median values do hardly change, but the variance of all quality measurements decreases without these datasets.

The data-driven virtual strain gauges have a higher signal-to-noise ratio than the strain gauge, and it is not prone to intermittent noise spikes whenever the wind turbine produced an average power near zero or negative. They seem reliable and more robust than the actual strain gauges in the long-term monitoring.

4 | CONCLUSION

In this paper, PCA was applied to generate a data-driven strain estimation technique in a supervised learning approach. It requires training data with temporary strain gauges in all relevant locations that are removable after the data-driven system model is generated. In this way, the monitoring system becomes independent of the strain gauges, which can be corrupted and are prone to sensor fault over time.

The approach was applied to an offshore wind turbine within the framework of a 2-month monitoring campaign; 40 min of training data resulted in a data-driven system model that could estimate the strain response, using only geophones, within the 2 months with high precision. In this study, the virtual sensors performed reliably and robustly, while their output exhibited a high resemblance to the measured strain response with coefficients of determination ($R^2$) above 0.99 on average. Furthermore, the estimated fatigue damage was on average −1.76% of the measured strain, while the uncertainty of the estimated equivalent strain range was 2.95% of the measured.

Moreover, the virtual sensors have a higher signal-to-noise ratio than the strain gauges since they inherit the higher signal-to-noise ratio from the geophone. Furthermore, the virtual sensors seemed more reliable and robust than the actual strain gauges in this monitoring campaign.

In future research, the proposed technique should include the quasi-static response by utilising sensors that accurately measure low-frequency vibrations. Generally, the quasi-static strain response of wind turbines is a significant part of the fatigue life, and therefore, it should be included in a strain estimation technique. Furthermore, the process of setting up the frequency bands and the corresponding number of principal components should be automated. Additionally, a generalised assessment of the robustness and reliability of virtual sensing should be researched and established.

PEER REVIEW

The peer review history for this article is available at https://publons.com/publon/10.1002/we.2683.

DATA AVAILABILITY STATEMENT

The authors elect to not share data.

ORCID

Marius Tarpø https://orcid.org/0000-0002-7480-6408
Sandro Amador https://orcid.org/0000-0003-3045-2916
Evangelos Katsanos https://orcid.org/0000-0003-2574-5327
Rune Brincker https://orcid.org/0000-0002-5344-1649

REFERENCES

3. Karczub DG. The prediction of dynamic stress and strain in randomly vibrating structures using vibrational velocity measurements. Ph.D. Thesis; 1996. This thesis has been made available in the UWA Profiles and Research Repository as part of a UWA Library project to digitise and make available theses completed before 2003. If you are the author of this thesis and would like it removed from the UWA Profiles and Research Repository, please contact digitaltheses-lib@uwa.edu.au (1996). doi:10.26182/5ca2fdde34621