System Identification at the Extreme Edge for Network Load Reduction in Vibration-Based Monitoring

Federica Zonzini, *Graduate Student Member, IEEE*, Vasilis Dertimanis, Eleni Chatzi, and Luca De Marchi, *Senior Member, IEEE*

***Abstract*—Mechanical complexity, wide dimensions, and big data volume may hamper the implementation of Internet of Things (IoT)-enabled structural health monitoring (SHM) systems. In particular, one of the most important challenges is the reduction of the data payload to be transmitted over the monitoring network. Addressing the problem in the context of vibration-based SHM, this work explores system identification (SysId) as an innovative strategy for data compression at the extreme edge. Indeed, SysId is a signal processing technique aiming at finding a very reduced (i.e., less then one tenth of the total signal length) set of meaningful parameters, which can provide an alternative, but yet completely equivalent, frequency characterization of the structure. In the proposed approach, an embedded system-oriented adaptation of the sequential tall-skinny QR decomposition (eS-TSQR) from the dense linear algebra domain has been exploited to tackle both the memory and computational complexity of the involved algorithms. This yielded to the embodiment of input–output and output-only SysId models into a resource constrained device (i.e., an STM32L5 microcontoller unit), targeted on low-power and low-cost SHM applications, proving high effectiveness for the structural assessment of civil and industrial plants. Besides, a cost-benefit analysis is also presented, in which the energy saving brought by SysId running in a sensor-near manner is comprehensively measured against the power consumption due to data transmission, as implied by state-of-the-art communication protocols for IoT. Results demonstrate that SysId is 1.19× and 2.78× less energy demanding (with a payload reduction of 9× and 45×) w.r.t. compressed sensing-driven and compression-free solutions, respectively.**

***Index Terms*—Data compression, edge processing, parametric system identification (SysId), structural health monitoring (SHM), tall-skinny QR decomposition, vibration analysis.**

Manuscript received 11 February 2022; revised 12 April 2022; accepted 13 May 2022. Date of publication 20 May 2022; date of current version 7 October 2022. The work of Federica Zonzini and Luca De Marchi was supported by the European Union's Horizon 2020 Research and Innovation Program through Arrowhead Tools Project under Grant 826452. The work of Eleni Chatzi was supported by the Horizon 2020, the EU's Framework Programme for Research and Innovation, under Grant 769373 (Project: FORESEE). *(Corresponding author: Federica Zonzini.)*

Federica Zonzini is with the Advanced Research Center on Electronic Systems for Information and Communication Technologies "Ercole De Castro", University of Bologna, 40136 Bologna, Italy (e-mail: federica.zonzini@unibo.it).

Vasilis Dertimanis and Eleni Chatzi are with the Department of Civil, Environmental and Geomatic Engineering, ETH Zürich, 8093 Zürich, Switzerland (e-mail: v.derti@ibk.baug.ethz.ch; chatzi@ibk.baug.ethz.ch).

Luca De Marchi is with the Department of Electrical, Electronic and Information Engineering, University of Bologna, 40136 Bologna, Italy (e-mail: l.demarchi@unibo.it).

Digital Object Identifier 10.1109/JIOT.2022.3176671

I. Introduction

STRUCTURAL health monitoring (SHM) systems are becoming ubiquitous in multiple application scenarios due to the increasing demand for safer structures and infrastructures [1], [2].

Significant advancements have been achieved in recent years in terms of promoting instrumentation of structural systems with the purpose of autonomous and continuous diagnostics [3]. Indeed, the advent of cyber–physical systems, which can be considered as one of the most prominent results promoted by the Internet of Things (IoT) paradigm [4], made the real-time and over-time monitoring process a tangible reality, allowing for the continuous sharing of information between users, sensors, and structures [5].

Among the several available SHM techniques, vibration-based monitoring refers to the process of inferring the integrity status of a structural system by continuous and—to the degree feasible—automated monitoring of its dynamics. The diagnostic procedure relies on the extraction of a representative set of vibration-based features over an extended period of time and is classically adopted for structures that operate in the dynamic regime, such as bridges [6] or wind turbines (WTs) [7]. Most typically, such features pertain to frequency-related quantities; the so-called modal parameters [8]. Identification techniques that extract modal information from the measured vibration response, also referred to as output-only methods, fall within the class of operational modal analysis (OMA) solutions [9]. In this case, which is though a requirement in practical scenarios, no controlled stimulus is applied to the structure, which is conversely left to vibrate in its normal operative conditions.

In recent years, the IoT community has placed constantly increasing attention to engineering problems related to the specificity of vibration-based structural monitoring and control. This includes the publication by Burrello *et al.* [10], where the focus is on vibration diagnostics for a railway viaduct. Similarly, an IoT-based model for the real-time condition monitoring of electrical rotors was proposed in [11] together with a novel classification scheme for fault characterization. Effective IoT solutions have found further application for railway bridge tracking, as documented in [12], which concerns the realization of an on-demand sensing system equipped with a dedicated power management unit capable of performing train-induced vibration energy harvesting and control.

Nevertheless, a number of issues remain to be tackled in order to enhance the responsiveness and the resilience of the designed monitoring architectures [13]. Five main and mutually inter-related challenges can be listed: 1) the big data volume implied by dense sensor networks deployed on large-scale infrastructures; 2) the consequent probability of network congestion due to limited and shared communication channels between manifold devices; 3) the growing latency in the data-to-user transfer process, since long time series are cumbersome to manage both in terms of time delivery and processing; 4) device constraints, i.e., limited memory space and computational resources available on sensors in order to meet the requirements of 5) low-power and low-cost hardware that can ensure a long-lasting monitoring modality.

These aspects are even more crucial in case of wireless monitoring systems [14], where the presence of battery-operated devices allows for higher versatility in the deployment process, while posing more stringent limitations on the power consumption. For example, an energy-efficient sensor scheduling strategy for bridge SHM is described in [15], together with a discussion on the advantages granted by pervasive edge computing for decentralized and long-lasting battery-operated monitoring systems. Furthermore, a recently released prototype, Sensifi [16], represents an ultrahigh-rate wireless sensing system originally targeting spacecraft vibration analysis via a custom low-power node measuring vibrations. Aspects, such as power consumption optimization, data compression, time synchronization, and integrated circuit electronics are tackled by the authors and thoroughly investigated as crucial points of SHM-oriented and IoT-driven systems.

As such, recent work has been focused on data compression techniques as a means to jointly address the above-mentioned issues. In this context, it is worth mentioning the works [17]–[20]. Besides the promising results already achieved by the above-referenced works, a consistent deal of research has been spent in the last two decades, seeking to design the most effective compression scheme for vibration monitoring [10], [21]. However, the edge computing perspective, i.e., investigating whether and how these solutions could be practically implemented on self-contained sensor boards, has only recently gained attention [22].

A. Vibration Data Compression at the Extreme Edge

The adoption of conventional compression strategies is often ineffective for vibration data or, in other cases, the compression ratio (CR) these allow for is insufficient. In [10], a list of standard data reduction algorithms is provided, citing their main limitations. The downside of lossless strategies is that a very low CR (usually lower than $3\times$) can be achieved. On the other hand, lossy methods can be employed, ensuring a higher CR but at the cost of more power-hungry implementations. Alternatively, wavelet-based solutions have also been tested, where compression is achieved by thresholding the wavelet coefficients in the different spectral bands and retrieving only those above a certain energetic value. Despite their promising performance in terms of compression level, wavelet-based solutions suffer heavier computational costs, which render their integration in edge devices more challenging. Beside, the prevalent techniques explored for data compression in vibration analysis include spectral-based decomposition [23], compressed sensing (CS) [24], and eigenvalue-based approaches, since they can efficiently take advantage of the sparse and localized nature of features that is unique to vibration signals.

The former refers to the ensemble of methodologies built on the selection of a small batch of parameters out of the spectral representation of the input signals. This procedure can be straightforwardly performed by applying peak-picking algorithms directly on the Fourier-operated data and, then, by extracting the topmost peak spectral values. A seminal work in this field is provided in [25] and further validated with a near-sensor implementation on prototyping boards. However, these methods present a crucial limitation since they assume that the vibration components to be analyzed are well-spaced, highly energetic, and significantly decoupled, a condition which barely holds when dealing with the majority of real SHM scenarios under operative conditions. In this sense, their effect is to "decompose" a multidegree-of-freedom system into the linear summation of single structural components, which can be treated independently.

On the other hand, CS approaches define the problem on a pure mathematical basis by resorting to linear algebra transformations as a means for data reduction [26], [27]. The underpinning principle behind CS is that, under the premise that the processed class of signals is sparse in a given sparsity basis, only a few coefficients suffice to capture the signal content. If this condition applies, a shrunk version of a long time series can be obtained by projecting it onto a lower-dimensional subspace through a suitable compression matrix. Thus, CS performs a lossy compression and its effectiveness is conditional upon the selection of two fundamental ingredients, namely, the optimal compression matrix and the sparsity basis. Once these quantities have been properly defined, a near-sensor implementation of the CS encoder can be achieved by statically loading the compression matrix at the sensor start-up phase and using it to perform compression. Notably, since the signal sparsity may vary at a large extent due to structural and environmental factors, methods capable to adapt these defining features over time should be preferred. However, since this adaptation is difficult to be accomplished on the fly, the compression-accuracy tradeoff is commonly solved by relaxing the CR in favor of a lower noise-to-signal reconstruction error. As a consequence, typical CRs for vibration-based SHM hardly exceed one fifth [22] of the total amount of samples.

Addressing this issue, the scheme of history principal component analysis (HPCA) has very recently been proposed for network load reduction. It exploits the eigenvalue decomposition of the correlation matrix between signal components to extract the primary information to be preserved. Outstanding results were obtained in [10] via adoption of HPCA, where a $10\times$ compression factor was achieved with satisfying reconstruction accuracy, while embedding the algorithm on network end nodes.

As an alternative to standard Fourier-driven or eigenvalue-based algorithms, structural analysis built on parametric approaches relies on the idea that the mechanical and physical laws governing the equations of motion admit an abstract, but still completely equivalent, mathematical representation

as a causal linear time-invariant filter [28]. In line with this formulation, the objective of parametric system identification (SysId) implemented via time series models is to estimate that set of filter coefficients, also known as *model parameters*, which can exactly reproduce the measured input–output system relationship. The power of parametric models is, thus, to instill structure and to encapsulate, in this way, the meaningful portion of the signal content in a reduced set of values (the model parameters), which fully capture the underlying system dynamics. In particular, since the number of parameters typically settles below a couple of dozens [29], massive compression levels could be potentially attained considering the length of the time series to be collected.

It is worth pinpointing that the benefit in pursuing parametric identification strategies is not restricted to the reduction of the data payload to be transmitted, but more importantly extends to the significant enhancement in the quality of the retrieved spectral properties [30]. In this sense, a twofold advantage is brought. First, the spectrum is analytically generated from the computed filter coefficients, as opposed to the conventional approach of applying a Fourier transformation on the raw data, where the influence of noise might be detrimental. Thus, spectra deriving from parametric methods inherently allow for a significant increase in the signal-to-noise ratio (SNR). Second, it follows that the delivered spectral profiles are characterized by a much sharper and smoother trend with respect to nonparametric approaches; a trait which facilitates the subsequent feature extraction phase, particularly when dealing with peak-picking algorithms.

Despite these advantages, discussed in preliminary works in which the problem of SysId edge inference is considered and treated in a purely mathematical manner [31], [32], there is, to the best of the authors' knowledge, only one example of parametric SysId implementation on edge devices available in the literature. A possible explanation for this may be found in the high computational complexity of the involved algorithms, which renders their embodiment in resource-constrained sensors a nontrivial task.

This is the case of the work presented in [19], where Kim and Lynch exploited parametric system modeling, running on the Imote sensor platform, for the structural assessment of civil infrastructures. Nevertheless, despite promising results, the very restrictive memory footprint of this sensor board is not compliant with the execution of the algorithms involved by output-only SysId. To tackle this issue, the authors focused on input–output SysId, adopting simple correlation-based methods, at the expense of reducing communication efficiency, owing to the necessity of broadcasting a reference signal to multiple sensor nodes.

B. Contribution

In this manuscript, the practical embodiment of three different parametric SysId algorithms on a cut-off-the-shelf microcontroller unit (MCU) is proposed, with particular focus on the memory and algorithmic effort that is implied in this undertaking. The proposed implementation does not require transmission of a reference signal among sensor nodes, as opposite to [19]. The main contribution is summarized as follows.

1) On-sensor implementation of both input–output and output-only SysId schemes. From the former category, the autoregressive with the eXogenous input (ARX) model has been selected, whereas the standard autoregressive (AR) model and its smoothed version AR with moving average (ARMA) model have been chosen to deal with OMA-based scenarios. To the best of the authors knowledge, it is the first time that the viability of SysId at the extreme edge is demonstrated.
2) Development of a specific MCU-oriented implementation of the sequential tall-skinny QR factorization strategy as a means to restricting the memory occupancy implied by the execution of standard least-squares regression methods. The designed procedure represents the first edge implementation of this technique and, thus, differs from the original conceptualization tailored to high-performance parallel processors.
3) Analysis of the accuracy in the regressed model parameters when computed by the resource-constrained STM32L5 MCU and quantitative evaluation of the spectral estimation consistency.
4) Presentation of a cost-benefit analysis of the coded algorithms, discussing how time savings from transmission have to be counter-balanced against the increase in procedural burden due to the on-board processing. In this sense, the effect of the processing versus transmission on sensor battery life-cycle is evaluated under consideration of the most common communication protocols for wireless sensor networks.
5) Experimental validation of the coded solutions on the problem of vibration-based assessment of a small-scale WT, undergoing exposure to varying environmental conditions, as well as artificially induced damage.

This article is organized as follows. In Section II, the mathematical formulation of the adopted SysId models is introduced and followed in Section III by the linear algebra procedures which have been employed to fit such models into memory-constrained devices. The experimental validation phase pertaining to the structural assessment of a WT structure is described in Section IV. Results are provided in Section IV-C, first proving the algorithmic validity of the implemented models while porting them on the STM32L5 MCU, and then verifying their applicability for SHM-related tasks. A cost-benefit analysis is presented in Section V, followed by conclusion.

II. Parametric System Identification

SysId based on AR models makes use of regression techniques to identify the sought model parameters, as typically those minimizing the error between the predicted and actually measured system response according to certain heuristics.

In analytical terms, given $x[k]$ and $y[k]$ denoting the generic input–output pair gathered at time stamp kT_s (with T_s indicating the sampling period), a basic and most general variant of a univariate discrete-time parametric model at a generic sample

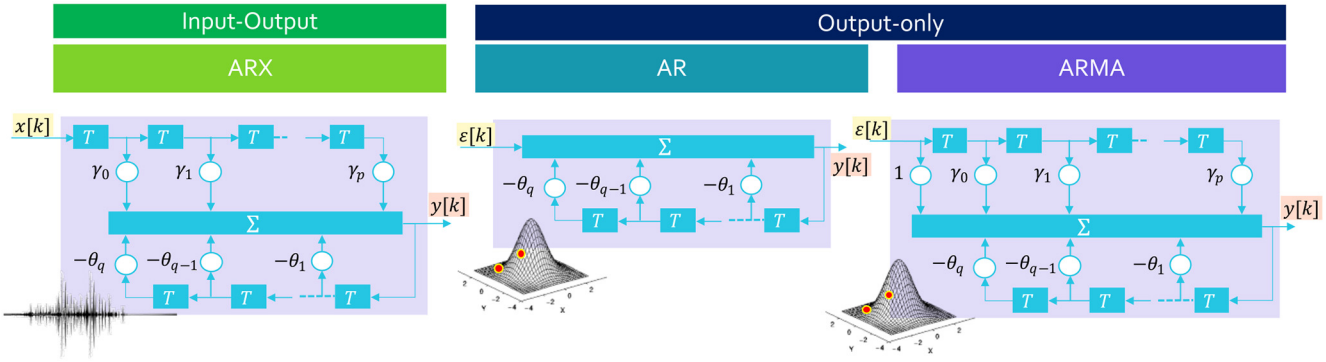


Fig. 1. Formulation of input–output (ARX) and output-only (AR, ARMA) parametric models.

TABLE I
TIME SERIES AND STATE-SPACE REPRESENTATION OF THE ARX, AR, AND ARMA PARAMETRIC MODELS WITH CORRESPONDING REGRESSION TERMS

Model type	Regression form	PSD	Regression terms
ARX	$y[k]+\sum_{i=1}^{q}\theta_i y[k-iT_s]=\sum_{s=0}^{p}\gamma_s x[k-sT_s]$	$S_y(f)=\left\lvert\frac{\sum_{s=0}^{p}\gamma_s e^{-j2\pi fsT_s}}{1+\sum_{i=0}^{q}\theta_i e^{-j2\pi fiT_s}}\right\rvert^2$	$\Phi[k]=\begin{bmatrix}-y[k-1]\\ \vdots\\ -y[k-q]\\ u[k]\\ \vdots\\ u[k-p]\end{bmatrix}\ \beta=\begin{bmatrix}\theta_1\\ \vdots\\ \theta_q\\ \gamma_0\\ \vdots\\ \gamma_p\end{bmatrix}$
AR	$y[k]+\sum_{i=1}^{q}\theta_i y[k-iT_s]=e[k]$	$S_y(f)=\frac{\sigma_e^2}{\left\lvert1+\sum_{i=0}^{q}\theta_i e^{-j2\pi fiT_s}\right\rvert^2}$	$\Phi[k]=\begin{bmatrix}-y[k-1]\\ \vdots\\ -y[k-q]\end{bmatrix}\ \beta=\begin{bmatrix}\theta_1\\ \vdots\\ \theta_q\end{bmatrix}$
ARMA	$y[k]+\sum_{i=1}^{q}\theta_i y[k-iT_s]=e[k]+\sum_{s=0}^{p}\gamma_s e[k-sT_s]$	$S_y(f)=\left\lvert\frac{1+\sum_{s=0}^{p}\gamma_s e^{-j2\pi fsT_s}}{1+\sum_{i=0}^{q}\theta_i e^{-j2\pi fiT_s}}\right\rvert^2$	$\Phi[k]=\begin{bmatrix}-y[k-1]\\ \vdots\\ -y[k-q]\\ e[k]\\ \vdots\\ e[k-p]\end{bmatrix}\ \beta=\begin{bmatrix}\theta_1\\ \vdots\\ \theta_q\\ \gamma_1\\ \vdots\\ \gamma_p\end{bmatrix}$

$k\in\{0,\ldots,N-1\}$ reads

$$y[k]+\sum_{i=1}^{q}\theta_i y[k-iT_s]=\sum_{s=0}^{p}\gamma_s x[k-sT_s] \tag{1}$$

in which q and p specifically determine the number of parameters preserving memory of the past p input and q output instances, while θ_i and γ_s are the feedback and feed-forward taps of the corresponding filter. p and q are also known as the orders of the filter numerator and denominator polynomials, while their summation $N_p=p+q+1$ equals the total amount of model coefficients to be determined.

It is, therefore, from the algebraic manipulation of (1) that all the structural features of interest can be obtained, either in the time or frequency domain, by virtue of the dual relationship between the filter impulse response function (IRF) and its associated frequency response function (FRF)

$$H_y(f)=\frac{\sum_{s=0}^{p}\gamma_s e^{-j2\pi fsT_s}}{1+\sum_{i=0}^{q}\theta_i e^{-j2\pi fiT_s}}. \tag{2}$$

Finally, an estimate of the system's power spectral density (PSD) $S_y(f)$ can be delivered via the square of the magnitude of the FRF as

$$S_y(f)=|H_y[f]|^2 \tag{3}$$

from which the frequency features are retrieved.

A. Parametric Models for Modal Identification

Different identification strategies have been defined depending on the nature of the processed signals and the features of interest. In the following, three schemes will be reviewed, which are classically applied in the context of modal analysis. For the sake of clarity, Table I summarizes the mathematical expressions involved in (1) and (2).

1) Autoregressive Models With eXogenous Input: ARX models are applicable for experimental modal analysis (EMA), i.e., when both the input stimulus and the output response are measured. The block diagram representation of ARX, depicted in Fig. 1, stems from the state-space filter definition and clearly

shows the feedback-feedforward nature of this system model, whose characteristic equations coincide with those offered in (1) and (2).

Despite this being an extremely accurate tool, two main factors limit broad applicability of the ARX scheme. First, the practical difficulty in measuring, with sufficient precision, the input signal (excitation) of the structure under operational conditions, due to unmeasured, arbitrary and/or very weak excitation sources. Second, a decentralized processing requires that the input signal is made available to all the sensing nodes during a preprocessing step, thus increasing the amount of data to be transmitted.

2) Autoregressive Models: Conversely, OMA aims to identify structural properties on the basis of output-only information, under the assumption that the input signal can be described as a white Gaussian noise term $e[k] \sim \mathcal{N}(0, \sigma_e^2)$ with zero-mean and prescribed variance σ_e^2. More specifically, an AR model essentially comprises an all-pole IIR filter obtained by zeroing the contribution of the external input $x[k]$ in (1). The drawback of this model is that a high number of parameters is typically required in order to produce accurate results.

3) Autoregressive Models With Moving Average: Among the output-only methods, ARMA models are superior to basic AR in that they introduce a moving average term in the output equation, which provides smoother and clearly defined spectral curves.

III. Model Parameter Estimation

The expression provided in (1) can straightforwardly be converted into a linear regression formulation, as follows:

$$y[k] = \phi[k]^T \beta + \varepsilon[k] \tag{4}$$

with $\phi[k]^T \in \mathbb{R}^{1\times N_p}$ designating the regression vector and $\beta \in \mathbb{R}^{N_p\times 1}$ denoting the coefficient vector to be estimated. Assuming that the time series spans an observation window of N samples, a full-scale variant of (4) is given as

$$\mathbf{Y} = \Phi\beta + \mathbf{E} \tag{5}$$

where $\Phi = [\phi[1] \dots \phi[N]]^T \in \mathbf{R}^{N\times N_p}$ is a rectangular matrix with regression vectors arranged as horizontal entries, per row; $\mathbf{Y} = [y[1] \dots y[N]]^T \in \mathbf{R}^{N\times 1}$ and $\mathbf{E} = [e[1] \dots e[N]]^T \in \mathbf{R}^{N\times 1}$ correspond, instead, to the observation and error vector. Hence, a final estimate of the sought coefficient vector is yielded via ordinary least squares (OLS), according to

$$\beta = \left(\Phi^T\Phi\right)^{-1}\Phi^T\mathbf{Y} \tag{6}$$

while a recovery of the prediction error is returned as

$$\mathbf{E} = \mathbf{Y} - \Phi\beta \tag{7}$$

with variance $\sigma_e^2 = \mathbf{E}^T\mathbf{E}$. As such, any parametric model is completely characterized by a set of $N_p + 1$ values.

It is worth noting that, rather than resorting in standard linear algebra operations (6), artificial intelligence (AI) methods have been investigated to solve this task [33]. We here capitalize on the straightforward formalization offered by SysId methods. The AI alternatives form a pioneering field of research, which has recently garnered attention [34]: it leverages the algebraic similarities between the structure of conventional time regression methods and the convolution operations at the basis of convolutional neural networks. It is though still not clear whether such networks can be distilled to be ported on resource-constrained devices. Moreover, another point which prevents the full exploitation of AI solutions for the task of the SysId model parameter identification is that the above-discussed mathematical analogy only holds for a limited class of time series models.

4) OLS for ARMA Models: The Hannan–Rissanen Algorithm: The regression technique described above is only applicable for single-stage parametric models, such as ARX and AR, and may not be implemented for evaluation of the ARMA counterpart. In fact, in the latter case, $y[k]$ is regressed not only on its past values, but also on the preceding unobserved quantity $e[k]$, which thus needs to be implicitly calculated. In this case, the Hannan–Rissanen (HR) algorithm [35] provides a simple and yet asymptotically stable solution. HR is based on the cascade of two successive OLS steps: first, a high-order AR model is fitted to the measured response and an estimate of the noise term is derived, as dictated by (7). Knowing E, the next step involves matching a low-order ARX model to the same time series, finally returning an estimate of the ARMA parameters. To be consistent, the order of the first-step AR model should be at least twice the one adopted in the second ARX stage.

A. From OLS to QR Decomposition

The canonical OLS algorithm, which is given in (5), might be prone to numerical instability, rounding effects, and bad conditioning, primarily due to the required inverse matrix operation. To partly alleviate these effects, the QR factorization of the regression matrix is usually suggested as a viable procedure. Indeed, the QR [36] factorization aims at decomposing a full-rank matrix in the product of two independent matrices, namely, an orthogonal matrix Q and an upper triangular matrix R, with the advantage of converting any complex linear system in a simple back-substitution procedure.

For the problem at hand, $\Phi = QR$ can thus be computed and, once plugged into (5), the QR-based variant of OLS (QR-OLS) becomes

$$\beta = R^{-1}Q^T\mathbf{Y}. \tag{8}$$

The dimensions of the two factorizing matrices depend, in turn, on the arrangement of the matrix to be decomposed. In our case, the ratio between the number of rows ($N = N_{s1p}N_p$) and columns (N_p) of the regression matrix exactly amounts to N_{s1p}, i.e., the number of samples per parameter, which is empirically suggested to be a quantity larger than 20 in order to guarantee a sufficiently accurate estimation of the model parameters. Given this, the upper triangular structure of R imposes that only its upper $[N_p \times N_p]$ partition differs from zero. As such, an economy-size variant of the standard QR has to be preferred, returning $Q \in \mathbb{R}^{N\times N_p}$ and $R \in \mathbb{R}^{N_p\times N_p}$.

Several algorithms are available to accomplish QR decomposition. The Householder reflection method [36] is specifically implemented here, granting the most-favorable compromise among the modified Graham–Schmidt orthogonalization, which is readily implementable but extremely prone to numerical errors, and gives rotation, that, conversely, shows great stability but sensitivity to overflow/underflow in single-precision floating-point values [10], [37]. The choice was driven by the necessity to handle very weak and faint signals, sometimes close to the sensor sensitivity, as would be the case for vibration responses that are induced by ambient loads (e.g., traffic and wind).

B. Tall-Skinny QR Decomposition

QR-OLS is efficient in terms of processing, owing to its conceptual and algorithmic simplicity. However, in this form, it appears impractical for near-sensor implementations because of its elevated memory requirements imposed by the large dimensions of the matrices to be processed. It should be noted that the dimension of Φ increases with the square power of the number of parameters, i.e., $\dim\{\Phi(\cdot)\} \propto N_p^2 N_{s1p}$, while the random access memories (RAM), in low-power and low-cost MCUs, are typically below a couple of hundreds of kB, even for the devices with the largest storage capabilities. It follows that, when N_p is in the order of few tens and a minimum number of samples per parameter N_{s1p} is set, the available memory is rapidly consumed. As an example, assuming a single piece of data is represented as a word of 4 B (i.e., 32-bit parallelism), the combination of $N_p = 20$ and $N_{s1p} = 20$ requires at least $N_p^2 N_{s1p} 4 = 32$ kB of memory entirely dedicated to the storage of the regression matrix.

To overcome these restrictions, an MCU version of the classical sequential tall-skinny QR decomposition (S-TSQR) [38] is proposed in this work. S-TSQR was originally designed for parallel architectures (e.g., MapReduce) to provide a communication-avoiding solution for dense linear algebra problems enabling data transfer reduction by means of local grid operations. In this work, we have adapted S-TSQR to single-core embedded computing platforms, in which the computing power and the memory allocation policy of the processor are dramatically lower. Such goal was achieved by exploiting efficient coding techniques, such as loop unrolling, register blocking, buffered multiplications, vector outer product and matrix addition merging, and transposed multiplications enabling fast arithmetic and optimal memory reuse.

In general terms, S-TSQR leverages the key concept of *reproducibility*, i.e., the ability to obtain bit-wise identical results from different runs of the same algorithm given identical input data, regardless of how the computing resources are scheduled. In this sense, the ruling principle at the basis of S-TSQR (schematically depicted in the block diagram of Fig. 2) is to partition the full-scale decomposition of Φ in the subsequent decomposition of small-size $\check{\Phi}_i \in \mathbf{R}^{N_r \times N_p} (i \in \{1, N_c\})$ matrices comprising at most $N_r = N/N_c + N_p$ rows dictated by the selected number of chunks N_c. The procedure is described as follows. Apart from the initial step acting directly on the first N/N_c rows of Φ, in all the remaining $N_c - 1$ iterations QR is performed on

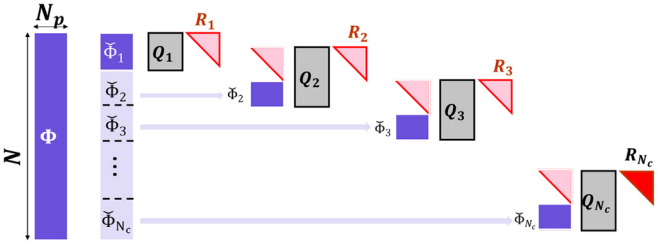


Fig. 2. Processing flow at the basis of the S-TSQR decomposition approach adopted in this work for the sake of matrix dimension reduction.

the newly generated matrix $\check{\Phi}_i = [R_{i-1}|\Phi_i]^T$ obtained from the horizontal concatenation of the previously computed R_{i-1} matrix and the current block rows $\check{\Phi}_i$. Accordingly, the original Q and R terms, referring to the complete regression matrix, can be recovered as $R = R_{N_c}$ and $Q = Q_1 Q_2 \ldots Q_{N_c}$. This means that, while R can be taken directly at the output of the last iteration in a very efficient way, the computation of Q adopted in the canonical S-TSQR procedure [38] is not affordable because it consumes a memory space exactly equal to the original regression matrix to be decomposed, since it implies the storage of all the intermediate Q_i matrices.

To overcome this limitation, a new and memory-efficient procedure was implemented in this work. The proposed solution (which will be referred to as eS-TSQR, i.e., embedded S-TSQR) is inspired by the sparse structure of the Q_i matrices, whose nonnull and nonunitary entries are the *Householder reflectors* α_i [36], i.e., those vectors which are used to perform the orthogonal triangularization of the matrix R. In particular, at the end of each TSQR iteration, an additional step (the coefficients vector update) is introduced, so that the matrix product $Q^T\mathbf{Y} = Q_{N_c}^T \ldots Q_2^T Q_1 \mathbf{Y}$ is substituted by two dot-products $\mathbf{Y_i} = \alpha_i \alpha_i^T \mathbf{Y_{i-1}}$, $(\mathbf{Y_0} = \mathbf{Y})$.

A complete description of the implemented eS-TSQR-OLS procedure is depicted in Fig. 3, where the two main phases, namely, sequential tall-skinny QR decomposition (eS-TSQR) and SysId, are underlined. Note that ARX and AR form direct methods meaning that one single cycle of eS-TSQR-OLS is necessary to obtain the sought model parameters. Conversely, ARMA models imply a recursive two-stage procedure. In this case, the entire procedure needs to be repeated twice: first, the AR modeling procedure is adopted to retrieve the (unknown) noise exciting force, which is then used in a second eS-TSQR-OLS iteration built on the ARX model in order to derive the ARMA parameters.

1) Chunk Size: The optimal number of partitions N_c for the eS-TSQR decomposition is a function of the selected number of samples per parameter. In order for the Householder algorithm to be applicable, it must be ensured that the number of rows in the regression matrix is strictly higher than the number of columns, corresponding to N_p. This condition is always satisfied in the second iteration, due to the fact that the regression matrix constitutes of the horizontal concatenation of R and the previously computed Householder matrix. While, in the first iteration, N_c should be selected such that

$$\frac{N_{s1p} N_p - N_p}{N_c} \geq N_p \tag{9}$$

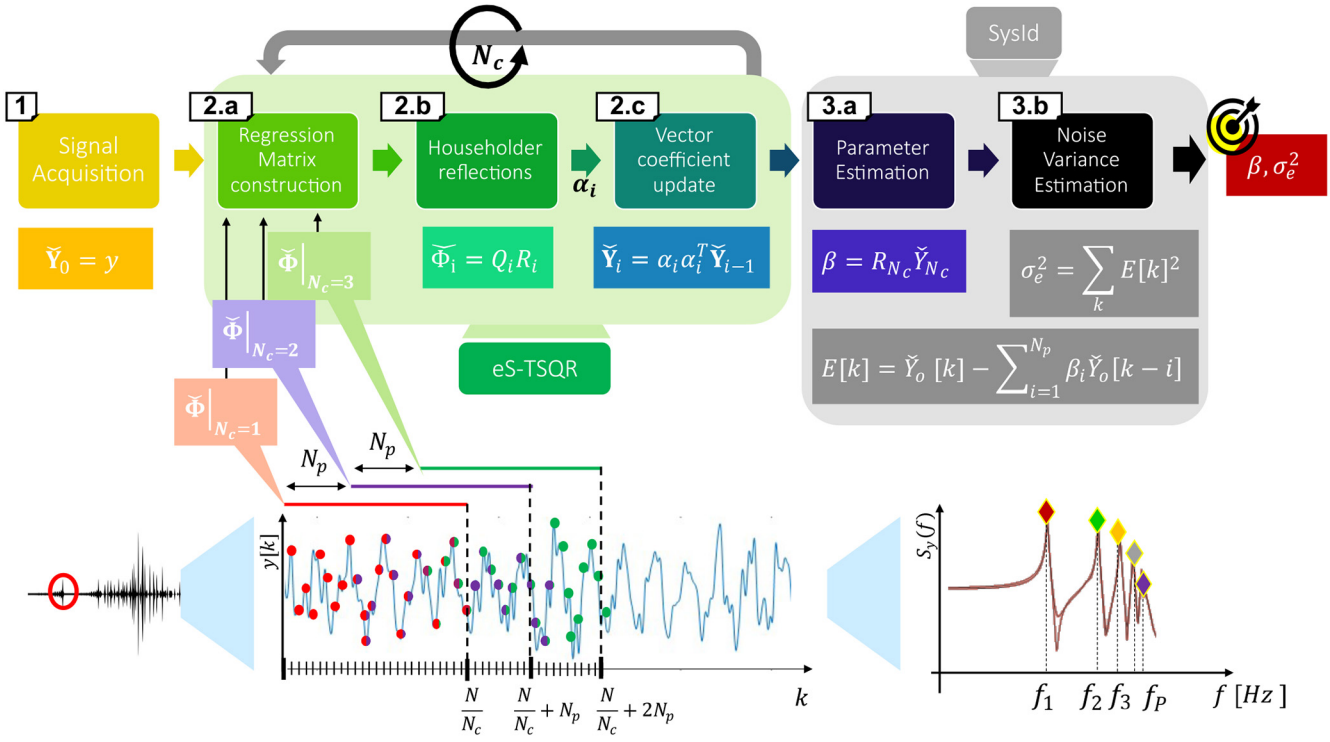


Fig. 3. eS-TSQR-OLS processing flow for model parameter estimation. From left to right: once signals have been acquired (step 1), the eS-TSQR decomposition process is entered and repeated N_c times: at each ith iteration, the regression matrix is first created (as illustrated in Table I) by selecting the proper signal frame (step 2.1), each comprising a window of N/N_c samples which is shifted in steps equal to the number of parameters N_p. Then, the Householder decomposition of $\breve{\Phi}_i$ is applied (step 2.2), paving the way for the subsequent update of the coefficients vector $\breve{\mathbf{Y}}$. At the end of the N_c cycle, the upper triangular matrix R_{N_c} and $\mathbf{Y}_{N_c}$ are used in the SysId phase to compute the model parameter β (step 3.1) and the residual noise density σ_e^2 (step 3.2). The finite set of N_p quantities can, thus, be transmitted at the receiving side, where the spectrum profile $S_y(f)$ of the acquired signal can be reconstructed and the sensor-related modal information are then extracted (e.g., the peak spectral values f_p).

from which it is easy to derive that $N_c^* \leq N_{s1p} - 1$. Hereinafter, $N_c = N_c^*$ will be assumed.

IV. Experimental Validation

A. Materials

The parametric models presented in Section II-A were embedded in the STM32L522ZE-Q Nucleo board, which is one of the latest products released by STMicroelectronics for the prototyping of embedded applications requiring ultralow-power consumption and higher security levels. It integrates, at its core, an STM32L5 MCU [39] based on an ARM Cortex-M33 processor with a single-precision floating-point unit (FPU) and upgraded level of performances thanks to the enhanced DSP functionalities. The equipped memory amounts to 256 kB of RAM and 512 kB of FLASH, which are enough to accommodate both static and volatile data for typical duty cycles of SHM scenarios. Furthermore, it is worth mentioning that this novel family of devices is particularly apt at addressing IoT-related challenges since it achieves excellence in ultralow-power consumption, while ensuring improved security features compared with the preceding L4/L4+ Series (e.g., memory encryption, optimized power management unit, instruction cache supporting both internal and external memories).

B. Methods

1) Model Order Selection: The selection of the proper model order is a critical point for the efficacy of parametric models, since both under or over-estimation may hamper the actual retrieval of the hidden structural information [40]. A plurality of methods has been proposed to tackle this challenge, which are usually based on statistical metrics, such as the Bayesian information criterion (BIC) adopted in this work [41]. Once estimated on a meaningful batch of data, the model order is assumed constant; noteworthy, this is a reasonable choice considering the slow-varying structural properties characterizing the majority of civil and industrial structures [42].

2) Performance Metrics: From the computed set of parameters, modal information can be retrieved by analyzing the associated PSD. As such, the quality of the identified structural properties was assessed by means of the Itakura–Saito spectral divergence (ISD) [43]. ISD represents a cumulative measure of the point-wise spectral distance between two different PSD curves. For N-long frequency vectors, it is defined as

$$\text{ISD} = \frac{1}{N} \sum_{c=1}^{N} \left[\frac{S_y(f)}{\hat{S}_y(f)} - \log\left(\frac{S_y(f)}{\hat{S}_y(f)} \right) - 1 \right]. \tag{10}$$

In our case, $S_y(f)$ and $\hat{S}_y(f)$ are the PSDs computed via SysId as detailed in Section II, by using the model parameters

TABLE II
ISD VALUES (MULTIPLIED BY 10^{-2}) FOR VARYING MODEL ORDER AND NUMBER OF SAMPLES PER PARAMETER

N_{s1p}	N_p 9	17	25	33	41	49	57
25	1.36	1.42	1.40	1.42	1.39	1.40	1.36
30	1.15	1.15	1.16	1.12	1.12	1.17	1.12
35	0.95	0.99	0.95	0.93	0.93	0.93	0.93
40	0.82	0.83	0.82	0.82	0.80	0.79	0
45	0.73	0.69	0.71	0.71	0.71	0	0

computed by the MCU with the eS-TSQR-OLS approach, and via built-in MATLAB functions addressing the same task, respectively. ISD ranges between 0 and 1: spectral superposition is considered perfect in case the ISD equals to zero, whereas higher values highlight possible misalignments.

C. Results

1) Algorithmic Validation: The effectiveness of the implemented extreme-edge processing with respect to off-line computation has been verified in the first phase of the experimental validation. In particular, we focused on the validation of the ARMA model estimation because, given the dual-stage structure of the HR algorithm, the retrieval of ARMA parameters implicitly confirms the validity of both the AR and ARX implementations.

This was accomplished by loading into the STM32L5 FLASH memory one noise-corrupted vibration signal, which was generated via simulation of a six-storey shear frame under white noise base excitation. All the possible combinations of N_p and N_{s1p} values were explored by varying the former quantity in between 9 and 57 (step size equal to 8), whereas the latter was swept in the interval [25; 50] (step size equal to 5). The performance was evaluated in spectral terms via the ISD and the corresponding results are reported in Table II.

As can be observed, the ISD values are below $1.5 \cdot 10^{-2}$ even for the worst-performing configuration, while reaching perfect superposition in some cases (e.g., $N_p = 49$ and $N_{s1p} = 45$).

2) Execution Time: To measure the execution time, the N_p and N_{s1p} pairs discussed in Section IV-C1 were selected, obtaining the processing times depicted in Fig. 4 for the AR/ARX[1] (red scale curves) and ARMA (blue scale curves) case.

The reported trends confirm that the time consumed by the ARMA model is nearly double the time required by the MCU for execution of the AR variant, when a mutual number of samples per parameter and the total amount of parameters is used. This outcome is, once again, consistent with the AR-ARX nature of the adopted HR algorithm. From Fig. 4, it can be seen that the relationship between the processing time and N_p is cubic, whereas the variation due to N_{s1p} is a linear function of the selected number of samples per parameter.

The maximum reported computation time amounts to 129 s and is associated with an ARMA model involving $N_p = 57$ and $N_{s1p} = 45$, i.e., 57 parameters are to be extracted from the time series (2565 samples). For in-field deployment, where

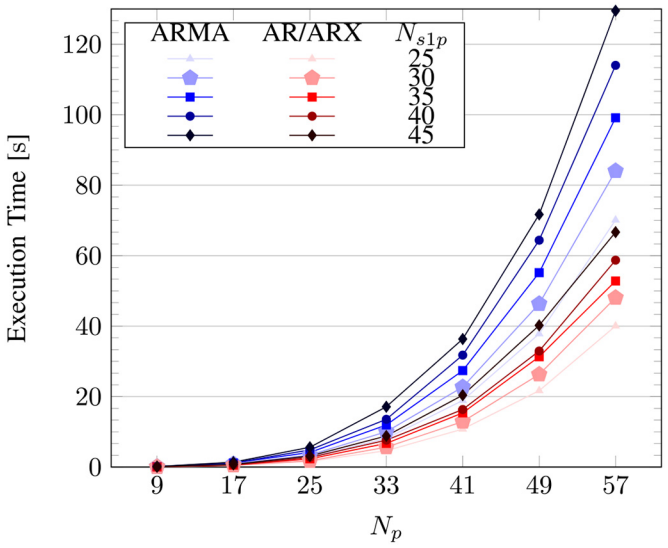


Fig. 4. Execution time for ARMA (blue scale curves) and AR (red scale curves) model running on STM32L5 MCU under different N_p and N_{s1p} configurations.

algorithms are to be executed with low latency, such computation time is barely compatible with high sampling rates. For the sake of an example, acquiring 2565 samples at 50 sps requires about 50 s, which is slightly more than one third of the time taken for processing them.

It is worth noting that, in many practical applications, a viable way to speed up identification is to apply a band-pass filtering operation before running the actual parametric identification task. This reduces the true content of the signal, owing to the focus on some selected spectral bands and in turn, lowers the number of parameters that are necessary to accurately model the system dynamics, which implies a decrease of the computation time according to N_p^3. In addition, as mentioned above, computing such a large number of parameters is hardly required in typical SHM scenarios, where model orders are typically confined below a couple of dozens even for the most complicated vibration patterns, such as the ones characterized by highly coupled modes or very rich profiles [29].

D. Practical Use Case: Wind Turbine Monitoring

The proposed edge solution for data compression in SHM deployments was validated on an actual operating structure. This objective was pursued by exploiting field data collected for a small scale WT hosted in the IBK laboratory at ETH Zürich. More in detail, the considered test-bed consists in a 3.5-kW Windspot blade, manufactured by Sonkyo Energy [44]. This is a small-scale prototype blade element whose structural behavior has been extensively investigated against artificially induced dynamic excitation, as well as varying environmental conditions under both pristine ("healthy") and damaged scenarios.[2] In this work, the vibration response signals, induced by white noise excitation (effective frequency bandwidth between 0 and

[1]Execution time for the ARX model is equal to the one required by AR for the same total number of parameters, for this reason the two single-stage models are provided jointly.

[2]The collected signals have been made publicly available at https://zenodo.org/record/3229743#.YLpz8vkzaUm).

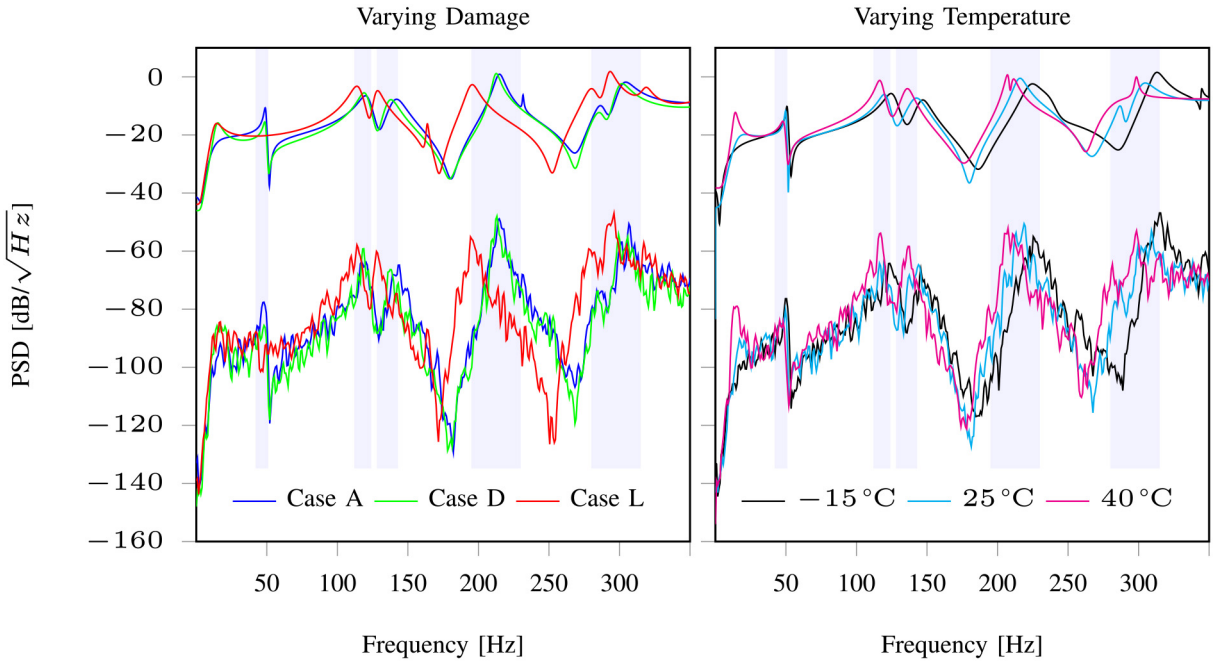


Fig. 5. Spectra of the WT blade working at the reference temperature of +25 °C under progressive damage tests (left), while the effects of temperature changes are underlined in the right-hand image for the damage-free status. Signals at higher density levels, characterized by a smoother profile, are obtained by SysId running on sensor, whereas the ones appearing at the bottom part are computed by standard Welch's method for PSD.

400 Hz) are considered to emulate a practical OMA scenario, which requires use of a broadband (ambient) excitation.

Similarly to the approach adopted in the preliminary validation step, data were statically loaded into the MCU nonvolatile memory at the start-up after conversion to the `float32` bit format. Moreover, since the response signals were acquired via use of commercial instrumentation involving PCB Piezotronics accelerometers that feature high sensitivity and high-resolution levels, which are not compliant with long-term and low-cost monitoring systems, the data sets were corrupted with the white Gaussian noise. This is meant to replicate the intrinsic electronic and mechanical drifts that are common in cut-off-the-shelves digital MEMS devices for low-cost and low-power embedded applications. In particular, the following features were considered: sensor noise equal to 80 mg/Hz^2, a constant offset bias of 40 mg, 16-bit ADC resolution corresponding to 0.061 mg/LSB, zero-g level and sensitivity change versus temperature of ± 0.1 mg/°C and ± 0.01 %/°C, respectively.

In [44], an experimental study has been performed on the considered blade structure, which reveals that the vibration pattern experienced by this structure is remarkably complex, as it is characterized by multiple and closely spaced spectral regions undergoing significant changes due to varying temperature and operational effects. This suggests that simple AR models would be either ineffective in capturing all the significant components with enough resolution or, conversely, too complex to approximate a reasonable solution. Hence, an ARMA model was applied, whose model order—according to the BIC criterion—has been estimated equal to 20 (N_p = 41), for processing time frames of 3000 samples, acquired at a sampling frequency of 833 Hz. The corresponding compression factor amounts to $CR = 3000/41 \approx 75$.

The effectiveness of ARMA models for data compression has been evaluated by verifying whether the spectral signatures, that were reconstructed by the ARMA parameters that were computed by the STM32L5 device under varying conditions, are capable to track the corresponding shifts in the peak spectral values. The rationale behind this choice is that variations in the frequencies that are associated with the most energetic modal components form important indicators of possible damages, or in other words are proxies of anomalies (defects).

Two different analyses were performed and the obtained spectral profiles are shown in Fig. 5. In the left panel of Fig. 5, the capability of the adopted ARMA model to follow the frequency variations induced by man-made damages is investigated. Three reference cases, denoted in the figure with label A, D, and L and characterized by the same temperature value of +25 °C, correspond to three different damaged states simulating, in sequence, the presence of one added mass (case A), the formation of one single crack (case D) and the concurrent occurrence of three cracking phenomena (case L). On another study, three signals for the healthy blade were processed while varying the temperature range between −15 °C, +25 °C, and +40 °C [see Fig. 5, right panel].

In both cases, the perturbation in the spectrum is clearly evident and increases for higher natural frequencies. This is additionally noted via use of gray background boxes whose width increases while moving toward higher frequencies.

Comparing the spectral curves derived from ARMA parameters and the ones computed via the more conventional Welch estimator (lower part of the spectrum), a good agreement is noticeable: indeed, despite a vertical shift due to a bias in the estimated noise density σ_e^2, the peak locations remain clearly centered as well as the global trends superimpose in quite a

precise manner. The main difference between the two spectral estimators is given by the filtering effect of the parametric method, which finally provides a PSD plot that can be more reliably used for extraction of the structural modes in both regimes.

V. Cost-Benefit Analysis

The energy consumption of a sensor node strongly depends on the executed tasks, which may pertain to data acquisition, data processing, and data transmission, with each task contributing to the overall power budget.

In this work, the specific impact to the power budget of edge signal processing for SysId is comprehensively evaluated in conjunction with the one of transmission, by taking into consideration the communication protocols that are best suited for IoT applications. In doing this analysis, the energy spent for sampling can be neglected, as it is proportional to the length of the signal to be acquired. Further to SysId solutions (label SysId), compression-free scenarios (label No DSP), as well as compressive sensing solutions (label CS) are considered, the latter being the main (and most commonly adopted) competitor for data compression in this field.

To accomplish this goal, the IoT analyzer toolbox[3] presented in [45] has been specifically exploited since it provides an open-source platform which allows to simulate the working principles of different IoT-oriented protocols, and to quantify the energy consumption of the corresponding hardware modules. The complete list of protocols and related hardware considered in this work includes: the nrf52840 multiprotocol System on Chip [46] supporting both BLE 5.0 with Long Range connectivity and the 802.15.4 stack, the MAX2830 module [47] enabling 802.11 power-saving mode (PSM), the very recent SX-NEWAH [48] module implementing the communication based on 802.11ah Wi-Fi HaLoW, whereas the SX1272 transceiver [49] was chosen for the LoRaWAN technology. These devices differentiate both in terms of maximum power consumption (from 30 to 700 mW in transmission mode), available data rates (from 125 kbps to 10 Mbps) and maximum packet size (from 120 to 1280 Bytes).

The periodic acquisition of N-sample time series per hour from a tri-axial accelerometer device is simulated to mimic real vibration-based monitoring scenarios. Here, $N = N_{s1p}N_p$ is the total length of the waveforms to be acquired in case of the SysId method. The communication-related energy consumption computed by the analyzer (for a transmission distance of 200 m to be compliant with the communication ranges supported by all the considered protocols) is thus complemented with the one associated to the DSP task. To this end, the execution times reported in Fig. 4 were specifically employed and multiplied by the average power consumption of the STM32L5 device in normal operative mode, which has been experimentally measured equal to 15 mA, while powered at 3.3 V; a CR equal to 5 has been chosen for the CS case. In what follows, among the various tried configurations, results are presented only for the most critical one, corresponding to the ARMA model with $N_{s1p} = 45$: this leads to a gain in the compression factor of $9\times$ and $45\times$ comparing with CS and No DSP, respectively.

[3]https://gricad-gitlab.univ-grenoble-alpes.fr/morinelo/iot-analyzer

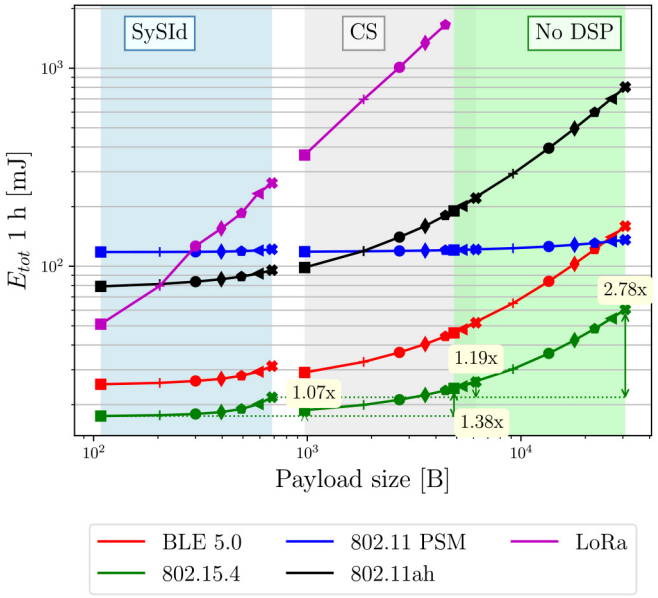


Fig. 6. Total energy expenditure for one hour duty cycle for different IoT protocols, taking into account expenditure due to data processing and outsourcing. The acquisition of triple-channel signals are assumed with $N_{s1p} = 45$, while sweeping N_p in the interval $[9, 57]$ (depicted with different markers). Three different background colors are used to indicate different data compression scenarios: SysId-based processing (blue), CS-based processing (gray), and compression-free (green). Missing points mean that the corresponding payload in the given T× time is not supported by the corresponding protocol.

The trends representing the total energy consumption deriving from the communication and processing operations are shown in Fig. 6. The different background colors are used to identify the three considered working configurations, namely, blue, gray, and green are associated to SysId, CS, and No DSP, respectively. Additionally, the markers are used to indicate the same transmission payload (per given N_{s1p}, while varying N_p in the interval $[9, 57]$, in steps of 8), such as it is easier to compare the considered approaches. The plot indicates that the power saving of SysId w.r.t. CS can reach $10\times$, increasing up to $100\times$ in case of no data compression.

It is further worth mentioning that SysId yields the most efficient performance for all the considered communication protocols. Apart from a horizontal bias due to hardware characteristics, the same consumption curve characterizes BLE 5.0, 802.15.4, and WiFi HaLoW by exhibiting a sharp increase for data payload higher than 1 kB. The trend is slightly different for 802.11 PSM, where the estimated energy profile is almost constant with a minimal increase in case of very large packet sizes. The reason is that this protocol works at a very high data rate (11 Mbps) with a large packet size (1280 B).

The gain in the saved energy for the least power-hungry protocol (i.e., 802.15.4) has been highlighted in Fig. 6 for the two extreme cases of $N_p = 9$ and $N_p = 57$. As can be observed, the energy savings are always favorable, moving from a minimum gain of $1.07\times$ up to a maximum improvement of $1.19\times$ while comparing with CS-driven solutions as dictated by the minimum and maximum number of parameters. Notably, these gains arise to $1.38\times$ (minimum N_p) and $2.78\times$ (maximum N_p) while considering compression-free scenarios.

The SysId-based approach yields a significant advantage w.r.t. the other solutions especially in the LoRa case. As a general observation, the restrictions in terms of sub-band occupancy imposed by ETSI for protocols working in the sub-1 GHz band makes LoRaWAN less effective for this kind of applications. Despite this practical limitation, denoted by the absence of markers in the purple curve for high payload size, the chart shows that, when SysId data compression is leveraged, even this long-range communication technology could become feasible for the assumed transmission rates.

VI. Conclusion

This work investigates the implementation of output-only SysId models at the extreme edge, as a mean to reducing the network congestion in large-scale structural monitoring. The capability to perform structural analysis via output-only methods is crucial since, in practical scenarios, it is often impossible to measure the input stimulus. The pursuit of this goal requires adaptation and customization of SysId algorithms, leveraging the potential of computation at the edge for monitoring solutions. This manuscript presents, in an explicit way, the algorithms and implementation procedures for doing so.

In particular, an MCU version of the eS-TSQR combined with least-squares estimators, termed as eS-TSQR, has been implemented and embedded on a Nucleo board equipped with an STM32L5 MCU. Validation on both synthetic and experimental vibration responses has been performed, proving accurate results in the frequency tracking of structural changes.

The potential power savings due to the network load reduction achieved by running SysId at the extreme edge have been thoroughly evaluated, taking into account the energy expenditure necessary for the model parameter computation. Different wireless transmission protocols that are commonly adopted in the IoT framework have been considered for this purpose. It has been demonstrated that SysId is, even in the most adverse network configurations (i.e., for very long payload sizes), 1.19 - 2.78 times more advantageous with respect to CS-driven and compression-free scenarios, thus ensuring a longer-lasting monitoring system.

Future works will cover the embodiment of the same algorithms in parallel, low-power architectures so as to speed up the execution time. Finally, via the investigated SysId method we aspire to pave the path to extreme-edge inference, by using the estimated parameters as features for the structural health status characterization.

References

[1] C. A. Tokognon, B. Gao, G. Y. Tian, and Y. Yan, "Structural health monitoring framework based on Internet of Things: A survey," *IEEE Internet Things J.*, vol. 4, no. 3, pp. 619–635, Jun. 2017.

[2] D. Misra, G. Das, and D. Das, "An IoT based building health monitoring system supported by cloud," *J. Rel. Intell. Environ.*, vol. 6, no. 3, pp. 141–152, 2020.

[3] K. Worden, E. J. Cross, N. Dervilis, E. Papatheou, and I. Antoniadou, "Structural health monitoring: From structures to systems-of-systems," *IFAC-papersonline*, vol. 48, no. 21, pp. 1–17, 2015.

[4] G. Hackmann, W. Guo, G. Yan, Z. Sun, C. Lu, and S. Dyke, "Cyber-physical codesign of distributed structural health monitoring with wireless sensor networks," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 1, pp. 63–72, Jan. 2014.

[5] G. Mei, N. Xu, J. Qin, B. Wang, and P. Qi, "A survey of Internet of Things (IoT) for geohazard prevention: Applications, technologies, and challenges," *IEEE Internet Things J.*, vol. 7, no. 5, pp. 4371–4386, May 2019.

[6] C. R. Farrar, S. W. Doebling, and D. A. Nix, "Vibration–based structural damage identification," *Philosoph. Trans. Royal Soc. London. A Math. Phys. Eng. Sci.*, vol. 359, no. 1778, pp. 131–149, 2001.

[7] S. Bogoevska, M. Spiridonakos, E. Chatzi, E. Dumova-Jovanoska, and R. Höffer, "A data-driven diagnostic framework for wind turbine structures: A holistic approach," *Sensors*, vol. 17, no. 4, p. 720, 2017.

[8] M. P. Limongelli, E. Chatzi, M. Döhler, G. Lombaert, and E. Reynders, "Towards extraction of vibration-based damage indicators," in *Proc. 8th Eur. Workshop Struct. Health Monitor.*, 2016, pp. 1–10.

[9] E. Reynders and G. D. Roeck, "Continuous vibration monitoring and progressive damage testing on the Z24 bridge," in *Encyclopedia Structural Health Monitoring*. Chichester, U.K.: Wiley, 2009.

[10] A. Burrello, A. Marchioni, D. Brunelli, S. Benatti, M. Mangia, and L. Benini, "Embedded streaming principal components analysis for network load reduction in structural health monitoring," *IEEE Internet Things J.*, vol. 8, no. 6, pp. 4433–4447, Mar. 2021.

[11] D. Ganga and V. Ramachandran, "IoT-based vibration analytics of electrical machines," *IEEE Internet Things J.*, vol. 5, no. 6, pp. 4538–4549, Dec. 2018.

[12] Y. Liu, T. Voigt, N. Wirström, and J. Höglund, "EcoVibe: On-demand sensing for railway bridge structural health monitoring," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 1068–1078, Feb. 2019.

[13] F. Lamonaca, C. Scuro, P. F. Sciammarella, R. S. Olivito, D. Grimaldi, and D. L. Carnì, "A layered IoT-based architecture for a distributed structural health monitoring system," *Acta Imeko*, vol. 8, no. 2, pp. 45–52, 2019.

[14] S. A. Putra, B. R. Trilaksono, M. Riyansyah, D. S. Laila, A. Harsoyo, and A. I. Kistijantoro, "Intelligent sensing in multiagent-based wireless sensor network for bridge condition monitoring system," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5397–5410, Jun. 2019.

[15] L. Yi, X. Deng, L. T. Yang, H. Wu, M. Wang, and Y. Situ, "Reinforcement-learning-enabled partial confident information coverage for IoT-based bridge structural health monitoring," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3108–3119, Mar. 2021.

[16] C.-C. Li, V. K. Ramanna, D. Webber, C. Hunter, T. Hack, and B. Dezfouli, "Sensifi: A wireless sensing system for ultra-high-rate applications," *IEEE Internet Things J.*, vol. 9, no. 3, pp. 2025–2043, Feb. 2022.

[17] X. Dong, D. Zhu, Y. Wang, J. P. Lynch, and R. A. Swartz, "Design and validation of acceleration measurement using the martlet wireless sensing system," in *Proc. Smart Mater. Adapt. Struct. Intell. Syst.*, vol. 46148, 2014, Art. no. V001T05A006.

[18] T. Srisooksai, K. Keamarungsi, P. Lamsrichan, and K. Araki, "Practical data compression in wireless sensor networks: A survey," *J. Netw. Comput. Appl.*, vol. 35, no. 1, pp. 37–59, 2012.

[19] J. Kim and J. P. Lynch, "Autonomous decentralized system identification by Markov parameter estimation using distributed smart wireless sensor networks," *J. Eng. Mech.*, vol. 138, no. 5, pp. 478–490, 2012.

[20] Z. Zou, Y. Bao, H. Li, B. F. Spencer, and J. Ou, "Embedding compressive sensing-based data loss recovery algorithm into wireless smart sensors for structural health monitoring," *IEEE Sensors J.*, vol. 15, no. 2, pp. 797–808, Feb. 2015.

[21] F. Zonzini, M. Zauli, M. Mangia, N. Testoni, and L. De Marchi, "Model-assisted compressed sensing for vibration-based structural health monitoring," *IEEE Trans. Ind. Informat.*, vol. 17, no. 11, pp. 7338–7347, Nov. 2021.

[22] R. Klis and E. N. Chatzi, "Vibration monitoring via spectro-temporal compressive sensing for wireless sensor networks," *Struct. Infrastruct. Eng.*, vol. 13, no. 1, pp. 195–209, 2017.

[23] M. F. Duarte and R. G. Baraniuk, "Spectral compressive sensing," *Appl. Comput. Harmon. Anal.*, vol. 35, no. 1, pp. 111–129, 2013.

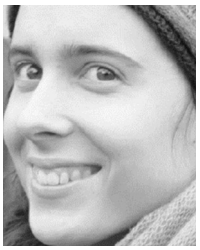
[24] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inf. Theory*, vol. 52, no. 4, pp. 1289–1306, Apr. 2006.

[25] A. T. Zimmerman, M. Shiraishi, R. A. Swartz, and J. P. Lynch, "Automated modal parameter estimation by parallel processing within wireless monitoring systems," *J. Infrastruct. Syst.*, vol. 14, no. 1, pp. 102–113, 2008.

[26] E. J. Candes and Y. Plan, "A probabilistic and RIPless theory of compressed sensing," *IEEE Trans. Inf. Theory*, vol. 57, no. 11, pp. 7235–7254, Nov. 2011.

[27] E. J. Candès, J. Romberg, and T. Tao, "Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information," *IEEE Trans. Inf. Theory*, vol. 52, no. 2, pp. 489–509, Feb. 2006.

[28] S. R. Venkatesh and M. A. Dahleh, "On system identification of complex systems from finite data," *IEEE Trans. Autom. Control*, vol. 46, no. 2, pp. 235–257, Feb. 2001.
[29] F. J. Cara, J. Juan, E. Alarcón, E. Reynders, and G. De Roeck, "Modal contribution and state space order selection in operational modal analysis," *Mech. Syst. Signal Process.*, vol. 38, no. 2, pp. 276–298, 2013.
[30] P. Stoica and R. L. Moses, *Spectral Analysis of Signals*. Hoboken, NJ, USA: Pearson Prentice Hall, 2005.
[31] S. P. Pakala, "Microprocessor implementation of autoregressive analysis of process sensor signals," M.S. thesis, Dept. Grad. School, Univ. Tennessee, Knoxville, TN, USA, 2006.
[32] A. Ghaddar, T. Razafindralambo, I. Simplot-Ryl, D. Simplot-Ryl, and S. Tawbi, "Towards energy-efficient algorithm-based estimation in wireless sensor networks," in *Proc. 6th Int. Conf. Mobile Ad-Hoc Sens. Netw.*, 2010, pp. 39–46.
[33] Q. Tang, J. Zhou, J. Xin, S. Zhao, and Y. Zhou, "Autoregressive model-based structural damage identification and localization using convolutional neural networks," *KSCE J. Civil Eng.*, vol. 24, pp. 2173–2185, Jun. 2020.
[34] M. Ngom and O. Marin, "Fourier neural networks as function approximators and differential equation solvers," *Stat. Anal. Data Min. ASA Data Sci. J.*, vol. 14, no. 6, pp. 647–661, 2021.
[35] T. R. Pillai, S. Palaniappan, A. Abdullah, and H. M. Imran, "Predictive modeling for intrusions in communication systems using GARMA and ARMA models," in *Proc. 5th Nat. Symp. Inf. Technol. Towards New Smart World (NSITNSW)*, 2015, pp. 1–6.
[36] F. Merchant, T. Vatwani, A. Chattopadhyay, S. Raha, S. Nandy, and R. Narayan, "Efficient realization of householder transform through algorithm-architecture co-design for acceleration of QR factorization," *IEEE Trans. Parallel Distrib. Syst.*, vol. 29, no. 8, pp. 1707–1720, Aug. 2018.
[37] L. Ma, K. Dickson, J. McAllister, and J. McCanny, "QR decomposition-based matrix inversion for high performance embedded MIMO receivers," *IEEE Trans. Signal Process.*, vol. 59, no. 4, pp. 1858–1867, Apr. 2011.
[38] J. Demmel, L. Grigori, M. Hoemmen, and J. Langou, "Communication-avoiding parallel and sequential QR factorizations," 2008, *arXiv:0806.2159*.
[39] *STM32L5 Nucleo-144 Board (MB1361), Rev 4*, ST Microelectron., Geneva, Switzerland, Sep. 2020.
[40] I. A. Rezek and S. J. Roberts, "Parametric model order estimation: A brief review," in *IEE Colloquium Model Based Digit. Signal Process. Techn. Anal. Biomed. Signals Dig.*, 1997, pp. 3.1–3.6.
[41] A. Mariani, A. Giorgetti, and M. Chiani, "Model order selection based on information theoretic criteria: Design of the penalty," *IEEE Trans. Signal Process.*, vol. 63, no. 11, pp. 2779–2789, Jun. 2015.
[42] K. Tatsis, V. Dertimanis, Y. Ou, and E. Chatzi, "GP-ARX-based structural damage detection and localization under varying environmental conditions," *J. Sens. Actuat. Netw.*, vol. 9, no. 3, p. 41, 2020.
[43] C. Magnant, E. Grivel, A. Giremus, L. Ratton, and B. Joseph, "Classifying autoregressive models using dissimilarity measures: A comparative study," in *Proc. 23rd Eur. Signal Process. Conf. (EUSIPCO)*, 2015, pp. 998–1002.
[44] Y. Ou, K. E. Tatsis, V. K. Dertimanis, M. D. Spiridonakos, and E. N. Chatzi, "Vibration-based monitoring of a small-scale wind turbine blade under varying climate conditions. Part I: An experimental benchmark," *Struct. Control Health Monitor.*, vol. 28, no. 6, 2021, Art. no. e2660.
[45] E. Morin, M. Maman, R. Guizzetti, and A. Duda, "Comparison of the device lifetime in wireless networks for the Internet of Things," *IEEE Access*, vol. 5, pp. 7097–7114, 2017.
[46] *nRF52840 Product Specification V1.2*, Nordic Semicond., Oslo, Norway, Jan. 2021.
[47] *MAX2830 2.4 GHz to 2.5 GHz 802.11g/b RF Transceiver, PA, and Rx/Tx/Antenna Diversity Switch, Rev 2; 3/11*, Maxim Integr., San Jose, CA, USA, Jan. 2019.
[48] *SX-NEWAH(US), Rev 4*, Silex Technol., Kyoto, Japan, Sep. 2020.
[49] *SX1272/73—860 MHz to 1020 MHz Low Power Long Range Transceiver, Rev 4*, Semtech, Camarillo, CA, USA, 2019.



Federica Zonzini (Graduate Student Member, IEEE) received the B.S. and M.S. degrees in electronic engineering and the Ph.D. degree in structural and environmental health monitoring and management from the University of Bologna, Bologna, Italy, in 2016 and 2018, and 2022, respectively.

Her main research interests include advanced signal processing techniques for structural health monitoring application, encompassing graph signal processing, compressive sensing, and artificial intelligence.

Vasilis Dertimanis received the Ph.D. degree from the National Technical University of Athens (NTUA), Athens, Greece, in the area of modeling and identification of faults in mechanical and structural systems.

He has also participated as a Marie Curie Experienced Researcher to the EU funded SmartEN ITN Project and self-employed as a Freelancer Engineer and an Inspector. Since January 2014, he is a member of the Chair of the Structural Mechanics, ETH Zürich, Zürich, Switzerland, and as of May 2017, he is a Senior Assistant. His research interests lie in the areas of structural identification and health monitoring, linear and nonlinear state estimation, active and passive structural control, hybrid testing, and optimization.

Eleni Chatzi received the Diploma and M.Sc. degrees in civil engineering from the Department of Civil Engineering, National Technical University of Athens (NTUA), Athens, Greece, in 2004 and 2006, respectively, and the Ph.D. degree (with distinction) from the Department of Civil Engineering & Engineering Mechanics, Columbia University, New York, NY, USA, in 2010.

She is currently an Associate Professor and the Chair of Structural Mechanics with the Department of Civil, Environmental and Geomatic Engineering, ETH Zürich, Zürich, Switzerland. Her research spans a broad range of topics, including applications on emerging sensor technologies and structural control, methods for curbing uncertainties in structural diagnostics and life-cycle assessment, as well as advanced schemes for nonlinear/nonstationary dynamics simulations. She led the recently completed ERC Starting Grant WINDMIL on the topic of "Smart Monitoring, Inspection and Life-Cycle Assessment of Wind Turbines."

Prof. Chatzi's work in the domain of self-aware infrastructure was recognized with the 2020 Walter L. Huber Research Prize, awarded by the American Society of Civil Engineers.

Luca De Marchi (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in electronics engineering from the University of Bologna, Bologna, Italy, in 2002 and 2006, respectively.

He is an Associate Professor of Electronics with the Department of Electrical, Electronic, and Information Engineering, University of Bologna. He has published more than 140 papers in international journals or in proceedings of international conferences, and holds two patents. His current research interests are in multiresolution and adaptive signal processing, with a particular emphasis on structural health monitoring applications.

Open Access funding provided by 'Alma Mater Studiorum - Università di Bologna' within the CRUI-CARE Agreement