On the use of Symbolic Regression for Population-Based Modelling of Structures

G. Tsialiamanis[1], N. Dervilis, K. Worden

Dynamics Research Group, Department of Mechanical Engineering, University of Sheffield
Mappin Street, Sheffield S1 3JD

Abstract

Modelling of structures is an important tool for decision making regarding infrastructure. In the absence of sufficient knowledge of the physical phenomena that affect the structure, one can follow a data-driven approach to model its behaviour, relying exclusively on data acquired from it. However, a common problem of this approach is the scarcity of data or biased data. To deal with these two problems, approaches have been considered to transfer knowledge via machine-learning models from one domain to another. The current work considers the case of *population-based structural health monitoring* (PBSHM) of structures. Such an approach is motivated by the common physics that dictate the behaviour of similar structures, which could offer an opportunity to exploit information from a population to create more robust and trustworthy models of data-poor structures of the same population. More specifically, the approach followed here is that of symbolic regression and the transfer is attempted between an extensively-monitored structure and a data-poor structure for a regression application. The methodology is applied in a prognosis problem of crack growth in metal plates and the results reveal the potential of symbolic regression to perform knowledge transfer.

Keywords: Population-based structural health monitoring, machine learning, transfer learning, symbolic regression.

1. Introduction

Structures play an important role in everyday life. They take part in almost every activity and, as a result, maintaining their condition and ensuring their safety is of high importance. To achieve high levels of safety for structures, it is quite important to model them and/or monitor them. Modelling of structure has been an objective for researchers for many years. Traditional

[1] Corresponding Author: George Tsialiamanis (g.tsialiamanis@sheffield.ac.uk)

approaches have followed a *physics-based* path; i.e. the researchers use their physical intuition to define mathematical models which explain the behaviour of materials and structures and can make accurate predictions. For example, the *finite element method* (FEM) [1], is a very common approach to model both existing and under-development structures. The modelling approach is especially useful during the design of the structure, the testing of what-if scenarios, and the optimisation of the structure.

Equivalently to using modelling techniques to establish the safety of under-development structures, monitoring is important to ensure that existing structures are safe. The various tasks performed for monitoring structures and maintaining their safety are part of the field of *structural health monitoring* (SHM) [2]. The functionalities of SHM are summarised by Rytter's hierarchy [3] as,

1. Is there damage in the system (*existence*)?
2. Where is the damage in the system (*location*)?
3. What kind of damage is present (*type/classification*)?
4. How severe is the damage (*extent/severity*)?
5. How much useful (safe) life remains (*prognosis*)?

The above hierarchy offers a convenient classification of the various projects that might be developed under the framework of SHM. It also provides a hierarchy of the difficulty of the above tasks; as one ascends, the tasks become, arguably, more difficult to perform.

The increasing difficulty of the tasks becomes evident by examining each of the levels in the hierarchy, and considering the various ways that such problems can be dealt with. The first step, the acknowledgement of the existence of damage is often translated into a *novelty detection* problem. Approaches to deal with this issue vary, including simple definition of a statistical novelty index and a corresponding threshold, which shall indicate alarm states [4]. For the second and third levels, a distinction between *physics-based* and *data-driven* approaches may be made. Following the physics-based approach, one seeks to exploit one's understanding of damage mechanisms and to define models which shall assist in finding the location of damage or the type of damage. An example is found in [5], where the physics of a crack in a structure are modelled via the XFEM method. For a data-driven approach, one would have to train a model according to data acquired from the structure in order to perform the desired functionality [6].

The last two levels, which refer to damage as an evolving-in-time phenomenon, are often considered the most difficult to deal with. The definition of the severity of damage and its evolution in the future, implies extensive understanding of the damage mechanism in the case of a physics-based approach (in [5] the size of the crack is inferred, so it could be considered a definition of the severity). However, such phenomena are difficult to model, because that requires modelling at different levels of detail (microscopic and macroscopic), and because damage evolution is an inherently random process [2]. Moreover, if a data-driven approach is followed, the models would be called to *extrapolate*, i.e. make predictions outside their training domain, which could affect their performance.

The current work is an attempt to deal with the problem of damage prognosis. As mentioned, data scarcity is quite often an issue of such problems. As a potential solution to lack of data, the current approach follows a *population-based* approach. Common approaches consider multiple structures; however, the current work is an attempt to transfer knowledge from one structure, which has been observed until failure, to another structure, for which the damage evolution process has begun. To perform this transfer, *symbolic regression* [7] is employed, as an attempt at a robust function-transferring approach. The layout of the paper is as follows. In Section 2, a brief introduction is given about the population-based approach to structural health monitoring. In section 3, a description of symbolic regression is provided. In Section 4, a damage-prognosis experimental dataset is presented, as well as the results of applying the transfer of knowledge from one structure to another using symbolic regression. Finally, in Section 5, the results and future steps are discussed.

2. Population-based structural health monitoring

Data-driven approaches are powerful tools when one deals with a problem whose underlying physics are unknown or maybe even impossible to define [8]. However, such approaches are often accompanied by difficulties of different types; the most common example is the scarcity of data. To apply a machine learning algorithm and build an accurate model, one needs data. Different volumes of data are required for different algorithms; for example, for *neural networks* and *deep learning* the need for data is imperative, while for other methods, such as a Gaussian process (GP) [9], inference can be made with a smaller amount of observed data. In any case, data should be available in order to follow the data-driven approach of modelling a phenomenon.

Motivated by *epidemiology* [10], the discipline of *population-based structural health monitoring* (PBSHM) has been recently proposed [11] as an attempt to introduce a population-based philosophy to the data-driven approaches of structural dynamics, including also physics-based approaches, or even combining the two. The approaches developed within the framework aim at performing SHM in *homogeneous* or *heterogeneous* populations. The first category refers to structures that are quite similar, or even nominally identical. Parts of structures which are manufactured similarly fall in this category, e.g. the blades of a helicopter [12]. In this case, an approach which may be followed is that of defining a *form* [11]. A form is a model, preferably a generative model, which defines the way that the structures within the population behave. The form may define the normal-condition characteristics of the structures and may be used to perform novelty detection in the population. The advantage of using such an approach is that there is no need for data from every structure of the population in order to perform inference.

Another approach to population data-driven modelling may be the attempt to transfer a model within a heterogeneous population. The use of *transfer learning* [13] is quite common within the PBSHM domain, because the two fields share a common goal, the transfer of knowledge from data-rich tasks to data-scarce tasks. An application of knowledge transfer within a heterogeneous population is presented in [11], where a damage-classifier model is transferred from one structure to another, achieving quite a high level of prediction accuracy.

An interesting task of PBSHM is that of classifying the structures into categories, with a view to identifying between which structures transfer of knowledge is feasible. In [11, 14], an algorithm for transforming structures into *irreducible elements* (IEs) is presented. The abstract IE models are constructed in order to encapsulate the information about the connectivity of the structures and about the properties of the various structural elements. After transforming the structures into IE models, various classification algorithms can be used to define clusters or communities of structures, within which transfer of knowledge is more likely, because of their similarity.

The aim of the current work is slightly different to most of the aforementioned approaches, where transfer is mainly attempted from an existing population to a newly-considered structure. In the current work, the transfer from one structure to another shall be be considered. Such a transfer is also performed in [11] but for a classification problem. The problem which is studied here is the regression problem of the definition of the remaining useful lifetime.

3. Symbolic regression

The process of learning, and especially machine learning, is quite often focussed on an optimisation procedure. During the optimisation, one often tries to minimise an error metric between the predictions of a model and the actual observations of a system, thus *fitting* the model to explain the system and make predictions about it. This optimisation procedure may be performed in several ways. A common way to perform such a procedure is via *gradient descent* [15]. This approach requires the definition of a *loss function*, which quantifies the error, and of the gradients of this loss function. In many cases, such as *neural networks* [15], the derivatives can be calculated in closed form and the procedure is feasible. However, in many cases, one might try to impose some properties on algorithms which may not be expressed in mathematical terms or may not be differentiable.

A case of such a model, whose loss function may not be differentiable is that of *symbolic regression* [7]. The method of symbolic regression aims at creating interpretable models given some data observations. The interpretability comes from the fact that the models are created by combining various mathematical expressions from a pool of such potential expressions. The motivation for symbolic regression is the equation discovery procedure which one might follow to create a model for a specific phenomenon.

The algorithm of fitting symbolic expressions to data is a *genetic algorithm.* A general framework of genetic algorithms includes a population of individuals, a *fitness function*, a *mutation* process and a *crossover* process. The individuals are often the models which are fitted to the data. The fitness function defines how well or how poorly each individual is fitted to the data, but also how well it satisfies other properties induced by the analyst. The mutation process is defined as random changes in the parameters of each individual of the population. These changes aim at a more thorough exploration of the parameter space close to the current state of the population. Finally, the crossover process combines the parameters of two (or more) individuals to create a new individual. Such a process aims at better exploration of the parameter space further away from the current state of the population.

For the purpose of symbolic regression and equation discovery, the candidate model-functions are encoded as graphs and the procedures of the genetic algorithms are defined as graph operations. The exact procedures are extensively described in [7]. The approach in the aforementioned work includes modifications to the traditional genetic-algorithm framework which can be quite beneficial for the current application. A first modification is that the parameters of the candidate functions are optimised using a gradient descent procedure at every step of the genetic-algorithm optimisation. Performing such optimisation assists in the exploitation of the parameter space, finding the optimal values of the tunable parameters of the candidate functions and reducing the dependency on the mutation process. A second modification is that the evolution algorithm is performed for separate complexities of the model. The complexity is roughly defined as how deep the graph of each function is and optimising at different complexity levels results in a pool of potential solutions of different complexities, which can be considered as potential *forms* [11].

The motivation for using symbolic regression to transfer knowledge from one structure to another comes from its robustness compared to other machine learning algorithms. A candidate model coming from a pool of potential models is a combination of continuous and smooth functions. Moreover, because the model is initially fitted on data coming from one structure, it is expected that the candidate models will enable the transferring of a model with similar and stable behaviour for unseen data of the new structure. An approach like this is also motivated by the way that a human might approach the problem. A potential approach would be to seek to define a closed form function to model all the structures within the population and allow some tunable parameters to fit the model to different structures. Similarly symbolic regression as defined in [7] defines candidate models/forms and allows some parameters to be tuned.

4. Application

To test the performance of symbolic regression in damage prognosis, an experimental setup is considered herein, as shown in Figure 1. The plates are of aluminium and representative of a real helicopter fuselage; they are submitted to cyclic loading with a sinusoidal load of 12Hz frequency and with maximum amplitude 35kN. The damage type is a skin crack, which has been artificially initiated to start the damage evolution process. For more information, the interested reader can refer to [16, 17].

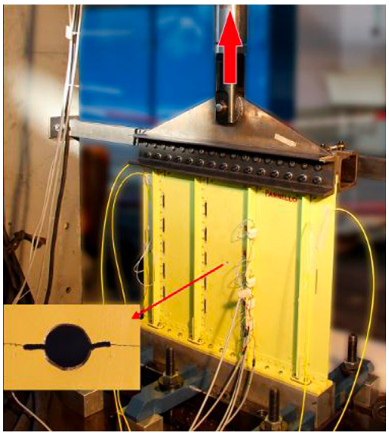

Figure 1: The experimental setup used to acquire the data of the current application [16].

The experiments involved six metal plates and the lengths of the cracks developed in the plates are shown in Figure 2 on the left as a function of the loading cycles needed to achieve the specific crack length. From the aforementioned figure, it becomes clear that, although the metal plates are considered identical, the evolution of the crack is quite different and different values of crack length are achieved for quite different numbers of loading cycles.

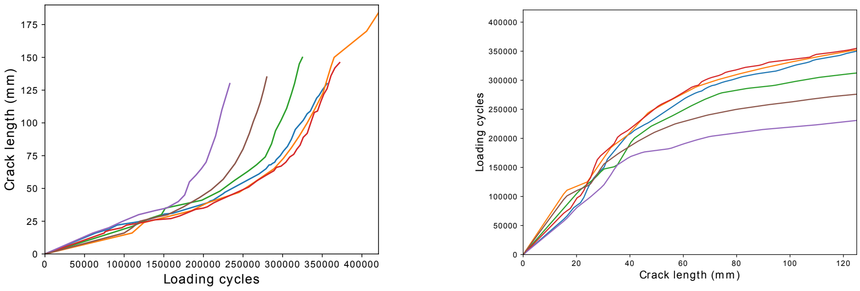

Figure 2: The crack-growth paths of the six plates in their original form (left) and with inverted axes (right), for the purposes of the current work. For reference, the colours of the crack-growth curves corresponding to plate-numbers one to six is considered to be blue, orange, green, red, purple, brown.

In reality, one may define an upper limit of the crack length, values above which could be considered too dangerous to operate the structure and repair would be necessary. In the current work, this limit is considered to be 125mm. Moreover, the relationship which is modelled herein is the number of cycles as a function of the crack length. Inverting the variables in such a way, as in Figure 2 on the right, shall be convenient for the current application. The reason is that the observations are now defined for every value of the input variables, i.e. the crack length, given that all the plates reach the aforementioned limit of 125mm. Moreover, because a closed-form function is fitted to the data, curves in the form of Figure 2 on the left would need a function which would have an asymptote to infinity.

To test the efficiency of the algorithm, one of the crack-growth curves of Figure 2 was considered and the symbolic regression model was fitted to it. As mentioned, the symbolic regression algorithm as developed in [7] seeks for solutions for different levels of complexity, which in the current case is defined as the depth of the equation graph. The result is a set of potential forms of varying equation-graph depth. An example of two solutions is shown in Figure 3. In the specific example, the orange curve equation is $N = 1.0221437\sqrt{x}$ and the green curve equation is $N = 1.9781963594752x - x^2$, where N is the number of cycles and x is the crack length.

The resulting set of forms, is then considered as potential equations that would describe the behaviour of the other plates. To test the assumption, two other plates from the available dataset are considered. The testing plates were the ones corresponding to the green (third plate) and purple curves (fifth plate) in Figure 2. This selection was made in an attempt to transfer knowledge between a plate whose behaviour is slightly different to the training one (the third plate) and a plate whose behaviour is quite different (fifth plate). Considering the parameters of the resulting set of

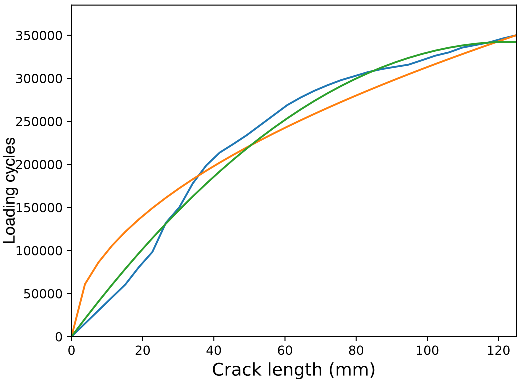

Figure 3: Damage-evolution curve (blue) and two function results of fitting the symbolic regression model to the data (orange and green curves).

equations of the symbolic regression algorithm as tunable, a simple gradient decent algorithm was used to fit the equations to partial observations of the damage-evolution curves. For the example equations of Figure 3, the parametric equations have the form $N = c_1 * \sqrt{x}$ and $N = c_2x - c_3x^2$, where $c_{1,2,3}$ are the tunable parameters. For the current work, two scenarios of available samples from the damage evolution phenomenon were considered. The first scenario comprises observations up to 40% of the total crack growth and the second up to 60% of the total crack growth. In order to get a single prediction, the average value of the fitted equations was considered as the prediction. The results for the two plates for the two scenarios of data availability are shown in Figures 4 and 5.

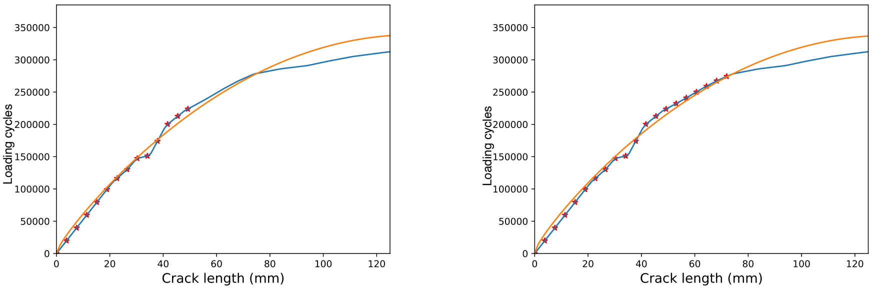

Figure 4: Predictions (orange curves) of fitting the set of equations of the symbolic regression to partial observations (red stars) of the damage path of the third plate (blue curve).

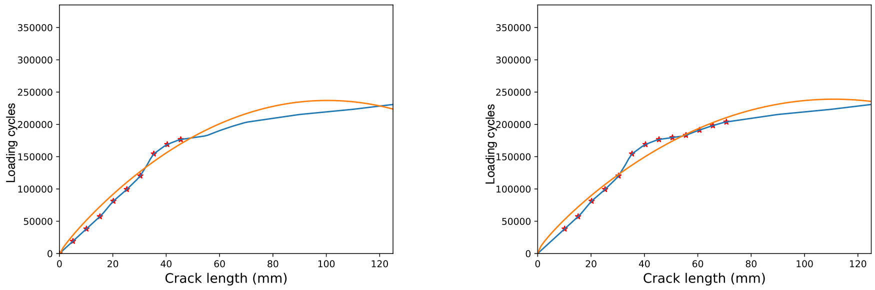

Figure 5: Predictions (orange curves) of fitting the set of equations of the symbolic regression to partial observations (red stars) of the damage path of the fifth plate (blue curve).

It can be seen from the previous figures, the predictions are quite close to the actual curve, especially for the total lifetime of the plates. More specifically, for the third plate, the predictions of the total lifetime are 337495 and 336773 loading cycles having available observations of 40% and 60% of the maximum crack length respectively. The actual total lifetime is 312500 loading cycles, which corresponds to errors of 24995 and 24273 respectively, overestimation in both cases. To provide a normalised version of the errors, their values are divided by the squared standard deviation of the lifetimes of the six plates, which yields errors of 24.45% and 23.06%. The respective errors for the fifth plate are quite lower and equal to 1.99% and 0.87%.

The results reveal that for transferring knowledge from the first to the fifth plate, the method works quite well, while for transferring from the first to the third there is a considerable error. The high error might be observed because of a quite different underlying physical system of the third plate compared to the first. Moreover, it can be observed that in all cases the predictive functions are not monotonically increasing, which contradicts one's understanding of the crack growth mechanism. It is left for future work to modify the loss function of the symbolic regression algorithm in order to yield results which are monotonically increasing.

5. Conclusions and next steps

In the current work, a symbolic-regression approach was presented as an attempt to transfer knowledge between two structures in a regression problem within a PBSHM framework. The approach is motivated by the stability and predictability of functions resulting from a symbolic regression algorithm. Moreover, the application is motivated by the traditional approach to model creating, where a person defines an analytical function to explain the relationship between two or more variables. The method is applied on a damage-prognosis problem, a common regression problem of SHM, and the results are presented.

The algorithm was used as an attempt to transfer knowledge between an aluminium plate, whose crack-growth path is fully available, to two other plates, whose damage-growth paths are only partially observed. In both attempts of transferring knowledge, the plates are considered to belong to the same population, thus they are considered to have similar underlying physics.

The results reveal that, by fitting a symbolic-regression model to the fully-observed damage path of one plate, a set of potential regression functions is acquired. Subsequently, the tunable parameters of these functions are fitted to the partial observations of the crack-growth curves of the testing plates. The results, in terms of the prediction of the total lifetime of the testing plate, were in one case good and in the other case decent. However, it is observed that the resulting functions do not satisfy the analyst's understanding of the physics of the crack-growth problem (in this case the functions are not monotonically increasing). In future work, a modification of the fitness function of the symbolic regression algorithm could be considered to impose such a restriction, aiming at better results. Another modification with a view to better results could be the inclusion of bias terms in the equations. Moreover, the method may benefit if more than one structures are included in the training set, attempting to inform further the algorithm about the common physics of the population.

References

[1] K.-J. Bathe. *Finite Element Procedures.* 2006.

[2] C.R. Farrar and K. Worden. *Structural Health Monitoring: A Machine Learning Perspective.* John Wiley and Sons, 2011.

[3] A. Rytter. *Vibrational Based Inspection of Civil Engineering Structures.* PhD thesis, Aalborg University, 1993.

[4] K. Worden, G. Manson, and D. Allman. Experimental validation of a structural health monitoring methodology: Part I. Novelty detection on a laboratory structure. *Journal of Sound and Vibration*, 259(2):323–343, 2003.

[5] K. Agathos, E. Chatzi, and S.P.A. Bordas. Multiple crack detection in 3D using a stable XFEM and global optimization. *Computational Mechanics*, 62:835–852, 2018.

[6] G. Manson, K. Worden, and D. Allman. Experimental validation of a structural health monitoring methodology: Part III. damage location

on an aircraft wing. *Journal of Sound and Vibration*, 259(2):365–385, 2003.

[7] M Cranmer. Interpretable machine learning for science with PySR and SymbolicRegression. jl. *arXiv preprint arXiv:2305.01582*, 2023.

[8] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021.

[9] C.E. Rasmussen. Gaussian processes in machine learning. In *Summer School on Machine Learning*, pages 63–71. Springer, 2003.

[10] S. Deering, G. Manson, K. Worden, D.W. Allen, C.R. Farrar, and J.S Lombardo. Syndromic surveillance as a paradigm for SHM data fusion. In *4th European Workshop on Structural Health Monitoring*, pages 998–1014, 2008.

[11] P. Gardner, L.A. Bull, J. Gosliga, N. Dervilis, E.J. Cross, E. Papatheou, and K. Worden. Population-based structural health monitoring. *Structural Health Monitoring Based on Data Science Techniques*, pages 413–435, 2022.

[12] T.A. Dardeno, L.A. Bull, R.S. Mills, N. Dervilis, and K. Worden. Modelling variability in vibration-based PBSHM via a generalised population form. *Journal of Sound and Vibration*, 538:117227, 2022.

[13] K. Weiss, T.M. Khoshgoftaar, and D. Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):1–40, 2016.

[14] D.S. Brennan, T. Rogers, E.J. Cross, and K. Worden. On quantifying the similarity of structures via a graph neural network for population-based structural health monitoring. In *Proceedings of ISMA2022 including USD2022*. KU Leuven, 2022.

[15] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, 2006.

[16] M. Corbetta, C. Sbarufatti, A. Manes, and M. Giglio. On dynamic state-space models for fatigue-induced structural degradation. *International Journal of Fatigue*, 61:202–219, 2014.

[17] C. Sbarufatti, A. Manes, and M. Giglio. Performance optimization of a diagnostic system based upon a simulated strain field for fatigue damage characterization. *Mechanical Systems and Signal Processing*, 40(2):667–690, 2013.