This is a repository copy of *Sparse, interpretable and transparent predictive model identification for healthcare data analysis*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/146278/

Version: Accepted Version

Proceedings Paper:
Wei, H. orcid.org/0000-0002-4704-7346 (2019) Sparse, interpretable and transparent predictive model identification for healthcare data analysis. In: Rojas, I., Joya, G. and Catala, A., (eds.) Proceedings of the 2019 International Work-Conference on Artificial Neural Networks (Advances in Computational Intelligence). 2019 International Work-Conference on Artificial Neural Networks (Advances in Computational Intelligence), 12-14 Jun 2019, Gran Canaria, Spain. Lecture Notes in Computer Science, 15 (11506). Springer , pp. 103-114. ISBN 9783030205201

https://doi.org/10.1007/978-3-030-20521-8_9

This is a post-peer-review, pre-copyedit version of an article published in Lecture Notes in Computer Science. The final authenticated version is available online at:
https://doi.org/10.1007/978-3-030-20521-8_9

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Accepted by 2019 International Work-Conference on Artificial Neural Networks (Advances in Computational Intelligence), Gran Canaria, Spain, 12-14 June, 2019

Sparse, Interpretable and Transparent Predictive Model Identification for Healthcare Data Analysis

Hua-Liang Wei[1,2]

[1] Automatic Control and Systems Engineering, University of Sheffield, Sheffield, S1 3JD, UK
[2] INSIGNEO Institute for in Silico Medicine, University of Sheffield, UK
w.hualiang@sheffield.ac.uk

Abstract. Data-driven modelling approaches play an indispensable role in analyzing and understanding complex processes. This study proposes a type of sparse, interpretable and transparent (SIT) machine learning model, which can be used to understand the dependent relationship of a response variable on a set of potential explanatory variables. An ideal candidate for such a SIT representation is the well-known NARMAX (nonlinear autoregressive moving average with exogenous inputs) model, which can be established from measured input and output data of the system of interest, and the final refined model is usually simple, parsimonious and easy to interpret. The performance of the proposed SIT models is evaluated through two real healthcare datasets.

Keywords: System Identification, Data-driven Modelling, Prediction, Healthcare, Machine Learning, NARMAX.

1 Introduction

Data analysis and data modelling are perhaps the most commonly used approaches to acquiring insightful understanding and characterization of complex systems or phenomena where the changes of relevant factors or variables can be quantitatively measured and recorded but the inherent mechanisms or first principle models are not available. Traditionally speaking, data analysis is a technique to achieve insight into data (e.g. organization's data, business data, or whatever other data of interest). Tasks of data analysis range from data management, pre-processing and evaluation to data mining and data modelling. Data modelling provides necessary techniques for understanding and analyze data; one important aspect of data modelling is to understand the relationships between different features and factors of interest through mathematical, statistical and/or other quantitative analysis approaches. System identification and predictive modelling are two important classes of data-driven modelling techniques, the former concerns the development of mathematical models using data observed from dynamical systems, whilst the latter concerns the revealing of relationships of features of interest from any collected data.

The past decades have witnessed tremendous developments and applications of system identification and predictive modelling techniques [1-4], which have been

applied in diverse areas including space weather [6-13], climate and geophysics [14-18], medicine and healthcare [19-21], environments [23-26], societal wellbeing studies [27], and engineering [28-29]. In concept, there are some subtle differences in system identification and predictive modelling. System identification concerns how to find a good model, from measured input and output data of a system of interest, that is as closely as possible to represent the input-output behavior. In doing so, it requires that the identified model should be as accurate as possible to characterize the underlying dynamics hidden in the data. Predictive modelling concerns the detection of dependence relationships among a group of variables by analyzing and modelling the relevant data; the goal is to determine if the change in some variables would affect the other variables, or if the attribute of some specific variables of interest (e.g. response, dependent or output variables) can be characterized by other variables (commonly known as explanatory, predictor, independent or input variables). Classification problem solving with either parametric or non-parametric data modelling methods is a typical example of the application of predictive modelling.

In more detail, system identification [1-4] is different from the conventional concept of predictive modelling [5] in that the former pays more attention on system dynamics. In system identification, it is usually assumed that the measurements or recorded data come from dynamic systems, whose current behaviour (often referred to as the system output) depends on previous or historical states of both the inputs (stimuli or driving signals) and the output itself. In system identification or dynamic modelling, all the input and output signals should strictly be recorded chronologically. Unlike in predictive modelling where the order of data records can be altered and normally the change will not affect the overall modelling performance, in system identification altering data record order is not allowed as the data records virtually reflect the change of the system behaviour with 'time' (this is an implicit independent variable in all time-invariant dynamic systems).

Despite the difference between system identification and predictive modelling, they share many similarities in model construction and algorithm implementations. For example, the commonly used methods of generalized linear models in predictive modelling, including model variable/term selection, model structure detection, model validation and so on, can be borrowed to deal with nonlinear dynamical models e.g. NARMAX (nonlinear autoregressive moving average with exogenous input) model with an appropriate modification, and vice versa. As highlighted in [4], NARMAX model can be considered a dynamically driven single hidden-layer recurrent neural network, which include many neural network structures e.g. radial basis function neural networks (RBFNs) as a special case.

While NARMAX model has been extensively applied in many interdisciplinary fields, its application potential in healthcare and related area has not yet been well explored. So, this study aims to introduce a type of sparse, interpretable and transparent (SIT) model for healthcare and related data modelling problems. We propose to use the NARMAX model, which possesses a number of attractive 'smart' properties, namely, simple and simulatable, meaningful, accountable, reproducible, and transparent. Two examples are provided to show the performance of the proposed approach.

2 Model Representation

A wide range of dynamic systems or processes can be represented using NARMAX model [4]. Taking the case of multiple inputs (designated by u_1, u_2, …, u_r) and one output (designated by y) problem as an example, the NARMAX model that links the output y to the inputs u_1, u_2, …, u_r is of the form:

$$\begin{aligned} y(t) = f[&y(t-1),\cdots,y(t-n_y),u_1(t-1),\cdots,u_1(t-n_u),u_2(t-1),...,u_2(t-n_u),..., \\ &u_r(t-1),...,u_r(t-n_u),e(t-1),\cdots,e(t-n_e)]+e(t) \end{aligned} \tag{1}$$

where y(t), u(t) and e(t) are the measured system output, input and noise sequences respectively at time instant t; n_y, n_u, and n_e are the maximum lags for the system output, input and noise; f[•] is some non-linear function to be estimated from data. Note that the noise e(t) is unmeasurable but can be replaced by the model prediction error in system identification procedure. The noise terms are included to accommodate the effects of measurement noise, modelling errors, and/or unmeasured disturbances.

Now define a group of new variables (i.e., lagged versions of the original input and output variables) as

$$x_m(t) = \begin{cases} y(t-m), & 1 \le m \le n_y \\ u(t-m+n_y), & n_y+1 \le m \le n_y+n_u \\ e(t-m+n_y+n_u), & n_y+n_u+1 \le m \le n \end{cases} \tag{2}$$

where $n = n_y + n_u + n_e$. Model (1) can then be written as

$$y(t) = f[x_1(t), x_2(t), \cdots, x_n(t)] + e(t) \tag{3}$$

In practice, many types of functions are available to approximate the unknown function f[•] in (1), including power-form polynomial models and rational models [28], radial basis function (RBF) [8, 31, 32], and wavelet expansions [33]. In this study, power-form polynomial basis is considered. Expanding model (1) by defining the function f[•] to be a polynomial of degree ℓ gives the representation:

$$\begin{aligned} y(t) = \theta_0 &+ \sum_{i_i=1}^{n} \theta_{i_1} x_{i_1}(t) + \sum_{i_1=1}^{n}\sum_{i_2=i_1}^{n} \theta_{i_1 i_2} x_{i_1}(t) x_{i_2}(t) + \cdots \\ &+ \sum_{i_1=1}^{n} \cdots \sum_{i_\ell=i_{\ell-1}}^{n} \theta_{i_1 i_2 \cdots i_l} x_{i_1}(t) x_{i_2}(t) \cdots x_{i_\ell}(t) + e(t) \end{aligned} \tag{4}$$

where $\theta_{i_1 i_2 \cdots i_m}$ are parameters. The degree of a multivariate polynomial is defined as the highest order among the terms. For example, the degree of the polynomial $h(x_1, x_2) = a_1 x_1 + a_2 x_1 x_2 + a_3 x_1^2 + a_4 x_1 x_2^2$ is $\ell = 1+2=3$, which is determined by the last term, $a_4 x_1 x_2^2$. Similarly, a polynomial model with degree ℓ means that the order of each term in the model is not higher than ℓ. Note that the polynomial representa-

tion (4) belongs to the family of linear-in-the-parameters (LIP) but nonlinear-in-the-variables (NIV) models.

In many applications, the noise signal e(t) in the NARMAX model (1) can be reasonably assumed to be an i.i.d. or white noise. In this case, model (1) can be reduced to a NARX model which only involves lagged input and output variables as below:

$$x_m(t) = \begin{cases} y(t-m), & 1 \le m \le n_y \\ u(t-m+n_y), & n_y+1 \le m \le n = n_y + n_u \end{cases} \tag{5}$$

With the above definition, (1) can easily be re-arranged to a LIP-NIV form.

Note that in nonlinear dynamical system identification, attention is always focused on building mathematical models, from experimental data, that can represent the inherent dynamics or system input-output relationship as accurate as possible. From a neural network perspective, a NARX model can be considered a dynamically driven 1-hidden-layer neural network, which is referred to as recurrent NARX network (R-NARX-NN) [4]. For example, for a NARX model of nonlinear degree ℓ =3, define

$$F_1[\mathbf{x}(t);\theta] = \sum_{i=1}^{n} \theta_i x_i(t)$$

$$F_2[\mathbf{x}(t);\theta] = \sum_{i=1}^{n}\sum_{j=i}^{n} \theta_{ij} x_i(t) x_j(t)$$

$$F_3[\mathbf{x}(t);\theta] = \sum_{i=1}^{n}\sum_{j=i}^{n}\sum_{k=j}^{n} \theta_{ijk} x_i(t) x_j(t) x_k(t)$$

The recurrent neural network structure of the NARX model is shown in Fig. 1.

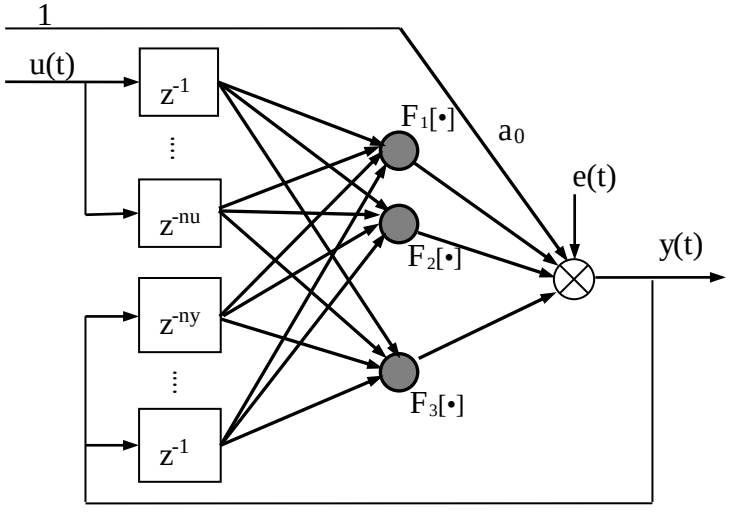

Fig. 1. The structure of the NARX model, which is a typical recurrent neural network.

3 Sparse Dictionary Learning and NARMAX Model Estimation

For convenience of description, we first focus on the NARX model estimation, for which the procedure starts with setting up a few parameters, namely, the maximum lags n_y and n_u, and the nonlinearity degree ℓ. Now, take a simple case as an example, where a system only has one input and one output signal, and assume n_y =1, n_u =1 and ℓ =3. We define the following distributed lag sub-dictionaries:

$$
\begin{aligned}
&D_0 = \{1\} \text{ (constant term)} \\
&D_1 = \{y(t-1), u(t-1)\}, \\
&D_2 = \begin{Bmatrix} y(t-1)\,y(t-1) \\ y(t-1)u(t-1) \\ u(t-1)u(t-1) \end{Bmatrix}, \\
&D_3 = \begin{Bmatrix} y(t-1)\,y(t-1)\,y(t-1) \\ y(t-1)\,y(t-1)u(t-1) \\ y(t-1)u(t-1)u(t-1) \\ u(t-1)u(t-1)u(t-1) \end{Bmatrix}
\end{aligned}
$$

The four sub-dictionaries will then be used to form a dictionary:

$$D = D_0 + D_1 + D_2 + D_3 \tag{6}$$

The task of finding a good model is equivalent to selecting important model terms from the dictionary D, which well represents the input-output relation of the system.

For complex cases (e.g. with many inputs, and with large time lags), we can define D_0, D_1, D_2, D_3, etc. in the same way. Note that the total number of potential model terms in a polynomial NARX model is $M = (n+\ell)!/[n!\ell!]$. For example, if ℓ =3, n_y = 10, n_u = 5, then M = (15+3)!/(15!3!) = 153. For large n_y and n_u, the number of initial candidate model terms included in the initial full model can be very large. However, for a given system, the M candidate model terms in D are not necessarily equally important for representing the system. Some terms may be irrelevant or only make very tiny contribution to explaining the system input-output behavior, thus should not be included in the model, because an inclusion of irrelevant model terms can generally lead to model overfitting, and may adversely make it more difficulty to reveal the true system dynamics. The forward regression orthogonal least squares (FROLS) algorithm [4, 31, 32] and its variants [34-38] provide an efficient, powerful tool for nonlinear significant model term selection and model structure detection. Detailed discussions on the FROLS algorithm can be found in [4, 39, 40].

This study uses the FROLS algorithm with ridge regularization to select significant model terms and determine the model structure. Once a NARX model structure is determined, the noise variables e(t-1), …, e(t- n_e) and the model terms involving these noise variables are accommodated to the NARX model, to develop a NARMAX

model structure. Note that the noise signal $e(t)$ is not observable but can only be estimated from the prediction errors: $\xi(t) = y(t) - \hat{y}(t)$, where $\hat{y}(t)$ is the model prediction at time instant t. Detailed discussions may be found in [4].

4 Case Studies and Real Applications

4.1 The Relation Between Influenza-like Illness Incidence Rate and Deaths

The weekly influenza-like illness (ILI) incidence rate and deaths data were acquired from the Office for National Statistics (ONS), The Royal College of General Practitioners Research and Surveillance Centre and Public Health Wales. The dataset contains a total of 991 weekly records starting in week 31 of 1999 and ending in week 30 of 2018. The raw data are plotted in Fig. 2.

The objective here is twofold. One is to quantify the relation between the week mortality and the ILI incidence rate, and another is to do a week ahead prediction of the death mortality. We consider two types of models: one using autoregressive variables and another one without using autoregressive variables. For both cases, the 991 data points are split into two parts: the first 600 samples are used for model training and the remaining 391 are used for model testing.

The Model Without Including Autoregressive Variables

Voterra model is special case of NARMAX model, without including autoregressive variables. The best Volterra model identified by the FROLS algorithm with Ridge regularization is:

$$\left.\begin{aligned} y(t) = {} & 8636.0572 + 64.6550u(t-1) + 55.6953u(t-4) \\ & - 0.5110u^2(t-1) + 0.0015u^2(t-1) - 0.8304u^2(t-4) - 0.0026u^3(t-4) \end{aligned}\right\} \quad (7)$$

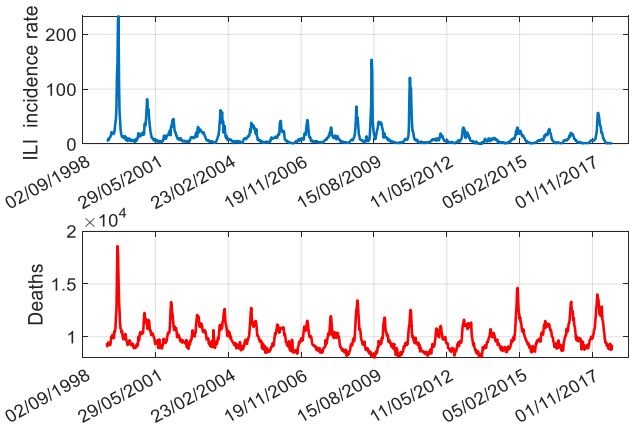

Fig. 2. Weekly influenza-like illness (ILI) incidence rate and deaths, England and Wales, between week 31 of 1999 and week 30 of 2018.

where u(t) represents the weekly ILI incidence rate and y(t) represents the number of weekly deaths. A comparison of the model predicted deaths and the corresponding true values, on the training and test data sets, are shown in Fig. 3 and Fig. 4, respectively. It can be seen that the simple NARMAX model (7) shows an overall very good prediction performance. Note that the model predictions of the short period centred on 15 August 2009 are quite bad, this is because there is some extremely odd behaviour in then ILI incidence rate as shown in Fig. 2.

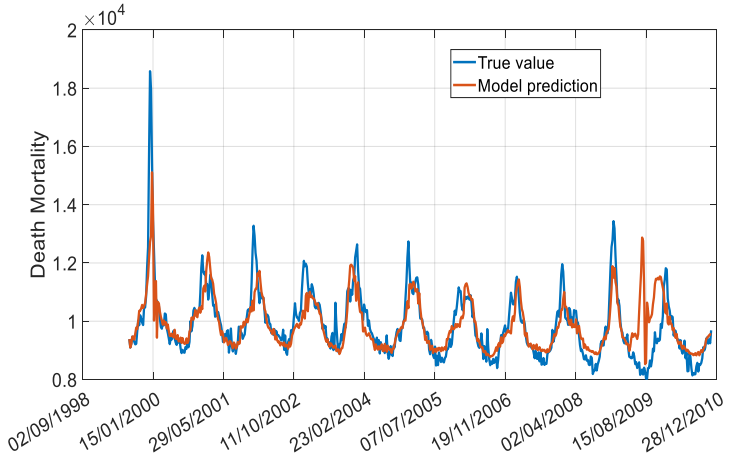

Fig. 3. A comparison of the model prediction with the corresponding true number of deaths, on the training dataset of the period between week 31 of 1993 and week 47 of 2010.

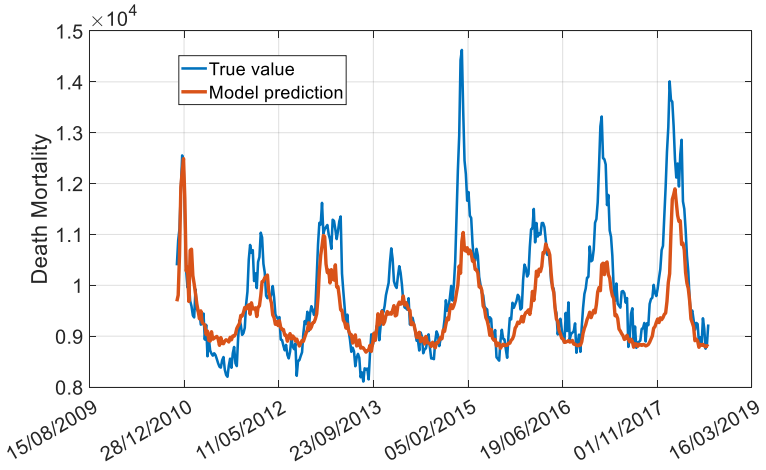

Fig. 4. A comparison of the model prediction with the corresponding true number of deaths, on the test dataset of the period between week 48 of 2010 and week 30 of 2018.

The Model with Autoregressive Variables

With the same training data, the FROLS algorithm produces the following NARMAX model:

$$y(t) = 616.435147 + 0.927840\,y(t-1) - 0.114871u(t-1)u(t-3) + 10.535455u(t-1) \quad (8)$$

Note that all the model terms involving noise variables such as u(t-1)e(t-1) are omitted and not included in the final model, because all these noise terms are not useful for model prediction but are only used to reduce bias in model estimation. A comparison of the model predicted deaths and the corresponding true values, on the training and test data sets, are shown in Fig. 5 and Fig. 6, respectively. Model (8) shows that the death mortality is closely correlated to the ILI incidence rate.

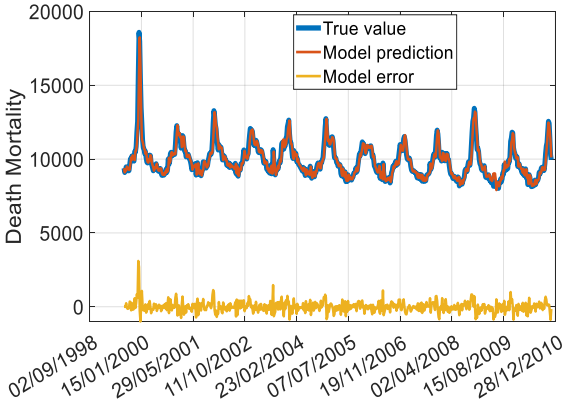

Fig. 5. A comparison of the model prediction with the corresponding true number of deaths, on the training dataset of the period between week 31 of 1993 and week 47 of 2010.

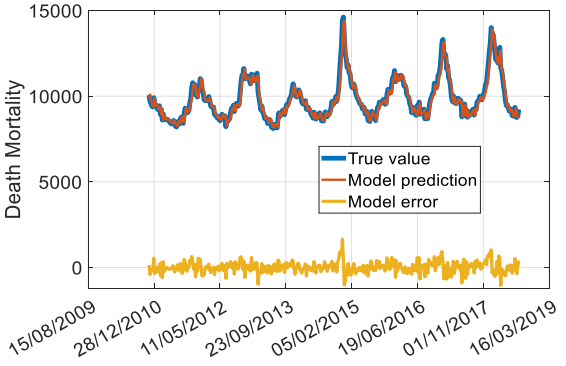

Fig. 6. A comparison of the model prediction with the corresponding true number of deaths, on the test dataset of the period between week 48 of 2010 and week 30 of 2018.

4.2 Analysis of Beijing Air Quality

A dataset of Beijing air quality is obtained from http://www.tianqihoubao.com/aqi/. The dataset contains six variables, namely, PM2.5 (μg/m^3), PM10 (μg/m^3), SO2 (μg/m^3), NO2 (μg/m^3), CO (mg/m^3), O3 (μg/m^3), all of which were measured daily. Here, in this study we are interested in understanding how PM2.5 depends on or is delated to the other five variables. We therefore treat PM2.5 as an output variable and the other five variables are treated to be the inputs.

We use 732 sample data of the period from 1 January 2016 to 31 December 2017 to train the model, and use the data of the period of 1 January 2018-31 January 2019 to test the model performance. The identified model is:

$$\left.\begin{aligned}
PM_{2.5}(t) = {} & 0.43615PM_{10}(t) + 38.07453CO(t) - 17.90321 \\
& + 0.31627CO(t)O_3(t) - 0.00119[O_3(t)]^2 - 7.48672[CO(t)]^2 \\
& + 0.11175PM_{10}(t)CO(t) + 0.00033[PM_{10}(t)]^2 \\
& + 0.01899NO_2(t)NO_2(t-1) - 0.007294PM_{10}(t)NO_2(t-1) \\
& + 0.06749CO(t-1)O_3(t-1) + 0.00100SO_2(t)O_3(t-2) \\
& + 0.01151PM_{10}(t)SO_2(t-1) - 0.02525SO_2(t-1)NO_2(t) \\
& - 0.03123PM_{10}(t)CO(t-2) - 0.39556NO_2(t) \\
& + 0.26760SO_2(t-2)CO(t) - 0.72491SO_2(t)
\end{aligned}\right\} \quad (9)$$

The values of RMSE (root mean squared error), MAE (mean absolute error), Correlation (between the measurement and model prediction), and R^2 (coefficient of determination) of model (9) over the training data are 15.4160, 10.8202, 0.9656, and 0.9924, respectively, and 16.9764, 11.2995, 0.9193 and 0.8450, respectively, over the test data. These statistics show that PM2.5 has a very strong relation or dependence with other five variables. A comparison between the model predictions and the corresponding true measured values over test datasets is shown in Fig. 7.

5 Conclusion

Dara-driven modelling and data based quantitative analysis play a key instrumental role in knowledge discovery from healthcare data. In many application scenarios, it is interested in knowing how a variable is explicitly related to other variables or factors. To answer this question, this study proposed a type of SIT (sparse, interpretable and transparent) approach, called NARMAX model, which possess 'smart' properties: simple/sparse/simulatable, meaningful, accountable, reproducible, and transparent. SIT-NARMAX model can be written as a LIP-NIV (linear-in-the-parameters and nonlinear-in-the-variables) form, which can easily be estimated using the state-of-the-art linear regression methods. By applying the forward regression orthogonal least squares (FROLS) algorithms to this type of model, it usually leads to parsimonious representations for most real data modelling problems. The main advantage of the proposed modelling approach is that the resulting model is not only parsimonious but also can be written down and easily interpreted. As illustrated in the case studies, the

proposed approach provides a powerful and effective tool for dealing with real healthcare and related data modelling problems.

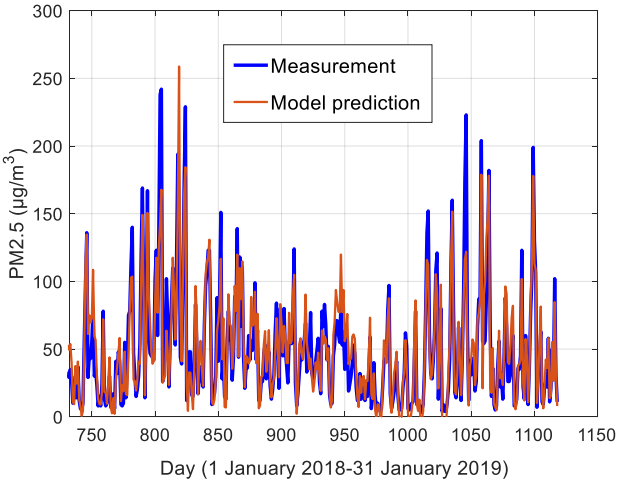

Fig. 7. A comparison between the model predicted values and the corresponding measurements of Beijing air quality (PM2.5) over the test dataset (RMSE = 16.9764, MAE =11.2995, r = 0.9193 and R^2 = 0.8450).

Acknowledgments

This work was supported in part by the Engineering and Physical Sciences Research Council (EPSRC) under Grant EP/I011056/1, the Platform Grant EP/H00453X/1, and the EU Horizon 2020 Research and Innovation Programme Action Framework under grant agreement 637302.

References

1. Ljung, L.: System Identification: Theory for the User, Prentice-Hall: Upper Saddle River, N.J. (1987).
2. Sodestrom, T., Stoica, P.: System Identification, Prentice Hall: Upper Saddle River, N.J. (1988).
3. Nelles, O.: Nonlinear System Identification. Heidelberg, Berlin: Springer-Verlag (2011).
4. Billings, S.A.: Non-linear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains, Wiley: London (2013).
5. Kuhn, M., Johnson, K: Applied Predictive Modeling, Springer (2013).
6. Wei, H.L., Billings, S.A., Sharma, A. S., Wing, S., Boynton, R. J., Walker, S.N.: Forecasting relativistic electron flux using dynamic multiple regression models. Annales Geophysicae 29(2), 415-420 (2011).

7. Wei, H.L., Billings, S.A., Balikhin, M.: Prediction of the Dst index using multiresolution wavelet models. Geophysical Research 109(A7), A07212 (2004).
8. Wei, H.L., Zhu, D.Q., Billings, S.A., Balikhin, M.A.: Forecasting the geomagnetic activity of the Dst index using multiscale radial basis function networks. Advances in Space Research 40(12), 1863-1870 (2007).
9. Balikhin, M.A., Boynton, R.J., Walker, S.N., et al.: Using the NARMAX approach to model the evolution of energetic electrons fluxes at geostationary orbit. Geophysical Research Letters 38(18), L18105 (2011).
10. Boynton, R.J., Balikhin, M.A., Billings, S. A.: Using the NARMAX OLS-ERR algorithm to obtain the most influential coupling functions that affect the evolution of the magnetosphere. Geophysical Research –Space Physics 116, A05218 (2011).
11. Gu, Y., Wei, H.L., Boynton, R.J., Walker, S.N., Balikhin, M.A.: System identification and data-driven forecasting of AE index and prediction uncertainty analysis using a new cloud-NARX model. Journal of Geophysical Research: Space Physics 124(1), 248-263 (2019).
12. Boynton, R., Balikhin, M., Wei, H.-L., Lang, Z.-Q: Applications of NARMAX in space weather. In Machine Learning Techniques for Space Weather (2018), 203–236.
13. Camporeale, E.: The challenge of machine learning in space weather nowcasting and forecasting. arXiv preprint arXiv:1903.05192 (2019).
14. Wei, H.-L., Billings, S.A.: An efficient nonlinear cardinal B-spline model for high tide forecasts at the Venice Lagoon. Nonlinear Processes in Geophysics 13(5), 577-584 (2006).
15. Karsten, S., Nitesh V.C., Auroop R.G.: Complex networks as a unified framework for descriptive analysis and predictive modeling in climate science. Statistical Analysis and Data Mining 4(5), 497-511 (2011).
16. Bigg, G.R., Wei, H.L., Wilton, et al.: A century of variation in the dependence of Greenland iceberg calving on ice sheet surface mass balance and regional climate change. Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences 470(2166), 2013066 (2014).
17. Pearson, R.G., Dawson, T.P.: Predicting the impacts of climate change on the distribution of species: are bioclimate envelope models useful? Global Ecology & Biogeography 12(5), 361–371 (2003).
18. Helmuth, B.: From cells to coastlines: how can we use physiology to forecast the impact of climate change? Journal of Experimental Biology 212, 753-760 (2009).
19. Billings, C.G. Wei, H.-L., Thomas, P., Linnane, S.J., Hope-Gill, B.D.M.: The prediction of in-flight hypoxaemia using non-linear equations. Respiratory Medicine 107(6), 841-847 (2013).
20. Shamanand, J., Karspeck, A.: Forecasting seasonal outbreaks of influenza. Proceedings of the National Academy of Sciences USA 109(50), 20425–20430 (2012).
21. Zhang, Y., Bambrick, H., Mengersen, K., Tong, S., Hu, W.: Using google trends and ambient temperature to predict seasonal influenza outbreaks. Environment International 117, 284–291 (2018).
22. Osthus, D., Gattiker, J., Priedhorsky, R., Del Valle, S.Y.: Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy. Bayesian Analysis, in press, 2019.
23. Pisoni, E., Farina, M. Carnevale, C., Piroddi, L.: Forecasting peak air pollution levels using NARX models. Engineering Applications of Artificial Intelligence 22(4-5), 593-602 (2009).

24. Feng, X., Li, Q., Zhu, Y., Hou, J., Jin, L., Wang, J.: Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmospheric Environment 107, 118–128 (2015).
25. Bai, Y., Li, Y., Wang, X., Xie, J., Li, C.: Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions. Atmospheric Pollution Research 7, 557–566 (2016).
26. Sun, W., Sun, J.: Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. Journal of Environmental Management 188, 144–152 (2017).
27. Gu, Y., Wei, H.-L.: Significant indicators and determinants of happiness: Evidence from a UK survey and revealed by a data-driven systems modelling approach. Social Sciences 7(4), art. 53 (2018).
28. Zhang, W., Zhu, J., Gu, D.: Identification of robotic systems with hysteresis using nonlinear autoregressive exogenous input models, International Journal of Advanced Robotic Systems 14 (3), 1729881417705845 (2017).
29. Santos, R.F., Pereira, G.A.S., Aguirre, L.A.: Learning robot reaching motions by demonstration using nonlinear autoregressive models, Robotics and Autonomous Systems 107, 182-195 (2018).
30. Billings, S.A., Zhu, Q.M.: Rational model identification using an extended least-squares algorithm. International Journal of Control 54 (3), 529-546 (1991).
31. Chen, S., Billings, S. A., Luo, W.: Orthogonal least squares methods and their application to non-linear system identification. International Journal of Control 50(5), 1873-1896 (1989).
32. Chen, S., Cowan, C., Grant, P.: Orthogonal least squares learning algorithm for radial basis function networks. IEEE Transactions on Neural Networks 2(2), 302–309(1991).
33. Billings, S.A., Wei, H.-L.: The wavelet-NARMAX representation: a hybrid model structure combining polynomial models with multiresolution wavelet decompositions. International Journal of Systems Science 36(3), 137-152(2005).
34. Chen, S., Hong, X., Luk, B.L., Harris, C.J.: Orthogonal-least-squares regression: A unified approach for data modelling. Neural Computation 72(10), 2670–2681(2009).
35. Zhang, L., Li, K., Bai, E.-W., Irwin, G.W.: Two-stage orthogonal least squares methods for neural network construction. IEEE Transactions on Neural Networks and Learning Systems 26(8), 1608–1621(2014).
36. Guo, Y., Guo, L.Z., Billings, S.A., Wei, H.-L.: Identification of nonlinear systems with non-persistent excitation using an iterative forward orthogonal least squares regression algorithm. International Journal of Modelling, Identification and Control 23,1-7 (2015).
37. Yaghoobi, M., Davies, M. E.: Fast non-negative orthogonal least squares. In Proc. Eur. Sig. Proc. Conf., pp. 479–483, Nice, France (2015).
38. Li, Y., Cui, W.G., Guo, Y.Z. et al.: Time-varying system identification using an ultra-orthogonal forward regression and multiwavelet basis functions with applications to EEG. IEEE Transactions on Neural Networks and Learning Systems 29 (7), 2960-2972 (2018).
39. Wei, H.-L., Billings, S.A., Liu, J: Term and variable selection for nonlinear system identification. International Journal of Control 77, 86–110 (2004).
40. Wei, H.-L., Billings, S.A.: Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information. International Journal of Modelling, Identification and Control 3(4), 341-356 (2008).