Proceedings Paper:
Billings, S. and Wei, H. (2019) NARMAX model as a sparse, interpretable and transparent machine learning approach for big medical and healthcare data analysis. In: Proceedings of the 5th IEEE International Conference on Data Science and Systems. 5th IEEE International Conference on Data Science and Systems, 10-12 Aug 2019, Zhangjiajie, China. IEEE, pp. 2743-2750. ISBN: 978-1-7281-2059-1.

https://doi.org/10.1109/HPCC/SmartCity/DSS.2019.00385

Accepted by 5th IEEE Int. Con. on Data Science and Systems, 10-12 August 2019, Zhangjiajie, China.

NARMAX Model as a Sparse, Interpretable and Transparent Machine Learning Approach for Big Medical and Healthcare Data Analysis

Hua-Liang Wei [a,b], Stephen A. Billings [a,b]
a. *Department of Automatic Control and Systems Engineering*
b. *INSIGNEO Institute for in Silico Medicine*
University of Sheffield
Sheffield, United Kindom
w.hualiang@sheffield.ac.uk, s.billings@sheffield.ac.uk

***Abstract*—Influenza and influenza-like illnesses are one of the leading causes of death in the world, resulting in heavy losses to individual families and nations. Accurate and timely forecasts of seasonal influenza would therefore crucially important to inform and facilitate public health decision-making for presenting and intervening influenza epidemics. System identification and data-driven modelling approaches play an indispensable role in analyzing and understanding complex processes including medical, healthcare and environmental time series. This paper aims to present a type of sparse, interpretable and transparent (SIT) model, which cannot only be used for future behavior prediction but more importantly for understanding the dependent relationship between the response variables of a system on potential independent variables (also known as input variables or predictors). 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 based on input and output data of the system of interest, and the final refined model is usually simple, parsimonious and easy to interpret. The general framework of the NARMAX model is presented, and the state-of-the-art algorithms for such a SIT model estimation are described. Two case studies are provided to illustrate how well the SIT-NARMAX model can work for medical, healthcare and related data.**

Keywords—machine learning, system identification, data-driven model, time series, forecasting, NARMAX model

I. INTRODUCTION

Nowadays big data ubiquitously exists everywhere such as in space weather, airport and airline management, internet of things (IoT), medicine and healthcare. It is an inexorable trend that individual organizations will have more and more large scale data collected from the services, applications and platforms they provide [1], and the growth of data volume will be exponentially growing [2]. Big data offers the chance to better understand the underlying basis and nature of the subject of interest. For example, big data enables improved information processing and knowledge discovery from medical and health records [3],[4] and promises better healthcare [5].

Influenza is a major worldwide public health problem. According to the new estimates by the United States Centres for Disease Control and Prevention (US-CDC), the World Health Organization and global health partners, the annual influenza-related deaths (including cardiovascular disease and diabetes) are between 290,000 and 650,000 [6]. Our ability to effectively respond to major influenza outbreaks heavily relies on the accurate and timely prediction of their occurrences and activities. However, presently, our ability to forecast the timing, magnitude and duration of influenza outbreaks, based on existing approaches, is still quite limited [7],[8]. In 2008, Google launched its breakthrough service called Google Flu Trends (GFT), aiming at using Google search queries to do accurate prediction for influenza activity [9]. GFT is perhaps the first ever and most successful example of making use big data for public services. Although the internet-based surveillance GFT was shut down in 2015, it stimulated the development of new tools for public healthcare through big data mining (see e.g. [10]-[12]).

Once the query information collected from millions of users is properly aggregated, the resulting data can then be used to build high-level models that make sense of the original scattered information. A variety of methods have been proposed for internet-based forecasting, including regularized regression [10], [13], multivariate time series modelling [14], [15], support vector machine [16], neural networks [17], random forest [18], dynamic Bayesian model [19], and deep learning [20].

More broadly, data-driven modelling approaches have well served as a powerful tool for medical, healthcare and environments, and related fields. The list of available methods in the literature is quite long, such as wavelet neural networks, [21], [22], support vector machines [23], deep neural networks [24]-[25], and system identification techniques [26], [27], just mention a few.

This study presents a type of sparse, interpretable and transparent (SIT) model for medical, healthcare and related data analysis. We propose to use the NARMAX (nonlinear autoregressive moving average with exogenous inputs) model, which possesses a number of attractive 'SMART' properties (namely, simple and simulatable, meaningful, accountable, reproducible, and transparent) [28]. Several examples are presented to show how well the proposed SIT-NARMAX model works for medical, healthcare and related data analysis problems.

The remainder of the paper is as follows. In Section II, a brief introduction to data-driven modelling is presented. In Section III, the NARMAX model structured is described in detail. In Section IV, two case studies are provided, one

concerning the relation between mortality and influenza-like illness incidence rate, and another focusing on influenza forecasting. The work is briefly summarized in Section V.

II. Data-driven Modelling – A System Identification Persipective

Many practical data modelling problems can be described as follows. There is a response variable y (also known as output or dependent variable) that depends on a set of independent variables $\mathbf{x} = \{x_1, x_2, ..., x_n\}$ (also known as input or explanatory variables). Usually, a number of observations of both the output and input variables are available, which are denoted by $\{y_k, \mathbf{x}_k\}$ (k =1, 2,…,N). The true quantitative representation of the relationship between the output y and the input $\mathbf{x}$ is in general not known. The central task of data modelling is to establish quantitative representations, e.g. mathematical models such as $\boldsymbol{y} = f(\mathbf{x}) + e$ (where e is model error), to approximate the input-output relationship as close as possible.

A variety of methods and algorithms are available in the literature for dealing with different types of nonlinear data based modelling problems, including system identification [28]-[30], data mining [31],[32], pattern recognition and classification [33], supervised statistical learning [34],[35]. Among these methods, system identification techniques provide a tool for deducing mathematical models from measured input and output data for dynamic processes. In general, the output signal y at time instant t depends on the past output values and exogenous input signals u_1, u_2, …, u_r in a form of $y(t)=F[y(t\text{-}1), \ldots, y(t\text{-}n_y), u_1(t\text{-}d), u_1(t\text{-}d\text{-}1), \ldots, u_1(t\text{-}n_u), \ldots, u_r(t\text{-}d), u_r(t\text{-}d\text{-}1), \ldots, u_r(t\text{-}n_u)] + e(t)$, where F is an unknown function that needs to be estimated from the measurements, r is the number of exogenous input variables, d is a time delay (usually $d = 1$), n_y is the time lag in the output, n_u is the time lag in the inputs, and $e(t)$ is noise or model error.

There are a diversity of methods and approaches for building a good function to approximate the function f or F for a given problem, such as polynomials [36]-[38], radial basis functions [39]-[41], and wavelet functions [42]-[46]. The polynomial based representation, due to its attractive properties [28], is the most commonly used basis functions.

III. NARMAX Model

A. Static Regression Model

Consider a multivariate regression problem, with n predictor variables, $x_1, x_2, \cdots, x_n$, and one response variable y. The modelling task is to investigate if there exits a function f that can map the predictor variables to the response variable such that

$$y(k) = f(x_1(k), x_2(k), \cdots, x_n(k)) + e(k) \tag{1}$$

where $x_i(\cdot)$ (i=1,2, …, n) and $y(\cdot)$ represent the sequence of the observed predictor and response variables, respectively, $e(\cdot)$ represents the model error; $f(\cdot)$ represents some linear or nonlinear functions.

In most cases the function $f(\cdot)$ is unknown, but can be approximated by different models. In this study, a polynomial based regression model is considered. Expanding model (1) by defining the function $f(\cdot)$ to be a polynomial of degree ℓ gives the representation

$$\begin{aligned} y(k) = \theta_0 &+ \sum_{i_1=1}^{n} \theta_{i_1} x_{i_1}(k) + \sum_{i_1=1}^{n} \sum_{i_2=i_1}^{n} \theta_{i_1 i_2} x_{i_1}(k) x_{i_2}(k) + \cdots \\ &+ \sum_{i_1=1}^{n} \cdots \sum_{i_\ell=i_{\ell-1}}^{n} \theta_{i_1 i_2 \cdots i_l} x_{i_1}(n) x_{i_2}(k) \cdots x_{i_\ell}(k) + e(k) \end{aligned} \tag{2}$$

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

$$h(x_1, x_2, x_3) = a_1 x_1^4 + a_2 x_2 x_3 + a_3 x_1^2 x_2 x_3^2$$

is ℓ = 2+1+2=5, which is determined by the last term, $a_3 x_1^2 x_2 x_3^2$. 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 representation (2) belongs to the class of linear-in-the-parameters models.

B. NARX Model

Taking the case of a one input (designated by u) and one output (designated by y) problem as an example, the NARX model that links the output y to the input u is written as

$$\begin{aligned} y(t) = F[&y(t-1), y(t-2), \cdots, y(t-n_y), u(t-d), \\ &u(t-d-1), \cdots, u(t-d-n_u)] + e(t) \end{aligned} \tag{3}$$

where $y(t)$, $u(t)$ and $e(t)$ are the measured system output, input and noise sequences respectively at time instant t (here we use t to represent dynamical system models), n_y, n_u and n_e are the maximum lags for the system output, input and noise; $F[\bullet]$ is some non-linear function to be determined, and d is a time delay (typically d = 0 or d = 1). The noise signal $e(t)$ is unmeasurable but can be estimated as the prediction errors: $\xi(t) = y(t) - \hat{y}(t)$, where $\hat{y}(t)$ is the predicted value at time instant k generated by an estimated model. The noise terms are included to accommodate the effects of measurement noise, modelling errors, and/or unmeasured disturbances. Note that in models (1) and (3), two different indices 'k' and 't' are respectively indicate that model (1) is static while model (2) is dynamic, but in the following the two indices will be unified to 't' for convenience of description.

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

$$x_m(t) = \begin{cases} y(t-m), & 1 \le m \le n_y \\ u(t-m+n_y), & n_y+1 \le m \le n = n_y + n_u \end{cases} \tag{4}$$

The NARX model (3) can then be written in the same form as (1), and it can further expressed as the linear-in-the-parameters representation (2). For example, for a simple case where d = 1, n_y =2, n_u =1, ℓ =2, the full NARX model is

$$\begin{aligned} y(t) = \theta_0 &+ \theta_1 y(t-1) + \theta_2 y(t-2) + \theta_3 u(t-1) \\ &+ \theta_4 y^2(t-1) + \theta_5 y(t-1) y(t-2) + \theta_6 y(t-1) u(t-1) \\ &+ \theta_7 y^2(t-2) + \theta_8 y(t-2) u(t-1) + \theta_9 u^2(t-1) + e(t) \end{aligned} \tag{5}$$

C. NARMAX Model

For NARX model (3), it is assumed that the noise signal $e(t)$ is an i.i.d. process or a white sequence. Such an assumption, however, may not always be true, since the noise $e(t)$ may be a correlated or coloured sequence for many real applications. For cases where the signal $e(t)$ is not a white sequence, an established NARX model may be biased. A solution to obtain an unbiased model is to introduce lagged noise variables $e(t\text{-}1)$, $e(t\text{-}2)$, …, $e(t\text{-}n_e)$ to the model to construct a NARMAX model [47]

$$\begin{aligned} y(t) = F[&y(t-1),\cdots,y(t-n_y),u(t-1),\cdots,u(t-n_u),\\ &u(t-n_u),e(t-1),\cdots,e(t-n_e)]+e(t) \end{aligned} \tag{6}$$

Model (6) include the NARX model (3) and several other linear and nonlinear representations e.g. AR, ARX, Volterra series models as special cases [48]. The NARMAX model (6) is easily accommodated in the linear-in-the-parameters form (2) by defining $x_k(t)$ in Eq. (4) as

$$x_m(t) = \begin{cases} y(t-m), & 1 \le m \le n_y \\ u(t-m+n_y), & n_y+1 \le m \le n = n_y+n_u \\ e(t-m+n_y+n_u), & n_y+n_u+1 \le m \le n \end{cases} \tag{7}$$

where $n = n_y + n_u + n_e$. Note that the noise signal $e(t)$ in model (7) is unmeasurable; in model identification procedure it is often replaced by the model residual sequence as follows. Let $\hat{F}[\cdot]$ be an estimator for the function $F[\bullet]$, the model errors $\varepsilon(t)$ can be estimated as

$$\begin{aligned} \varepsilon(t) &= y(t) - \hat{y}(t) \\ &= y(t) - \hat{f}(y(t-1),\cdots,y(t-n_y),u(t-1),\cdots, \\ &\qquad u(t-n_u),\varepsilon(t-1),\cdots,\varepsilon(t-n_e)) \end{aligned} \tag{8}$$

The variable $e(t)$ in (7) can then be replaced by $\varepsilon(t)$. Detailed discussions on how to iteratively calculate model parameters and update the model errors, and on how to verify and test the validity of a model can be found in [49]-[51].

D. NARMAX Model Estimation

It is known that for a linear model, the model terms and the variables are exactly the same. For a typical generalized linear model or a nonlinear model, however, variables and terms are generally distinct, and the distinction can be illustrated using the simple nonlinear model below [52]

$$\left.\begin{aligned} y = g(x_1,x_2) = a_0 + a_1x_1 + a_2x_2 \\ + a_3x_1^2 + a_4x_2^2 + a_5x_1e^{x_2} + a_6|x_1x_2| \end{aligned}\right\} \tag{9}$$

Here there are only two variables: x_1, x_2, but there are seven terms, that is, the *const*, x_1, x_2, x_1^2, x_2^2, $x_1e^{x_2}$, $|x_1x_2|$.

It is easy to know that the total number of potential model terms in the power-form polynomial model (2) is $M = (n+\ell)!/[n!\ell!]$, where again ℓ is the degree of nonlinearity. For example, if ℓ =3, $n_y = 2$, $n_u = 1$, $n_e = 3$, then $M = (6+3)!/(6!3!) = 84$. For the dynamic regression model (3), with power-form polynomials as basis functions, if n_y and n_u are large, the number of candidate model terms included in the initial full model can be very large. However, numerous practical applications show that in almost all real-data cases, generally only a small number candidate model terms are important for characterizing the underlying dynamics and all the other candidate terms are either not important or irrelevant and thus can be ignored.

The forward regression orthogonal least squares (FROLS) algorithms [28], [53]-[54] provides an efficient, powerful tool for nonlinear model term selection and model structure detection. A detailed discussion of the FROLS algorithm and ERR index can be found in [28], [53]-[54]. Here, we only give a very brief summary of the algorithm. FROLS searches through all the possible candidate model terms to select the most significant terms one by one. The significance of each of the model terms is measured by an index, called the error reduction ratio (ERR) [53], which evaluate how much of the variance change in the system response can be accounted for by including the relevant model term. There exist such cases where some model terms only make a very small contribution (measured by ERR values) but they may be statistically significant and are therefore also included in the models. Some statistical criteria, e.g. AIC, BIC, PESR (penalized error-to-signal ratio) [55], APRESS (adjusted (PRediction error sum of squares) [56], [57], can be used to monitor the model selection procedure and determine the model complexity.

IV. Case Studies

A. 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. 1.

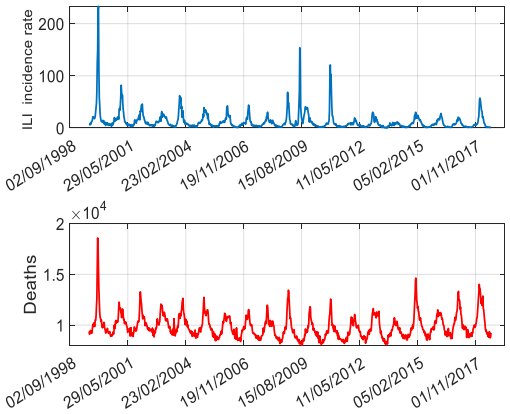

Fig. 1. Weekly influenza-like illness (ILI) incidence rate and deaths, England and Wales, between week 31 of 1999 and week 30 of 2018.

The objective here is twofold. One is to reveal how the week mortality relates to the ILI incidence rate through data-driven modelling approach, and another is to do one-week-ahead prediction of the death mortality. 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 NARMAX method is applied to the training data, and the best NARMAX model is:

$$\left.\begin{aligned} y(t) &= 616.435147 + 0.927840\,y(t-1) \\ &\quad -0.114871x(t-1)x(t-3) + 10.535455x(t-1) \end{aligned}\right\} \quad (10)$$

where $x(t)$ represents the weekly ILI incidence rate and $y(t)$ represents the number of weekly deaths. Note that all the model terms involving noise variables such as $x(t\text{-}1)e(t\text{-}1)$ 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. 2 and Fig. 3, respectively.

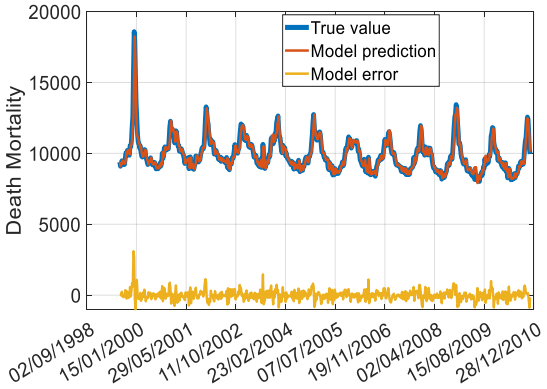

Fig. 2. A comparison of the model prediction with the corresponding true number of deaths, on the training dataset of the period between week 31 of 1993 and week 47 of 2010.

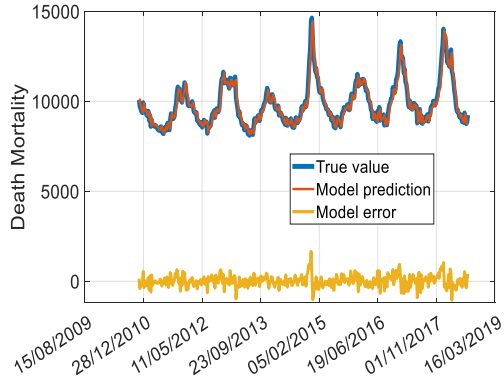

Fig. 3. A comparison of the model prediction with the corresponding true number of deaths, on the test dataset of the period between week 48 of 2010 and week 30 of 2018.

Model (10) shows that the death mortality is correlated to the influenza-like illness (ILI) incidence rate of one and three weeks ago. From Fig. 2 and Fig. 3, it can be seen that the simple NARMAX model shows an excellent prediction performance.

B. Influnza Forecasting

1) *Data*

Google Flu Trends (GFT) provides a successful example of making use big data to predict the future and to significantly improve forecasting performance. GFT involves a massive volume of structured or unstructured datasets which are very large and complex. Through a comprehensive processing procedure, workable and easy to use datasets GFT datasets became available. For example, 100 sub-datasets extracted from GFT datasets, containing the influenza information of the United States of America during the period from 1 June 2003 to 11 May 2008, were available and 45 of which were used for influenza forecasting [9].

2) *Models*

In this study, we consider the first sub-dataset used in [9]. The dataset contains digital information of the general influenza symptoms of the USA of during the period from 1 June 2003 to 11 May 2008. The sub-dataset contains values nine variables, representing the nine census divisions of the USA, which are geographically shown in Fig. 4. For convenience, the general influenza symptoms of the nine census divisions are represented by following variables: Y_1 (NE), Y_2 (MA), Y_3 (ENC), Y_4 (WNC), Y_5 (SA), Y_6 (ESC), Y_7 (WSC), Y_8 (Mountain), Y_9 (Pacific).

The sub-dataset used comprises 259 weekly data points in total. The first 135 data points of the period from 1 June 2003 to 25 December 2005 are used for model training and the remaining 124 for model performance test. The 135 training data for the nine census divisions are plotted in Fig. 5.

Fig. 4. The nine census divisions of the America. https://www.ncdc.noaa.gov/monitoring-references/maps/us-census-divisions.php

The NARMAX method, together with an iterative orthogonal forward regression algorithm (iOFR) [58],[59], is applied to the 135 weekly training data points of the period from from 1 June 2003 to 25 December 2005 (shown in Fig. 5). Our objective is to estimate nine predictive models for each of the census divisions of USA; these models will be used for one-week ahead forecasting for influenza trends. The initial target models are of the form:

$$Y_i(t) = F_i[Y_1(t-1),...,Y_1(t-3),...,Y_9(t-1),...,Y_9(t-3)] \quad (11)$$

where the functions $F_i[\cdot]$ (i=1,2,…, 9) are chosen to be polynomials of nonlinear degree ℓ =2. Note that initially each of the nine models in (11) involves a total of 406 candidate model terms, but the NARMAX estimation procedure will identify the most important model terms and produce a sparse model for each of the nine cases.

The predictive model identified for each of the nine census divisions are presented as follows.

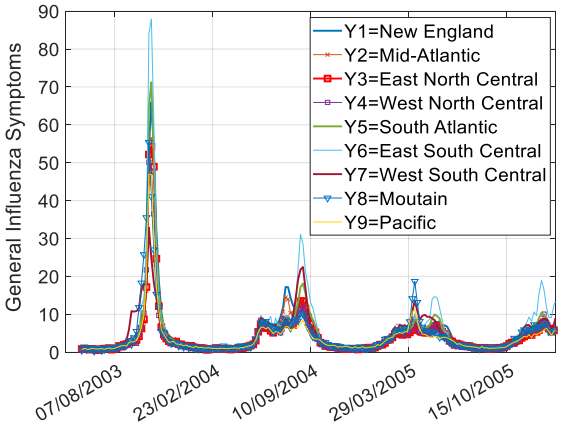

Fig. 5. The influenza information of the nine census divisions of the United States of America during the period from 1 June 2003 to 31 Dec 2006.

$$\left.\begin{aligned} Y_1(t) = {} & 0.7573Y_8(t-1) + 0.6645Y_4(t-1) - 0.4174Y_3(t-1) \\ & - 0.1188Y_7(t-1) + 1.2207Y_2(t-1) - 0.6408Y_9(t-2) \\ & - 0.0245Y_7(t-1)Y_9(t-1) + 0.0777Y_8(t-1)Y_8(t-1) \\ & - 0.1522Y_7(t-1)Y_8(t-1) + 0.0480Y_7(t-1)Y_7(t-1) \\ & - 0.4354 \end{aligned}\right\} \quad (12)$$

$$\left.\begin{aligned} Y_2(t) = {} & 0.4792Y_8(t-1) + 0.7703Y_4(t-1) \\ & - 0.0156Y_7(t-1)Y_9(t-1) \end{aligned}\right\} \quad (13)$$

$$\left.\begin{aligned} Y_3(t) = {} & -0.3450Y_8(t-1) + 0.0584Y_7(t-2)Y_8(t-1) \\ & - 0.0063Y_6(t-2)Y_8(t-1) + 1.3012Y_4(t-1) \\ & - 0.0456Y_4(t-2)Y_7(t-1) \end{aligned}\right\} \quad (14)$$

$$\left.\begin{aligned} Y_4(t) = {} & 0.7573Y_8(t-1) - 0.5256Y_3(t-3) \\ & + 0.3131Y_2(t-3) + 0.2234Y_8(t-2) \end{aligned}\right\} \quad (15)$$

$$\left.\begin{aligned} Y_5(t) = {} & 0.2593Y_8(t-1) + 0.0415Y_8(t-1)Y_8(t-1) \\ & + 0.1275Y_7(t-2) - 0.0845Y_7(t-1)Y_9(t-1) \end{aligned}\right\} \quad (16)$$

$$\left.\begin{aligned} Y_6(t) = {} & 1.4107Y_8(t-1) - 0.4475Y_9(t-3) \\ & + 0.0293Y_7(t-1)Y_7(t-3) \end{aligned}\right\} \quad (17)$$

$$\left.\begin{aligned} Y_7(t) = {} & 1.1638Y_7(t-1) - 0.0128Y_6(t-2)Y_7(t-1) \\ & + 0.0302Y_7(t-2)Y_8(t-2) + 0.0018Y_5(t-1)Y_7(t-2) \\ & - 0.6441Y_7(t-2) - 0.0061Y_5(t-1)Y_5(t-1) \end{aligned}\right\} \quad (18)$$

$$\left.\begin{aligned} Y_8(t) = {} & 1.6721Y_8(t-1) - 0.0279Y_1(t-1)Y_7(t-1) \\ & - 0.4851Y_3(t-3) + 0.0192Y_1(t-3)Y_7(t-2) \end{aligned}\right\} \quad (19)$$

$$Y_9(t) = 0.9814Y_8(t-1) - 0.0091Y_2(t-2)Y_7(t-1) \quad (20)$$

Note that model terms in each of the models are arranged in order of their significance for explaining the changing trend of the output variables (influenza symptoms). The CPU time used for generating each of these models is about 0.42 seconds on average. It is very interesting to notice that all the output variables (except Y_7 for West South Central) are closely related to the lagged variable $Y_8(t\text{-}1)$, this probably implies that the outbreaks of influenza in other census divisions are closely associated to the influenza epidemics of a week ago in the Mountain region.

3) *Model performance*

We use two statistics, namely, root mean squared error (RMSE) and mean absolute error (MAE), to evaluate the performance of the identified models. The values of the two statistics for each of the models (12)-(20), over the test data (1 January 2006 to 11 May 2008) are given in Table I.

TABLE I. THE TWO STATISTICS RMSE (ROOT MEAN SQUARED ERROR) AND MAE (MEAN ABASOLUTE ERROR), BOTH ARE THE TEST DATA. RAW 2: RMSE; RAW 3: MAE.

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Y_7	Y_8	Y_9
1.3429	1.5851	1.5818	2.0799	1.9912	4.3526	1.4697	1.9036	1.3154
0.8596	0.9896	0.8532	1.0472	1.1037	2.0421	0.8076	0.9416	0.7093

A comparison between the model prediction and the corresponding observations for the nine census divisions are illustrated in Fig.6 (for New England, Mid-Atlantic and East North Central regions) , Fig. 7 (for West North Central, South Atlantic and East South Central regions) and Fig. 8 (West South Central, Mountain and Pacific regions), respectively.

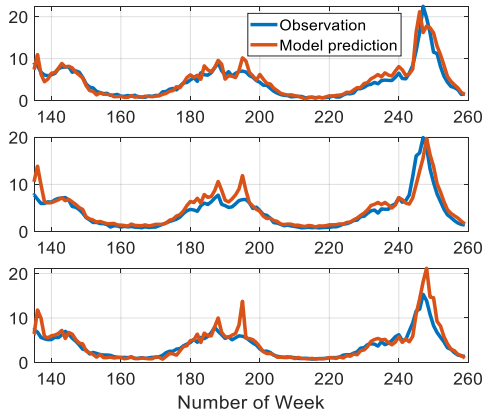

Fig. 6. A comparison between the model prediction and the observed influenza symptoms for the three census divisions: New England, Mid-Atlantic and East North Central regions, during the period from 1 January 2006 to 11 May 2008.

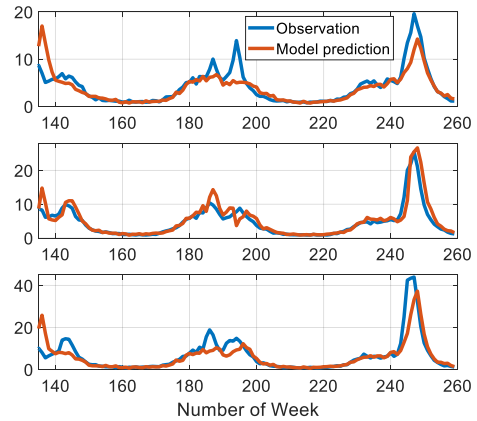

Fig. 7. A comparison between the model prediction and the observed influenza symptoms for the three census divisions: West North Central, South Atlantic and East South Central regions, during the period from 1 January 2006 to 11 May 2008.

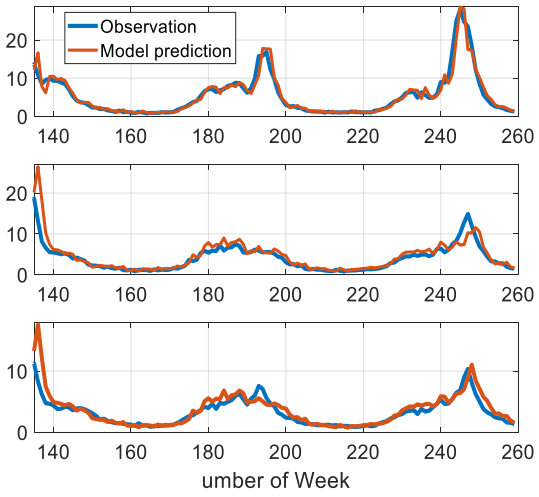

Fig. 8. A comparison between the model prediction and the observed influenza symptoms for the three census divisions: West South Central, Mountain and Pacific regions, during the period from 1 January 2006 to 11 May 2008.

V. Conclusion

The study focuses on a class of 'SMART' (simple and simulatable, meaningful, accountable, reproducible, and transparent) representations, called SIT-NARMAX model, which has a number of attractive properties for big medical and heathcare data-driven modelling problems. The main contributions of this study are as follows. Firstly, a simple but effective model was established to represent the relationship between weekly mortality and the influenza-like illness (ILI) incidence rate of England and Wales; as shown in the first case study, the model provides a perfect representation of the relation between the weekly influenza-like illness (ILI) incidence rate and death mortality. Secondly, nine transparent models have been built from Google flu trends data; the models reveal how the general influenza symptoms of each of the census divisions of the America are closely associated with that in the other regions. These promising results suggest that the NARMAX model and the associated methodology provide a powerful tool for medical and healthcare data modelling problems, especially for cases where the interest is not only in prediction and forecasting but also strongly in interpreting and explaining the dependent relation of the output (response) variables on a few important candidate input (explanatory) features.

Acknowledgment

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 EU Horizon 2020 Research and Innovation Programme Action Framework under grant agreement 637302.

References

[1] V. Mayer-Schonberger and K. Cukier,Big Data: A Revolution That Will Transform How We Live, Work, and Think, Boston: Houghton Mifflin Harcourt, 2013

[2] D. Howe et al., "Big data: the future of biocuration", Nature, vol. 455, pp.47-50, 2008.

[3] O. Ola and K. Sedig, "The challenge of big data in public health: An opportunity for visual analytics," Online J. Public Health Informat., vol. 5, art. no. 223, 2014.

[4] P. Y. Wu, C. W. Cheng, C. Kaddi, J. Venugopalan, R. Hoffman, M. D. Wang, "Omic and electronic health record Big Data analytics for precision medicine," IEEE Trans. Biomed. Eng., 64, pp. 263–273, 2017.

[5] A. F. Simpao, L. M. Ahumada, and M. A. Rehman, "Big data and visual analytics in anaesthesia and health care," British J. Anaesth. Vol. 115, pp. 350–356, 2015.

[6] World Health Organization (2014) Influenza (seasonal) (World Health Org, Geneva) [Access on 28 January 2019] https://www.who.int/news-room/detail/14-12-2017-up-to-650-000-people-die-of-respiratory-diseases-linked-to-seasonal-flu-each-year.

[7] J. Shamanand A. Karspeck, "Forecasting seasonal outbreaks of influenza," Proc Natl Acad Sci USA, vol. 109, no. 50, pp. 20425–20430, 2012.

[8] S. Yang, M. Santillana, and S. C. Kou, "Accurate estimation of influenza epidemics using Google search data via ARGO," Proc. Natl Acad. Sci. USA, vol. 112, no. 47, pp.14473–14478, 2015.

[9] J. Ginsberg, M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," Nature, vol. 457, no. 7232, pp. 1012–1014, 2009.

[10] V. Lampos, A. C. Miller, S. Crossan, and C. Stefansen, "Advances in nowcasting influenza-like illness rates using search query logs," Scientific Reports, vol. 5, art. no.12760, 2015.

[11] M. Wagner, V. Lampos, I. J. Cox, and R. Pebody, "The added value of online user-generated content in traditional methods for influenza surveillance," Scientific Reports, vol. 8, art. no. 13963, 2018.

[12] S. Yang, M. Santillana, J. S. Brownstein, J. Gray, S. Richardson, S. C. Kou, "Using electronic healthrecords and Internet search information for accurate influenza forecasting," BMC InfectiousDiseases, vol. 17, no. 1, art. 332, 2017.

[13] D. Osthus, A. R. Daughton, and R. Priedhorsky, "Even a goodinfluenza forecasting model can benefit from internet-based nowcasts, but those benefits are limited," PLOS Comp Bio, in press, 2019.

[14] T. Preis, H. S. Moat, "Adaptive nowcasting of influenza outbreaks using Google searches," Royal Society Open Science, vol. 1, no. 2, pp. 1-5, 2014.

[15] Y. Zhang, H. Bambrick, K. Mengersen, S. Tong, and W. Hu, "Using google trendsand ambient temperature to predict seasonal influenza outbreaks," Environment International, vol. 117, pp. 284–291, 2018.

[16] M. Santillana, A. T. Nguyen, M. Dredze, M. J. Paul, E. O. Nsoesie, J. S. Brownstein, "Combining search, social media, and traditional data sources to improve influenza surveillance," PLOS Computational Biology, vol. 11, no. 10, art. 1004513, 2015.

[17] S. Volkova, E. Ayton, K. Porterfield, and C. D. Corley, "Forecasting influenza-like illness dynamics for military populations using neural networks and social media," PLOS ONE, vol. 12, no. 12, pp.1-22, 2017.

[18] S. Kandula, D. Hsu, J. Shaman, "Subregional nowcasts of seasonal influenza using search trends," Journal of Medical Internet Research, vol. 19, no. 11, art. 370, 2017.

[19] D. Osthus, J. Gattiker, R. Priedhorsky, S. Y. Del Valle, "Dynamic Bayesian influenza forecasting in the United States with hierarchical discrepancy", Bayesian Analasys, in preess, 2019.

[20] G. Xi, L. Yin, Y. Li, and S. Mei, "A deep residual network integrating spatial-temporal properties to predict influenza trends at an intra-urban scale", In Proceeding GeoAI'18 Proceedings of the 2nd ACM SIGSPATIAL International Workshop on AI for Geographic Knowledge Discovery, pp.19-28, Seattle, WA, USA, November 06 - 06, 2018.

[21] X. Feng, Q. Li, Y. Zhu, J. Hou, L. Jin, J. Wang, "Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation," Atmos. Environ. vol. 107, pp. 118–128, 2015.

[22] Y. Bai, Y. Li, X. Wang, J. Xie, C. Li, "Air pollutants concentrations forecasting using back propagation neural network based on wavelet decomposition with meteorological conditions," Atmos. Pollut. Res., vol. 7, pp. 557–566, 2016.

[23] W. Sun and J. Sun, "Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm," Journal of Environmental Management, vol. 188, pp. 144–152, 2017.

[24] A. Esteva, B. Kuprel, R. Novoa, J. Ko, S. Swetter, H. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks,", Nature, vol. 542, no. 7639, pp.115–118, 2017.

[25] E. Choi, A. Schuetz, W. F. Stewart, J. Sun, "Usingrecurrent neural network models for early detec-tion of heart failure onset," J. American Med. Inform. Assoc. vol. 24, no. 2, pp. 361 –367, 2017.

[26] Y. Gu and H.-L. Wei, "Significant indicators and determinants of happiness: Evidence from a UK survey and revealed by a data-driven systems modelling approach," Social Sciences, vol. 7, no. 4, art. 53, 2018.

[27] C.G. Billings, H.-L. Wei, P. Thomas, S.J. Linnane, and B.D.M. Hope-Gill, "The prediction of in-flight hypoxaemia using non-linear equations," Respiratory Medicine, vol. 107, no. 6, pp. 841-847, 2013.

[28] S. A. Billings, Nonlinear System Identification: NARMAX Methods in the Time, Frequency, and Spatio-Temporal Domains, Chichester, West Sussex: Wiley, 2013.

[29] L. Ljung. System Identification: Theory for the User, Upper Saddle River, N.J.: Prentice Hall, 1999.

[30] O. Nelles, Nonlinear System Identification. Heidelberg, Berlin: Springer-Verlag, 2011.

[31] J. Han and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, Calif, USA, 2001.

[32] I. H. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques, Morgan Kaufmann, San Francisco, Calif, USA, 2005.

[33] C. M. Bishop, Neural networks for pattern recognition, Oxford University Press, 1995.

[34] V. Vapnik, The Nature of Statistical Learning Theory (2nd ed.), New York, NY: Springer, 1999.

[35] J. Friedman, T. Hastie, and R. Tibshirani, The Elements of Statistical Learning, New York: Springer, 2001.

[36] L. Aguirre and S. Billings, "Retrieving dynamical invariants fromchaotic data using NARMAX models," Int. J. Bifurc. Chaos, vol. 5,no. 2, pp. 449–474, 1995.

[37] L. A. Aguirre and S. A. Billings, "Improved structure selection fornonlinear models based on term clustering," Int. J. Contr., vol. 62, pp.569–587, 1995.

[38] J. R. Ayala Solares, H.-L. Wei, and S. A. Billings, "A novel logistic-NARX model as a classifier for dynamic binary classification," Neural Computing and Applications, vol. 31, no. 1, pp. 11–25, 2019.

[39] S. Chen, S. A. Billings, C. F. N. Cowan, and P. M. Grant, "Practicalidentification of NARMAX models using radial basis functions," Int. J. Control, vol. 52, no. 6, pp. 1327–1350, 1990.

[40] S. Chen, X. Hong, and C. J. Harris, "Sparse kernel regression model-ling using combined locally regularized orthogonal least squares and D-optimality experimental design," IEEE Trans. Autom. Control, vol.48, no. 6, pp. 1029–1036, Jun. 2003.

[41] S.A. Billings, H.-L. Wei, and M. A. Balikhin, "Generalized multiscale radial basis function networks," Neural Networks, vol. 20, no. 10, pp.1081-1094, 2007.

[42] S. A. Billings and H.-L. Wei, "The wavelet-NARMAX representation: a hybrid model structure combining polynomial models with multiresolution wavelet decompositions," International Journal of Systems Science, vol. 36, no. 3, pp.137–152, 2005.

[43] H.-L. Wei and S. A. Billings, "A unified wavelet-based modelling framework for non-linear system identification: the WANARX model structure," International Journal of Control, vol. 77, no. 4, pp.351–366, 2004.

[44] H.-L. Wei and S. A. Billings, "Long term prediction of non-linear time series using multiresolution wavelet models," International Journal of Control, vol. 79, no. 6, 569–580, 2006.

[45] H.-L. Wei, S. A. Billings, Y. Zhao, and L. Guo, "Lattice dynamical wavelet neural networks implemented using particle swarm optimization for spatio–temporal system identification," IEEE Transactions on Neural Networks, vol. 20, no. 1, pp. 181-185, 2009.

[46] H.-L. Wei, S. A. Billings, Y. Zhao, and L. Guo, "An adaptive wavelet neural network for spatio-temporal system identification," Neural Networks, vol. 23, no. 10, pp.1286-1299, 2010.

[47] I. J. Leontaritis, and S. A. Billings, "Input-output parametric models for non-linear systems—part I: deterministic non-linear systems; part II: stochastic non-linear systems," *Int. J. Control*, vol. 41, pp.303-344, 1985.

[48] R. K. Pearson, *Discrete-Time Dynamic Models*. Oxford: Oxford University Press, 1999.

[49] S. A. Billings, and W. S. F. Voon, "Correlation based model validity tests for nonlinear models," *Int. J. Control*, vol. 44, pp.235- 244, July 1986.

[50] S. A. Billings, and Q. M. Zhu, "Nonlinear model validation using correlation tests," Int. J. Control, vol. 60, pp.1107- 1120, Dec. 1994.

[51] L. F. Zhang, Q. M. Zhu, and A. Longden, "A correlation-test-based validation procedure for identified neural networks," IEEE Transactions on Neural Networks, vol. 20, no. 1, pp. 1-13, 2009.

[52] H.-L. Wei, S. A. Billings, and J. Liu, "Term and variable selection fornonlinear system identification," Int. J. Control, vol. 77, pp. 86–110, 2004.

[53] S. Chen, S. A.Billings, and W. Luo, "Orthogonal least squares methods and their application to non-linear system identification," *International Journal of Control*, vol. 50, no. 5, pp. 1873-1896, 1989.

[54] H.-L. Wei and S. A. Billings, "Model structure selection using an integrated forward orthogonal search algorithm assisted by squared correlation and mutual information," *International Journal of Modelling, Identification and Control*, vol. 3, no. 4, pp. 341-356, 2008.

[55] H.-L. Wei, S. A. Billings, Y. Zhao, and L. Z. Guo, "An adaptive wavelet neural network for spatio-temporal system identification," *Neural Networks*, vol. 23, no. 10, pp. 1286-1299, 2010.

[56] S.A. Billings and H.-L. Wei, "An adaptive orthogonal search algorithm for model subset selection and non-linear system identification," International Journal of Control, vol. 81, no. 5, pp. 714-724, 2008.

[57] Y. Gu and H.-L. Wei, "A robust model structure selection method for small sample size and multiple datasets problems," Information Sciences, vol. 451, pp. 195-209, 2018.

[58] Y. Guo, L.Z. Guo, S.A. Billings, and H.-L. Wei, "An iterative orthogonal forward regression algorithm," International Journal of Systems Science, vol. 46, no. 5, 776-789, 2015.

[59] Y. Guo, L.Z. Guo, S.A. Billings, and H.-L. Wei, "Identification of nonlinear systems with non-persistent excitation using an iterative forward orthogonal least squares regression algorithm," International Journal of Modelling, Identification and Control, vol. 23, no. 1, pp. 1-7, 2015.