2nd International Joint Conference on Water Distribution Systems Analysis & Computing and Control in the Water Industry
Valencia (Spain), 18-22 July 2022
doi: https://doi.org/10.4995/WDSA-CCWI2022.2022.14170

ON THE USE OF SINDY FOR WDN

Mario Castro-Gama[1]

[1] Specialist Infrastructure for Water Expertise Center (WEC) at Vitens N.V., Zwolle (The Netherlands)

[1] *Mario.Castrogama@Vitens.nl*

Abstract

With the growing interest of water utilities on digitalization, running multiple scenarios can become cumbersome with limited budget and short data collections. The total number of hydraulic simulations required (usually using commercial software), becomes a burden for near real-time operation. In order to circumvent the computational burden (limitation), since a couple of decades, several Machine Learning techniques have been used to create a meta-model or surrogates of a Water Distribution Networks (WDN) based on a subset of data available through SCADA. Among the many possible surrogates a Sparse Identification of Non-linear Dynamics (*SINDy*) method is presented. The method is applied to two datasets: i) to obtain a surrogate of a benchmark network and ii) real data of water consumption of different District Metered Areas (DMA) of a real water utility. The method is: i) computationally inexpensive, ii) less data demanding for calibration than other modern methods, iii) parsimonious, and iv) could be used to infer physical relations among data.

Keywords
Water demand, WDN, Surrogates, SINDy, DMD.

1 INTRODUCTION

Since a couple of decades ago there has been an increase on the number assets which are registered by utilities and used as elements of hydraulic models. Such models are then used to for real-time applications. With the development of Digital Twins [1] by utilities the number of scenarios and decisions which can be pursued by operators increases exponentially the requirement of hydraulic simulations. This creates a trade-off between model size and number of simulations which can be carried out to answer a specific question (i.e. leakage detection, anomaly detection, long-term and short-term planning, condition assessment). A way of circumventing this trade-off is the use of additional computer processing, parallelization of model runs, skeletonization or model simplification or the use surrogates or meta-models of the Water Distribution Network (WDN) model. Although pushing additional computer processing and capacity in data warehouses it's a possibility for some utilities, it is not sustainable in the long run. Within the meta-models category several applications for meta-modelling of WDN are available. Methods ranging from neural networks [2] [3], from simple types such as generalized and perceptron multilayers [4] [5], to more recent developments such as Deep learning networks (DNN) [6] are available. Most of these meta-models encapsulate a large amount of data (i.e. pipe flows, pressures or heads and demands) as black-box and their physical interpretation gets lost in the inner workings of these non-linear regressions. Since a few years ago some methods have been (re)discovered for the identification of principal modes from complex systems such as turbulent flows in the form of Koopman operators.

There is a large amount of such methods such as Principal Component Analysis (PCA), ERA, PDO, ICA, KIC, Dynamic Mode Decomposition (DMD) [7] [8] and Sparse Identification Non-Linear Dynamics (SINDy) [9] [10]. Here the latter and its possible applications for WDN are presented.

2 SPARSE IDENTIFICATION NON-LINEAR DYNAMICS - SINDY

SINDy states that given a set of measurement data $\{x(t)\}_{i \in I}$, it is possible to accurately learn a function $f(x(t))$ so that $\frac{dx}{dt} = f(x(t))$ is identified. Two assumptions are required, i) the full state measurements, and ii) that f only has a few active terms, (i.e. f is sparse) in the space of all possible functions of $x(t)$.

$$X = \begin{matrix} x^T(t_1) \\ x^T(t_2) \\ \vdots \\ [x^T(t_m)] \end{matrix} = [\begin{matrix} x_1(t_1) & x_2(t_1) & \cdots & x_n(t_1) \\ x_1(t_2) & x_2(t_2) & \cdots & x_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(t_m) & x_2(t_m) & \cdots & x_n(t_m) \end{matrix}] \tag{1}$$

The rate of change of X can be estimated using finite differences or total variation derivatives. Or simply by assuming an Euler update by taking the next time step as the outcome when dt is small.

$$\dot{X} = \begin{matrix} \dot{x}^T(t_1) \\ \dot{x}^T(t_2) \\ \vdots \\ [\dot{x}^T(t_m)] \end{matrix} = [\begin{matrix} \dot{x}_1(t_1) & \dot{x}_2(t_1) & \cdots & \dot{x}_n(t_1) \\ \dot{x}_1(t_2) & \dot{x}_2(t_2) & \cdots & \dot{x}_n(t_2) \\ \vdots & \vdots & \ddots & \vdots \\ \dot{x}_1(t_m) & \dot{x}_2(t_m) & \cdots & \dot{x}_n(t_m) \end{matrix}] \tag{2}$$

In order to solve this the first step is to construct library $\Theta(X)$ of candidate nonlinear functions of X :

$$\Theta(X) = [\begin{matrix} \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X & X^2 & X^3 & \cdots & X^p \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \end{matrix}]$$

$$\text{i.e. for } p = 2,\ X^2 = \begin{matrix} x_1^2(t_1) & x_1(t_1)x_2(t_1) & \cdots & x_2^2(t_1) & \cdots & x_n^2(t_1) \\ x_1^2(t_2) & x_1(t_2)x_2(t_2) & \cdots & x_2^2(t_2) & \cdots & x_n^2(t_2) \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ [x_1^2(t_m) & x_1(t_m)x_2(t_m) & \cdots & x_2^2(t_m) & \cdots & x_n^2(t_m)] \end{matrix} \tag{3}$$

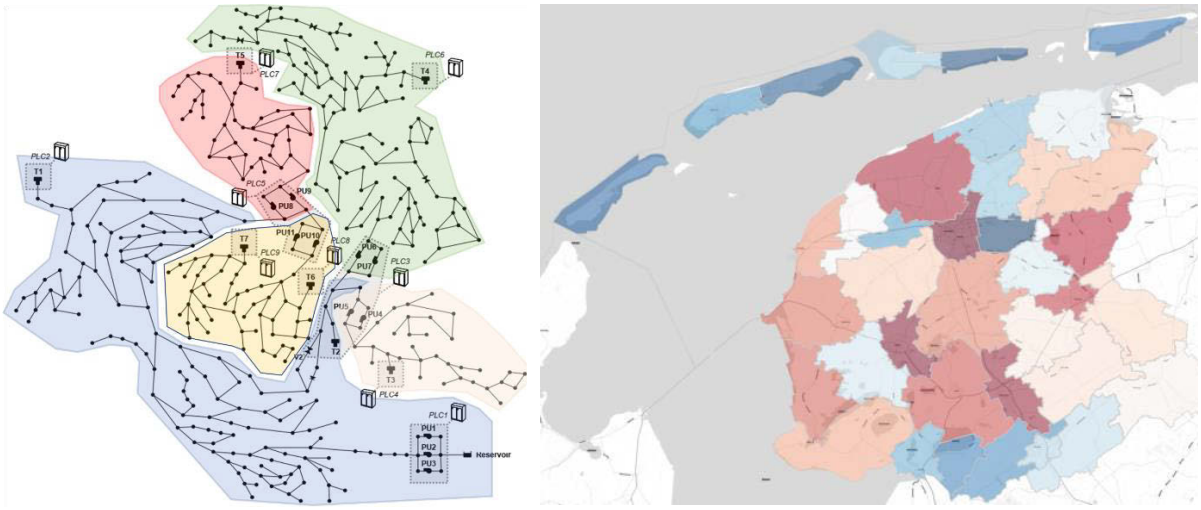
Then perform a sparse regression on $X = \Theta(X)\sum$. This is required to solve for all coefficients $\sum = [\sigma_1; \ldots ; \sigma_n]$, $\sigma_i \in R_p$. Then, let $\lambda > 0$ be the sparsity threshold and the following iterative procedure ensures that an sparse regression is obtained.

1. Initial guess: solve $X = \Theta(X)\sum$. via ordinary least squares
2. If $\sum(i,j) < \lambda$ set $\sum(i,j) = 0$
3. For $k = 1, 2, \ldots, n$

   solve $\dot{X}(:, k) = \Theta(X)(:, \Sigma(:, k) > \lambda)\Sigma(\Sigma(:, k) > \lambda), k)$ via least squares
4. Repeat steps 2-3 until the coefficients do not change (or for a fixed number of iterations)

3 CASE STUDIES

Two different case studies are used. Firstly, is the simulated data from a benchmark WDN known as C-Town. Secondly, the water balance data of a province located in the northern part of the Netherlands and operated by Vitens.



(a) C-Town- with 5 DMA's. Shows also selected monitoring locations for tanks and pumping stations.

(b) Areas of water balance (large DMA's) in a province of the Netherlands operated by Vitens N.V.

Figure 1. Case studies

3.1 C-Town

It corresponds to a small WDN with 5 DMA's from which flows and pressures at particular locations can be fetched from the system. In this case study two different configurations are assumed (Fig 1a).

- First, a configuration in which all variables are observed. In this case a total of 724 variables is considered. Although it is unrealistic to collect all variables related to demands, pressures and flows within a WDN, the goal here is to determine whether or not SINDy is able to reconstruct both mass and energy balance according to the Global Gradient Algorithm (GGA) [11] [12], without prior knowledge of the equations. A dataset of 4 weeks (2688 timestamps every 15 minutes) is created.
- Second, a configuration of SINDy in which only a subset (43 variables) in the system are collected in the SCADA system for each DMA is presented. Variables which show no variation during the total length of the dataset where eliminated resulting in only 37 variables. The goal in this case is to be able to determine whether or not anomalies can be detected. Anomalies can represent multiple behaviours such as change of valve status, leakages, or even cyber-physical attacks. Here SINDy is compared to another surrogate. Two datasets are obtained from BATADAL (Battle of Attack Detection Algorithms) one of *normal* operation of the WDN and one *abnormal* (with anomalies). A SINDy model of the normal operation data is trained and subsequently tested on the abnormal data. The hypothesis is that SINDy is able to capture the system dynamics and will be able to identify the timestamps of anomalies as such.

3.2 Water balance areas of a province

Data collected from the last 4 full calendar years (2018-2021) of a northern province operated by Vitens are analysed. Data corresponds to the water balance in each of the Water Balance Areas (WBA) of the province. It is not possible to assess WBA as District Metered Areas (DMA's) due to the former being much larger. Fig. 1b, presents the localization of each of the WBA, while Fig. 2 represents the total water consumption pattern of the province within a 24 hour period at an hourly resolution. This water consumption is obtained by taking into account all the production

locations of the province. The use of Automated Metered Readings (AMR's) is only available for a pilot area in the largest city of the province and for large customers, however given the low leakage percentage Fig. 2 is representative of the demand pattern. In this case the area is composed of 12 different DMA's where data is available between 2018 end 2021. The homogeneous period (Fig 2 below) where data is available for all variables is 11-Nov-2019 and 30-Oct-2020 (7840 timestamps) is highlighted.

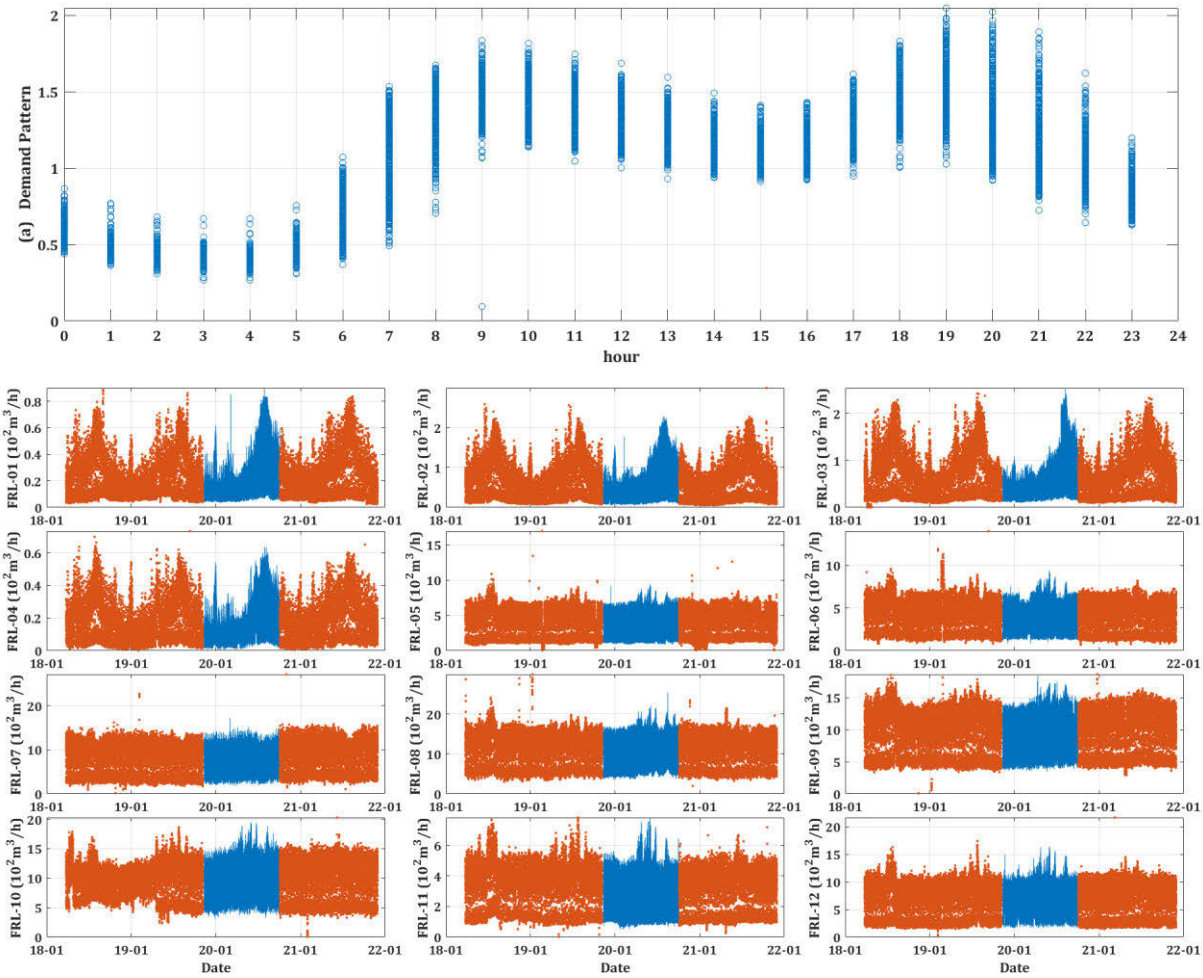


Figure 2. Water consumption province Period 2018-2021 in a North-Province of The Netherlands (a) daily water balance, (below) time series for each DMA's. Period in blue corresponds to a period of homogeneous data collection.

4 RESULTS

C-Town fully monitored

A fully monitored WDN implies installing volume meters, pressure sensors and AMR's on each location of the WDN. After training a SINDy model the results of RMSE for a full monitoring are presented in Fig. 3. Results have been ranked from higher to lower RMSE for pressure at nodes (A) and flows in links (B).

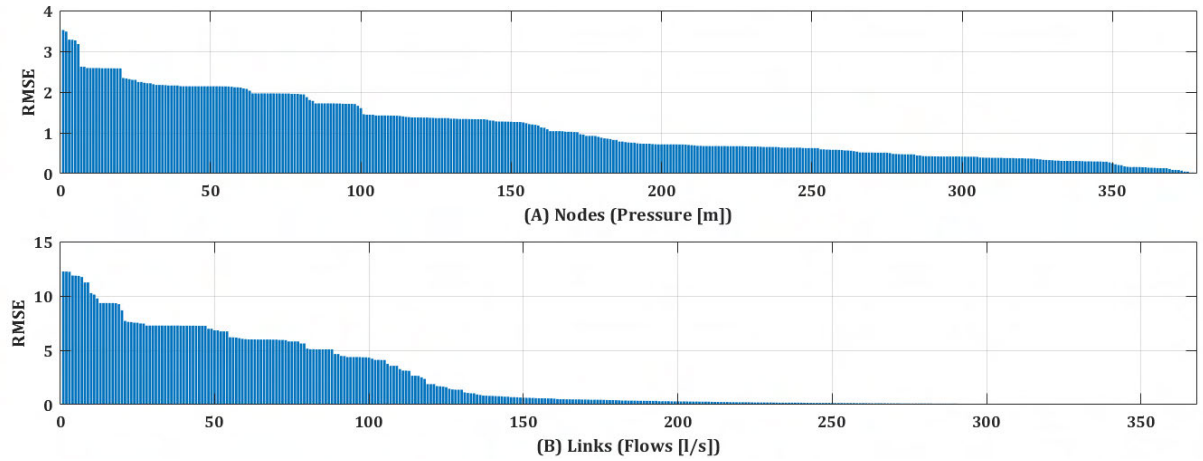


Figure 3. RMSE of all variables of C-Town. (A) Node pressures, (B) pipe flows.

This results indicate that SINDy is a suitable alternative for representation of a fully monitored WDN. To portrait the results obtained for each variable Fig. 4 presents the fitness of the time series for flow in pipe 11 (Q_{11}). This is a pipe where flows change direction throughout the simulation. The RMSE of this variable is high 11.86 l/s, however one can assess that most of the large errors occur during the change of trajectory of the flows in consecutive time stamps and the SINDy surrogate is able to return to the trend of the variable very fast. A similar behaviour of the application of SINDy has been obtained by other authors [4] for complex dynamical systems (such as Lorenz attractor). The errors are normally distributed as presented in Fig 4 (lower right).

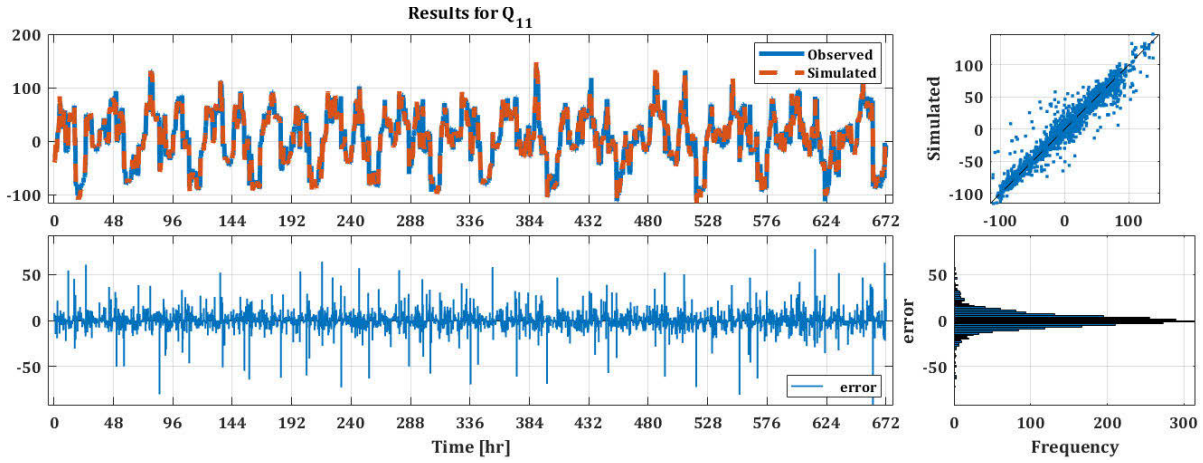


Figure 4. SINDy results for flows at pipe 11 (Q_{11}). Top left is the time series of Observed (EPANET) and Simulated (SINDy). Top right is the scatter of the time series Simulated vs Observed. Bottom left is the error obtained at each time stamp. Bottom right corresponds to the histogram of the errors obtained.

C-Town subset monitoring

On the second case, using a configuration with only 37 hydraulic variables the application of SINDy was not able to obtain similar results. The main issue on the formulation of the SINDy model is the fact that the data contains pump status as a variable. Such data was not used in the first case of analysis of the fully monitored network C-Town. Apparently, the inclusion of binary variables as independent variables in a SINDy model tends to create an overshoot in the behaviour of simulated dependent variables. Such behaviour is presented in Figure 5, where the time series shows that the status can have only values ϵ [0 or 1], while the estimation shows that the outcomes are real values in the range ϵ [-0.75, 1.30]. At this moment it is not known by the author whether or not there is a mechanism to handle binary variables within SINDy on the estimation of surrogates for WDN.

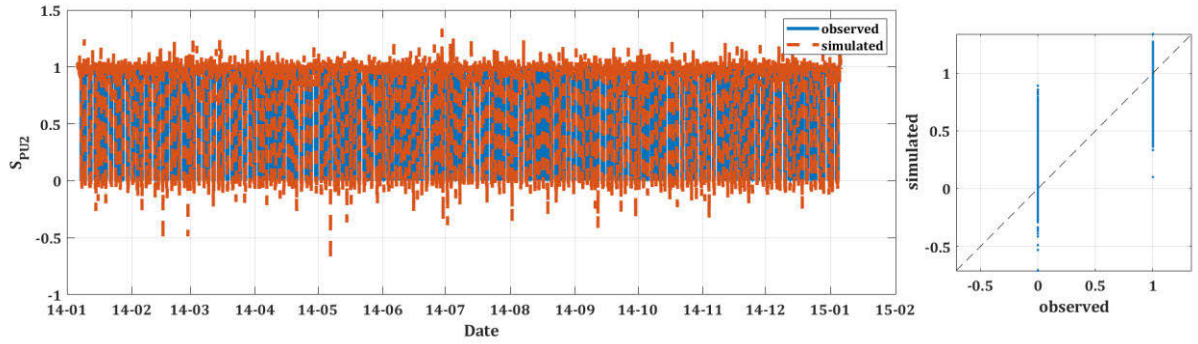


Figure 5. Results of time series for status of pump 2 and the corresponding error on the estimation.

Subsequently, the SINDy model obtained with normal data is applied to the data containing the anomalies (abnormal), however the results of the surrogate make it impossible to know whether the rapid variation of the time series is due to an event or due to the construct of the surrogate model itself.

Water balance of a province

Data of the water consumption is divided by 100 to obtain rescaled values. Subsequently a SINDy model is built. Lambda (λ) is used for sparsification and set as [0.01, 0.1, 1.0]. The maximum polynomial order is set to 2. The total number of variables considered in each case for Θ(X) is equal to 91 (i.e. order 0: 1; order 1: 12; and order 2: 78). Results of the RMSE for each value of λ are presented in Figure 6 where on the left are the sparse elements of $\Theta(X)^T$. Each row represents the coefficients which are active in the SINDy formulation for each DMA. Here *nnz* is equal to the total number of non-zero elements in each case. A higher value of λ will reduce the number of coefficients in Θ(X) which are non-zero from 954 to 131. On the right, the corresponding RMSE obtained after estimation of the consumption in each DMA. It needs to be mentioned that for values of $\lambda > 1.0$, the RMSEs increase continuously, while on the case that $\lambda < 0.01$ the additional gains on the error reduction are imperceptible. In addition, it is noticeable that the RMSEs are not linearly dependent on λ. This may lead to potentially select a different λ for the determination of the best surrogate of each area. Given the average consumption per area, it is expected that different areas with larger consumption will present larger errors.

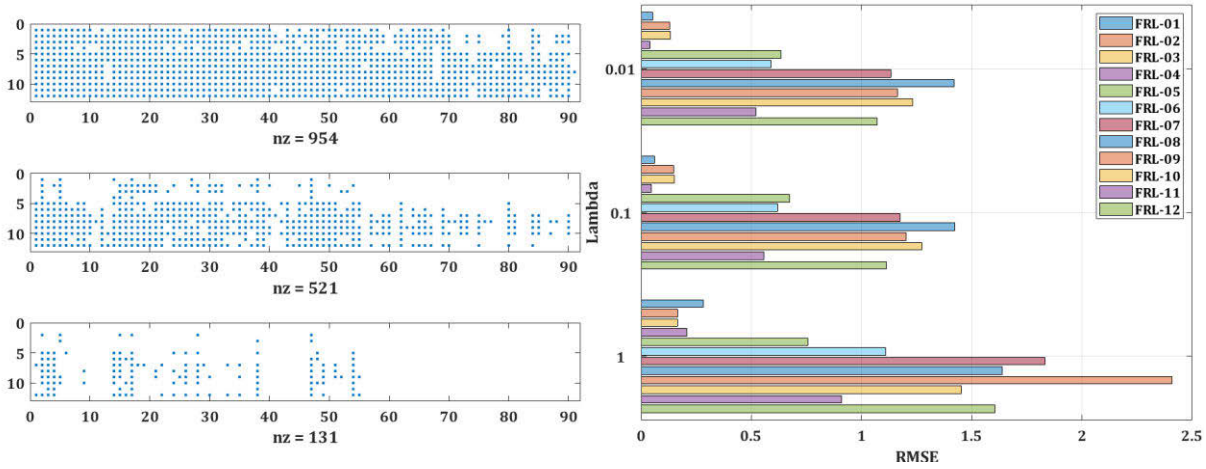


Figure 6. Left, matrix $\Theta(X)$ transposed. Right, the corresponding RMSE for each DMA.

Additional work is required to determine the minimum length of data required to generate comparable results. At this point 7,840 timestamps correspond to almost 1 year of continuous hourly data registration, this may not be possible for most utilities.

5 CONCLUSIONS

This article presents a new surrogate model for WDN. The method has been applied for both data of a benchmark WDN and for data of water consumption in a large province. In both cases SINDy was able to recreate the behaviour of the underlying system with low computational cost. The application for data of a fully monitored WDN shows the potential for the development of a very easy to setup surrogate model. Its application for a subset of monitoring variables of the same benchmark network were not able to reliably generate a surrogate model. In the case of the application of SINDy for the determination of a surrogate of water consumption in DMA's, once again the results show relatively good accuracy with respect to the observed values. Larger DMA's show larger RMSE and vice versa.

In addition, the possible application of the method as an anomaly detection algorithm for leakage detection or leakage localization are yet to be explored in a real system. Other aspect to consider as future work is the determination of the minimum length of the timeseries and the resolution required to build a trustworthy surrogate model.

6 REFERENCES

[1] F. Martinez-Alzamora, P. Conejos, M. E. Castro-Gama and I. Vertommen, "Digital Twins - A new paradigm for water supply and distribution networks," Hydrolink, pp. 48-54, 2021.

[2] Z. Rao and F. Alvarruiz, "Use of an artificial neural network to capture the domain knowledge of a conventional hydraulic simulation model," Journal of Hydroinformatics, vol. 9, p. 15–24, January 2007.

[3] D. R. Broad, G. C. Dandy and H. R. Maier, "Water Distribution System Optimization Using Metamodels," Journal of Water Resources Planning and Management, vol. 131, p. 172–180, May 2005.

[4] F. Martínez, V. Hernández, J. M. Alonso, Z. Rao and S. Alvisi, "Optimizing the operation of the Valencia water-distribution network," Journal of Hydroinformatics, vol. 9, p. 65–78, January 2007.

[5] E. Salomons, A. Goryashko, U. Shamir, Z. Rao and S. Alvisi, "Optimizing the operation of the Haifa-A water-distribution network," Journal of Hydroinformatics, vol. 9, p. 51–64, January 2007.

[6] Z. Y. Wu and A. Rahman, "Optimized Deep Learning Framework for Water Distribution Data-Driven Modeling," Procedia Engineering, vol. 186, p. 261–268, 2017.

[7] N. Kutz, S. Brunton, J. Proctor and B. Brunton, Dynamic mode decomposition : data-driven modeling of complex systems, Philadelphia: Society for Industrial and Applied Mathematics, 2016.

[8] Z. Wu, S. L. Brunton and S. Revzen, "Challenges in Dynamic Mode Decomposition," September 2021.

[9] S. Brunton and N. Kutz, Data-driven science and engineering : machine learning, dynamical systems, and control, Cambridge, United Kingdom New York, NY: Cambridge University Press, 2019.

[10] S. L. Brunton, J. L. Proctor and J. N. Kutz, "Discovering governing equations from data by sparse identification of nonlinear dynamical systems," Proceedings of the National Academy of Sciences, vol. 113, p. 3932–3937, March 2016.

[11] E. Todini and L. A. Rossman, "Unified Framework for Deriving Simultaneous Equation Algorithms for Water Distribution Networks," Journal of Hydraulic Engineering, vol. 139, p. 511–526, May 2013.

[12] E. Todini and S. Pilati, A Gradient Algorithm for the Analysis of Pipe Networks in: "Computer Applications in Water Supply. 1 - System Analysis and Simulation", B. Coulbeck and C. Orr, Eds., London, UK: Research Studies Press Ltd., 1988.

[13] B. Coulbeck, Computer applications in water supply, Letchworth, Hertfordshire, England New York: Research Studies Press Wiley, 1988.