Modeling Atmospheric Data and Identifying Dynamics: Temporal Data-Driven Modeling of Air Pollutants

Preprint · October 2020

CITATIONS
0

READS
78

3 authors, including:

Javier Rubio-Herrero
University of North Texas
15 PUBLICATIONS **87** CITATIONS
SEE PROFILE

Wai-Tong Louis Fan
Indiana University Bloomington
39 PUBLICATIONS **189** CITATIONS
SEE PROFILE

Modeling Atmospheric Data and Identifying Dynamics
Temporal Data-Driven Modeling of Air Pollutants

Javier Rubio-Herrero[a,*], Carlos Ortiz Marrero[b], Wai-Tong Louis Fan[c,d]

[a]*Department of Information Technology and Decision Sciences, G. Brint Ryan College of Business, University of North Texas, 1155 Union Circle, Denton, TX 76201*
[b]*Data Sciences and Analytics, Pacific Northwest National Laboratory, 902 Battelle Blvd, Richland, WA 99354*
[c]*Department of Mathematics, Indiana University, 831 East 3rd St. Bloomington, IN 47405*
[d]*Center of Mathematical Sciences and Applications, Harvard University, Cambridge, MA 02138*

Abstract

Atmospheric modelling has recently experienced a surge with the advent of deep learning. Most of these models, however, predict concentrations of pollutants following a data-driven approach in which the physical laws that govern their behaviors and relationships remain hidden. With the aid of real-world air quality data collected hourly in different stations throughout Madrid, we present a case study using a series of data-driven techniques with the following goals: (1) Find systems of ordinary differential equations that model the concentration of pollutants and their changes over time; (2) assess the performance and limitations of our models using stability analysis; (3) reconstruct the time series of chemical pollutants not measured in certain stations using delay coordinate embedding results.

Keywords: atmospheric modeling, nonlinear dynamics, sparse regression, delay embedding, stability analysis

1. Introduction

Chemists and environmental scientists refer to nitrogen oxides as the group of compounds that contain nitrogen and oxygen. The most important of those gases, the nitric oxide (NO) and

*Corresponding author
Email addresses: javier.rubioherrero@unt.edu (Javier Rubio-Herrero), carlos.ortizmarrero@pnnl.gov (Carlos Ortiz Marrero), lfan@cmsa.fas.harvard.edu (Wai-Tong Louis Fan)

Preprint submitted to XXXX

the nitrogen dioxide (NO_2), are of special interest because they are byproducts of many human activities. In 2013, 76.4% of the tropospheric NO_x had an anthropogenic source. Of that total, 75.5% originated from fossil fuel combustion and industrial processes [31].

The relevance of these two gases in how ozone (O_3) is distributed on earth makes the control of their emissions of paramount importance from a policy, environment, and health perspective. The chemical reactions that lead to the creation of ozone and the connection of this phenomenon with urban pollution have been largely studied [20, 21, 58]: Nitrogen dioxide reacts with oxygen and ultraviolet (UV) light to form ozone and nitric oxide. This is in fact the beginning of a looping process, as the NO reacts with free radicals that result from the interaction of UV light and volatile organic compounds (VOCs) in the atmosphere and produces NO_2 again, thus continuing with the creation of O_3 [13].

Ozone is critical in the stratosphere, as it protects the earth from the harmful effects of UV radiation by absorbing it. However, the creation of ozone in the troposphere poses a serious pollution problem as it is the main component of *smog* and thus becomes part of the air we breathe [13]. Ozone may cause health issues like airway inflammation and hyper-responsiveness, that can then lead to asthma [28]. In many inhabited places in the US, peak ambient ozone concentrations are sufficiently high as to provoke changes in lung function as well as respiratory symptoms in otherwise healthy people [39]. It is forecasted that it may lead to an increase in the median mortality rate in some areas of the US in the next decades [33]. In addition, the creation of tropospheric ozone is deeply affected by climate change, as the chemical reactions that produce it are very sensitive to temperature and lead to higher concentrations at higher temperatures [8]. It is also known that the creation of tropospheric O_3 is not a problem confined to urban areas, as winds may transport plumes of anthropogenic NO_2 to distant areas from their source before the creation of O_3 takes place.

The adverse effects of NO_x gases are not limited to the creation of ozone in the troposphere. In addition, these gases combine with sulfur oxides (SO_x) and moisture to produce *acid rain*. Low-pH rains have been often reported in China, Europe, and the United States [35, 27, 43], which have a prejudicial consequence on the acidity of soils, corrosion of materials, and human health.

In order to fight against the effects of pollution, many governments have implemented policies

at different levels, namely, national, regional, and local. For example, parts of California have seen a significant decrease in pollution over a period of 20 years despite "a 38% increase in regional motor vehicle activity, vigorous economic growth, and a 30% population increase" [40]. More recently, the Chinese government imposed strict measures to fight against the spread of the novel coronavirus, *COVID-19* [65]. This resulted in a cease of transportation and industrial activities in most parts of the country. A collateral effect of these strict measures was the sharp decrease in the emissions of NO_2 (see Figure 1).

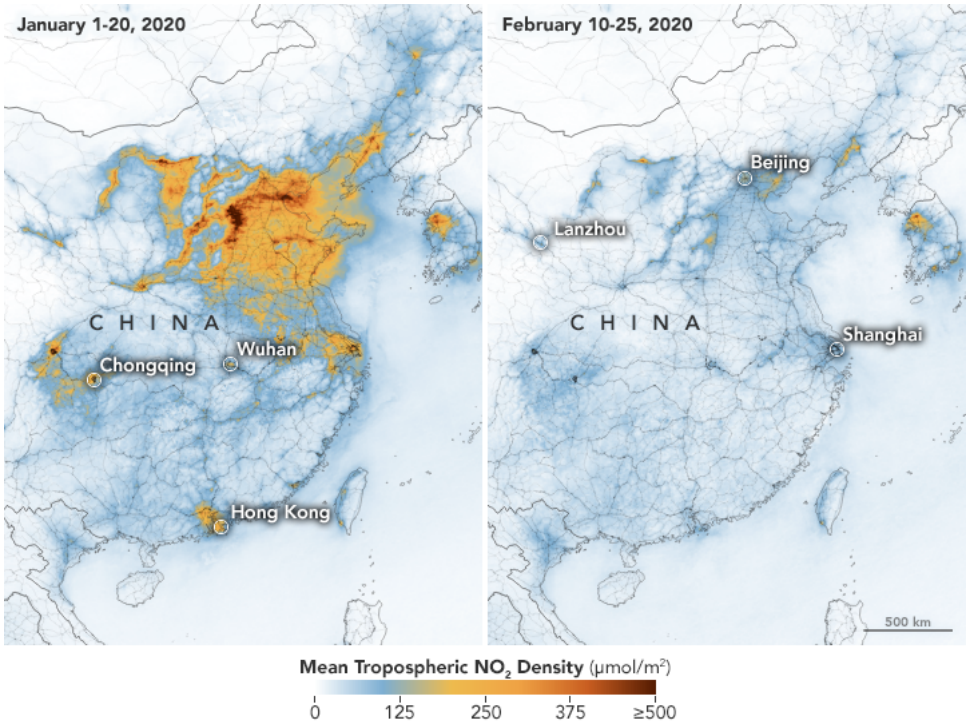


Figure 1: Drop in NO_2 emissions as a consequence of measures against COVID-19. Source: NASA Earth Observatory/ Joshua Stevens [5].

At a council level, some major capitals are introducing their own control policies. In an attempt at *pedestrianization*, many councils are moving towards transit models that hamper the use of cars in densely populated urban areas. In general, as noted in [36], "this kind of measures have not only environmental health impacts. They positively affect tourism development, job creation, improving safety, enhancing the appearance of urban areas." In 2018, Spain's capital, Madrid, designated parts of its downtown as *low-emission zones*, known as *Madrid Central* (MC) [3]. This city's

council limited the access of certain vehicles, albeit the transit of residents' vehicles remained permitted. Recent research shows that these limitations have decreased the concentration of NO_2 in downtown [36]. However, Madrid's new government is looking to reverse the implementation of MC in favor of less restrictive, yet geographically wider transit strategies, called *Madrid 360* [2].

Predicting the concentration of pollutants in the atmosphere is a well-studied topic [17, 22]. In [38] the authors rightfully indicate that there are two types of modeling efforts when it comes to forecasting the concentration of pollutants in the atmosphere. On the one hand, researchers can tackle this problem under a deterministic approach, in which the physics of the model come into play in the form of diffusion equations, the pollutants' chemical characteristics, or fluid dynamics. These models usually require the solution of highly-nonlinear mathematical relationships typically expressed in the form of partial differential equations [49, 64, 34]. Other physics-based models compute air parcels and track (or backtrack) the dispersion of atmospheric pollutants [59].

On the other hand, purely data-driven, statistical approaches bypass the physics that underlie the complicated behavior of these pollutants to find correlations in the data. The last few years have seen a surge of methods based on neural networks that can tackle some of these problems very effectively. Recurrent neural networks (RNNs) in general, and *Long Short-Term Memory* (LSTMs) in particular, have been explored profusely as a means to explain the evolution of unknown variables over time. While these types of neural networks have been studied for quite some time, the recent developments in hardware capabilities and the larger availability of data have put them again in the forefront of the machine learning scenery. Nonetheless, training neural networks in most practical purposes involve tuning a vast number of parameters. As a result of this flexibility, researchers may be able to forecast accurately the variable of interest over time, but they fail to extract an equation that can explain its behavior as a function of the predictors or the independent variables. This is why neural networks are classified as *black box models*. For instance, [52] used meteorological and chemical variables to forecast *maximum* daily ozone levels in Houston. The authors compared a regression model, a Box-Jenkins model, and fully-connected neural network with sigmoid activation functions and concluded that the neural network performed better. This problem was in turn tackled previously by [56] with the use of univariate deterministic/stochastic models, ARIMA models, and bivariate models. More recently, other black-box-based models were

used to forecast future concentrations of ozone and other pollutants in European cities [18, 44]. LSTMs were used by [50] to predict the levels of NO and NO_2 with a lags of 8, 16, and 24 hours. In addition, the authors used meteorological information as features to produce a time-series that aimed at helping with policy-making when it came to traffic restrictions. Their results were superior to those reported by *CALIOPE* [10], a simulation-based air quality forecasting system in use for temporal and spatial prediction of gas phase species and particulate matter (i.e. NO_2, SO_2, O_3 and PM_{10}) in Spain.

Unfortunately, these LSTM models require lots of training examples to attain accurate results and do not extrapolate well outside the regime in which they where trained [12]. Most importantly, these black box models produce uninterpretable relationships between the variables. Conversely, modeling by regression presents the opposite advantages and disadvantages: it forces the selection of a model or a family of models prior to calibrating the regression coefficients, thus reducing the modeling flexibility. However, it provides a clear, closed-form mathematical expression that relates the dependent variable and the set of independent variables.

To address some of the challenges outlined above, some researchers have started to look at ways to integrate the data-driven efforts with domain modeling. Most of the work is beginning to coalesce under the banner of "Scientific Machine Learning" [9] and promising approaches are continually being developed and improved [54, 14, 16, 53]. Inspired by these recent successes in the field, the goal of this work is to outline a data-driven, but domain-aware, pipeline to model atmospheric pollutants. Such a pipeline bridges the gap between these two classical perspectives and provides methods that can leverage data as well as produce relationships between atmospheric chemical species that capture the physics of the system being modeled. In this paper, we apply these techniques to find a system of ordinary differential equations (ODEs) from real-world measurments of NO_2 and O_3.

As previously discussed, the use of these techniques have clear applications in policy-making environments at the national, regional, and local levels where accurate quantitative tools are of vital importance to assess atmospheric contamination. In turn, these tools can help policy makers determine the benefits and impact of their pollution control techniques [51]. The structure of this paper is as follows:

1. In Section 2 we propose an alternative optimization approach for sparse identification of nonlinear dynamics (SINDy) [14]. We will apply this approach to real time-series data collected in various air quality stations located in Madrid and we will find systems of ordinary differential equations (ODEs) that will capture the dynamics of ozone and nitrogen dioxide at those geographical spots for a given time frame. We will discuss the implications of noisy datasets in this context and how that impacts the performance of SINDy.
2. In Section 3, we analyze some basic mathematical properties of the ODEs reconstructed from the data in Section 2 and offer some insight regarding the global behavior the dynamics of the concentrations of NO_2 and O_3. For each air quality station that measures both chemical species, we classify the critical points of the system of ODEs according to a stability analysis. This analysis provided us with a way to interpret the performance and limitations of our fitted equations.
3. In Section 4 we reconstruct the time series of the concentrations of O_3 at those stations were only NO_2 readings are available. To this end, we rely on Takens'*embedding* theorem to perform the reconstruction.

2. SINDy for Atmospheric Data

2.1. Motivation

The application in hand plays a crucial role in the dichotomy between flexibility and tractability posed by black-box models and regression models. Systems dynamics, for example, usually present a behavior that can be explained with low-order mathematical expressions. As a matter of fact, the dynamics of many physical and chemical systems can be modeled as a system of ODEs. One of those systems is *atmospheric chemistry*, which involves all the chemical reactions and interactions that take place on the Earth's atmosphere.

Chemical reactions are governed by their *rate equations*. The *kinetics* of a chemical reaction show how the concentrations of the reactants and the products vary during the reaction. Rate equations are easy to obtain from elementary reactions in closed systems (i.e., in systems where only those reactions occur and there is not an a flux of other molecules entering or exiting those

systems). For instance, *mass-action* kinetics suggest that the reaction

$$aA + bB \rightarrow cC + dD$$

consumes reactant A at a rate of $mole/(m^3 \cdot s)$ given by the differential equation

$$\frac{dA}{dt} = -k[A]^a[B]^b,$$

where k is the rate constant of the reaction [25]. In turn, the exponents of the concentrations, a and b, called the *partial orders* of the reaction are in this case the stoichiometric coefficients of the chemical reaction. Their sum is referred to as the *overall order* of the reaction.

Eventually, in a closed system, the reactants are exhausted. In open systems with influx and efflux of several chemical species, like the troposphere, where many chemical reactions occur simultaneously, molecules are constantly created and destroyed, either created or consumed as a result of many reactions. This phenomenon yields a steady-state in the concentrations that stems from a dynamic equilibrium [24]. Also, in open systems the reactions occur in multiple steps and the partial orders of rate equations usually do not match the stoichiometric coefficients. Moreover, since a reactant can be part of several reactions at the same time (i.e. can be consumed or produced as a result of other reactions), the time-series of its concentration is not necessarily given by the the rate equation of a single reaction.

As an example, consider the *Leighton cycle* [37] that explains the formation of ozone from nitrogen oxides in the troposhere in unpolluted conditions:

$$NO_2 + h\nu \xrightarrow{J_t} NO + O(^3P), \tag{1}$$

$$O(^3P) + O_2 \rightarrow O_3, \tag{2}$$

$$NO + O_3 \xrightarrow{k} NO_2 + O_2, \tag{3}$$

where $h\nu$ represents energy from solar radiation (as calculated by the product of Planck's constant, h, and the frequency of the wave of solar radiation, ν) and $O(^3P)$ denotes an oxygen atom in its fundamental state. The reaction rate J_t represents the *actinic flux*, which varies over time and depends largely on the incidence of photons, thus presenting different values according to other

factors such as cloud cover or season. The reactions (1)-(3) present the following kinetics [42]:

$$\frac{d[NO_2]}{dt} = -J_t[NO_2] + k[NO][O_3], \tag{4}$$

$$\frac{d[O_3]}{dt} = \frac{d[NO]}{dt} = J_t[NO_2] - k[NO][O_3]. \tag{5}$$

In unpolluted conditions, the ozone present in the troposphere is due to transport from the stratosphere and photochemical production. Its destruction is also due to photochemical reactions and from deposition on the earth's surface. These processes happen at a rate that maintains the level of ozone approximately constant in this layer of the atmosphere and, in these circumstances, the above represents a *null cycle* in which there is not any net production nor destruction of these chemical species. Therefore, their kinetics reach a pseudo-steady-state that can be expressed as

$$\frac{J_t}{k} = \frac{[NO][O_3]}{[NO_2]}.$$

This relationship explains why during daylight hours, when the actinic influx is large, there is an increment of the concentrations of NO and O_3 at the expense of a destruction of NO_2 [19]. It also explains why this trend is reversed during the night hours. During this series of reactions, some OH radicals from photolysis at wavelengths $\lambda < 319\ nm$ are created [23]. In *polluted* conditions, these radicals react with methane (CH_4), carbon monoxide (CO) and man-produced VOCs, leading to the creation of more ozone, amongst other molecules. Those new reactions complicate the variation of NO_2 and O_3 more than expressed in equations (4) and (5) and the system as a whole does not represent a null cycle anymore. A more complete (although not exempt from simplifications) dynamic model that illustrates this extended complexity is presented in [42].

In short, we conclude that the dynamics of ozone and nitrogen oxides in polluted environments result in a convoluted and intertwined series of chemical reactions. **We also justify that they can be largely explained with low-order polynomials**. Hence, the case above exemplifies an interesting research question in how we can **develop a data-driven methods that are able to capture the dynamics of a complex, atmospheric open system in a closed mathematical form** by identifying the values of the coefficients of the rate equations corresponding to some of the reactions that occur. Therefore, a data-driven method that allows to explain the changes in concentrations of

certain pollutants over time in such complicated dynamic systems turns out to be of higher importance, as it would be capable of finding relationships between chemical species that are "distant", in the sense that their concentrations are not directly related by a particular chemical reaction, but rather by a series of them. Examples of such distant species are nitrogen dioxide and ozone, to which we turn our attention to in this paper.

As mentioned in Section 1, there have been attempts to predict the concentration of ozone in the atmosphere. Our approach differs from all these in that we propose a regression approach that can capture the dynamics of the atmosphere in a way that different chemical species and their concentrations over time are interrelated, thus offering a closed form of the rate equations that govern the chemical reactions occurring during the selected time frame. In addition, in Sections 3 and 4 we will use those governing equations to offer analytical insight of the dynamics of these species and to reconstruct the time series of ozone in those stations that currently do not measure it.

2.2. Mathematical representation of the system dynamics

Our goal is to find a series of ODEs that describe the chemical dynamics in the troposphere. That is, a system of the form,

$$\dot{\mathbf{y}}(t) = \mathbf{F}(\mathbf{y}(t)), \quad i = 1, 2, \ldots, p, \tag{6}$$

where $\mathbf{y}(t) = (y_1(t), y_2(t), \ldots, y_p(t))^T \in \mathbb{R}^p$ is a vector containing the time response of the concentrations of the p chemical species under study, $\dot{\mathbf{y}}(t) = (\dot{y}_1(t), \dot{y}_2(t), \ldots, \dot{y}_p(t))^T$ is the vector of derivatives, and $\mathbf{F}(\mathbf{y}(t))$ is a vector field acting on $\mathbf{y}(t)$. In [14] the authors outlined a methodology to estimate the functional form $\mathbf{F}$ given samples from $\mathbf{y}$ and $\dot{\mathbf{y}}$ by solving the a series of least squares problems,

$$\min_{\boldsymbol{\beta_i}} ||\dot{\tilde{\mathbf{y}}}_\mathbf{i} - \tilde{\mathbf{F}}\boldsymbol{\beta_i}||_2^2, \quad i = 1, 2, \ldots, p. \tag{7}$$

where $\dot{\tilde{\mathbf{y}}}_\mathbf{i} = (\dot{y}_i(t_1), \dot{y}_i(t_2), \ldots, \dot{y}_i(t_m))$ and $\boldsymbol{\beta_i} \in \mathbb{R}^{n+1}$. The matrix $\tilde{\mathbf{F}} \in \mathbb{R}^{m\times(n+1)}$ contains information about candidate nonlinear basis functions (plus the intercept term) over a time horizon $m \gg n$. For example, if we have a system of two chemical species ($p = 2$) and we believe that the entries of

$\mathbf{F}(\mathbf{y}(t))$ are composed of second-order polynomials over $y_1(t)$ and $y_2(t)$ (i.e. $n = 5$), then

$$\tilde{\mathbf{F}} = \begin{bmatrix} 1 & y_1(t_1) & y_2(t_1) & y_1^2(t_1) & y_2^2(t_1) & y_1(t_1)y_2(t_1) \\ 1 & y_1(t_2) & y_2(t_2) & y_1^2(t_2) & y_2^2(t_2) & y_1(t_2)y_2(t_2) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & y_1(t_m) & y_2(t_m) & y_1^2(t_m) & y_2^2(t_m) & y_1(t_m)y_2(t_m) \end{bmatrix}.$$

Notice a solution to the minimization problem (7) yields an approximate solution to $\mathbf{F}$ in (6). Also note that when data from $\dot{\tilde{\mathbf{y}}}_\mathbf{i}$ is not available, we can approximate this quantity using the data vector $\tilde{\mathbf{y}}_\mathbf{i} = (y_i(t_1), y_i(t_2), \ldots, y_i(t_m))$.

In most occasions not all the terms in the chosen linear functional form are needed, as it is very likely that some are not relevant for explaining the dependent variable in question. Therefore, a frequent subproblem within regression is that of finding the subset of terms that provides the best representation of the independent variable. This search for a more parsimonious or sparse mathematical expression is the core idea explored in [14]. We analyzed the advantages and drawbacks of *LASSO* regression and *best subset* regression and concluded that for our application the latter was a better choice to obtain accurate representations of the dynamics of NO_2 and O_3 at different geographical points of Madrid. A description of both methods is detailed in Appendix A.

2.3. Method selected

Despite the fact that LASSO regression has been around for more than two decades, its successful application to *sparse identification of nonlinear dynamics* (SINDy) is rather new. Since the seminal paper in this subject [14], there have been multiple efforts to find sparse representations of physical and biological systems as well as population dynamics [14, 41, 32], amongst others. The tractability considerations discussed in the appendix suggested that LASSO regression was a sensible option for identifying the dynamics of chemical species in the troposphere. However, the aforementioned works found sparse identifications of dynamic systems from data that were previously generated. During the course of our present research it was our experience that LASSO does not perform well in the context of SINDy when dealing with real-world data from multiple geological locations, even after the stage of data preprocessing. We attribute this to the considerable amount of noise present in the data and the open nature of the system we are trying to

model. While LASSO certainly could find sparse systems of ODEs, trying to solve numerically the systems of ODEs denoted by (7) was very problematic because of numerical issues caused by singularities.

Given that, as mentioned in Subsection 2.1, the dynamics of the atmosphere can be explained with low-order polynomials, we conjectured that second-order polynomials would suffice to identify the dynamics of this system. When dealing with two chemical species, NO_2 and O_3 (i.e., $p = 2$), such polynomials have at most 5 variables and there are $2^5 = 32$ possible regressions for each equation in (7). In these circumstances, brute-force enumeration of all the possible regressions is computationally affordable and therefore we opted for a best subset regression approach in which for each chemical species i, the vector of optimal regression coefficients $\hat{\boldsymbol{\beta}}_{\boldsymbol{i}}$ was selected according to the *Akaike information criterion* (AIC) [6]. We thus solved the problem

$$\begin{aligned} BS(i): \ & \underset{\boldsymbol{\beta}_i}{\text{minimize}} \quad m \log\left(\frac{\|\dot{\tilde{\mathbf{y}}}_{\mathbf{i}} - \tilde{\mathbf{F}}\boldsymbol{\beta}_i\|_2^2}{m}\right) + 2\|\boldsymbol{\beta}_i\|_0 \\ & \text{subject to} \\ & \|\boldsymbol{\beta}_i\|_0 \leq n, \end{aligned} \tag{8}$$

where the ℓ_0-norm denotes the number of non-zero elements of the vector $\boldsymbol{\beta}_i$ (excluding the intercept). Selecting the AIC was motivated by the need of an equilibrium in the bias-variance tradeoff. This criterion seeks that equilibrium by taking into account both the sum of the square of the errors (hence tackling underfitting) and the number of variables in the regression (hence tackling overfitting). This was useful to compare all the possible models and allowed us to implement an algorithm (see Appendix B) that could bypass the numerical issues encountered when using the ODE solvers. Our algorithm was programmed with *MATLAB R2020a* and was built on the structure of the code developed in [14]. This code was adapted to solve the systems of ODEs with the output of the regressions obtained from the best subset method. MATLAB has different ODE solvers that can be used in various environments. In our particular case, we found that many of the systems of ODEs produced were stiff, and thus we used MATLAB's stiff solver *ode15s*. This improved the integrability of the systems of ODEs, although still presented numerical problems in some occasions. These problems stemmed from singularities or regions were the derivative changed very rapidly. To address this problem, we introduced in our algorithm an upper bound

for the values of the derivative in such a way that those regressions that eventually led to those ill-posed solutions would be discarded.

In the absence of numerical problems, our algorithm returns the optimal solution to Problem (8) for each $i = 1, 2, \ldots, p$ and provides representations of (7) that minimize the AIC. In the presence of numerical problems, it finds representations of (7) that minimize the AIC while also being numerically tractable.

2.4. Data

The data used in this manuscript were collected from *Kaggle*, an online community for data scientists [1]. This data set in particular was extracted from *Madrid's City Council Open Data website* [4] and contains hourly readings between 2001 and 2018 of various pollutants in 24 different stations located in across areas of downtown Madrid, as well as its outskirts (see Figure 2 and Table 1). Amongst these pollutants, there are readings for NO, NO_2, NO_x, and O_3 in $\mu g/m^3$. All stations have full readings of NO, NO_2, and NO_x for all hours. Only 14 stations (marked in bold in Table 1) have full readings O_3 for all hours; the remaining 10 stations did not capture readings for O_3. For this reason, our regressions were conducted for only the 14 stations for which both NO_2 and O_3 data were available. In Section 4, we will discuss how we used our results in these stations to reconstruct the time series of O_3 readings in the other 10 stations.

As mentioned in the previous section, the collected raw data were very noisy. In the presence of such noise, it is very complicated to find systems of ODEs that that can be solved numerically without issues. Therefore some data preprocessing was needed. The following summarizes the operations performed on the set of raw data:

- **Data normalization:** The order of magnitude of the concentrations of the different chemical species in the atmosphere may differ greatly. For this reason, the time series of all molecules were standardized (i.e., for each data point of the time series we subtracted the average concentration during the time frame considered and divided over the standard deviation). It was our experience that this lead to fewer numerical errors when we integrated the systems of ODEs represented by equation (7). We will denote our original M data samples of normalized concentrations of p chemical species as $\tilde{\mathbf{w}}_\mathbf{i}, i = 1, 2, \ldots, p$.

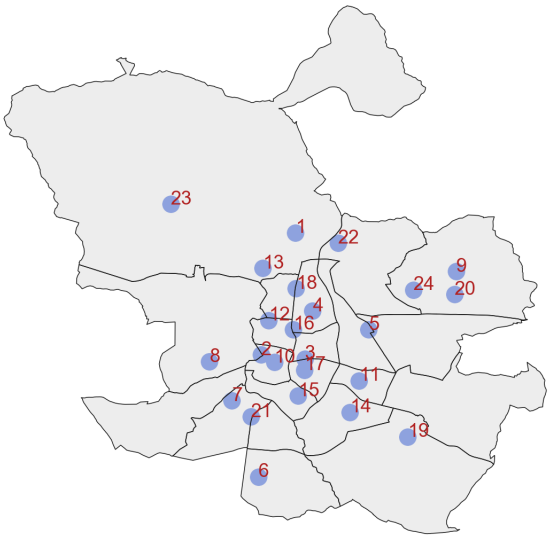


Figure 2: Stations for pollution control in Madrid

Number	Station name	Number	Station name
1	Pza. de España	13	Vallecas
2	**Escuelas Aguirre**	14	Mendez Alvaro
3	Avda. Ramón y Cajal	15	Castellana
4	**Arturo Soria**	16	**Parque del Retiro**
5	**Villaverde**	17	Plaza Castilla
6	**Farolillo**	18	**Ensanche de Vallecas**
7	**Casa de Campo**	19	Urb. Embajada
8	**Barajas Pueblo**	20	**Pza. Fernández Ladreda**
9	**Pza. del Carmen**	21	Sanchinarro
10	Moratalaz	22	**El Pardo**
11	Cuatro Caminos	23	**Juan Carlos I**
12	**Barrio del Pilar**	24	**Tres Olivos**

Table 1: List of stations for pollution control in Madrid

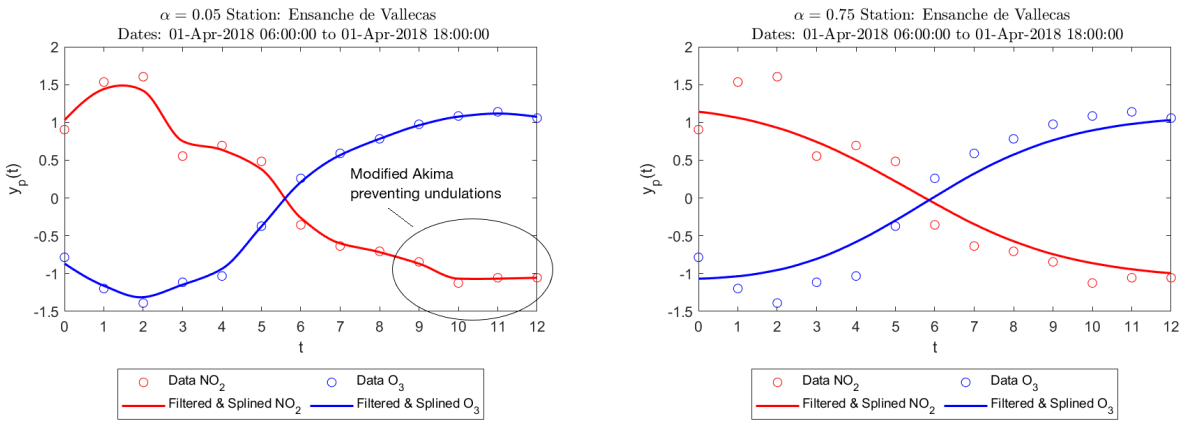
- **Data filtering:** the excess of data noise made it difficult to extract trends in the concentrations of pollutants over time. In order to address this issue we perform a Gaussian-weighted moving average filter over our data. This filter, as implemented in MATLAB, uses a window size determined heuristically that is attenuated according to a smoothing parameter $\alpha \in [0, 1]$. Values of α close to 0 reduce the window size (i.e., reduces the smoothing), whereas values of α close to 1 increase the window size (i.e., increases the smoothing). In very noisy and long time series, high values of α might be needed to extract meaningful patterns. However, this might come at the cost of excessive damping and the resulting time series might not be a good representation of the underlying data set. This parameter will be critical in our experiments, as we will discuss in subsection 2.5.

- **Data splining:** raw data were collected hourly, but the integration of a continuous-time system of ODEs requires a finer discretization of the time space. In consequence, we proceeded with the creation of 100 points between each original pair of data points, following a *modified Akima interpolation* [7]. This specific technique provides a cubic interpolation that works particularly well with sudden changes between flat regions. These changes, which usually result in excessive undulations with regular cubic splines, are very well handled with the modified Akima interpolation. Figure 3 shows an example of two normalized time series that were filtered with different values of α. Higher values of this parameter provide smoother but excessively dampened series. Lower values of α result in more realistic but also more unpredictable time series. In Figure 3a the modified Akima interpolation does a very good job in avoiding undulations in the last four data points of the NO_2 series.

After filtering and smoothing, the resulting vector of normalized concentrations will be our $\tilde{\mathbf{y}}_\mathbf{i} \in \mathbb{R}^m, i = 1, 2, \ldots, p$. Note that, by means of the splining operation, these vectors have notably more time samples than the original normalized vectors $\tilde{\mathbf{w}}_\mathbf{i}, i = 1, 2, \ldots, p$.

2.5. Optimal smoothing factor for regression

The goal of our best subset regression approach is not only to find the best system of differential equations in the sense of the AIC, but also that the solutions of those systems represent a good fit



(a) $\alpha = 0.05$

(b) $\alpha = 0.75$

Figure 3: Effect of filtering and splining on normalized data

with respect to the raw data. The regression coefficients in (7) are found by solving Problem (8) with the vectors $\tilde{\mathbf{y}}_\mathbf{i}, i = 1, 2, \ldots, p$. This means that the suitability or goodness of the fitted regressions is measured against data that have been previously manipulated. An excellent fit of an overly manipulated time series will probably not be very useful in practical terms. However, a good fit of noisy raw data seems difficult to obtain, especially under the reasonable and previously justified assumption of the suitability of second-order polynomials.

In this context, the severity of the data filtering phase is paramount. It is sensible to develop a framework in which the hyperparameter α is tuned adequately. For a given time window $[t_0, t_f]$ that contains M readings of NO_2 and O_3 in an air quality station s, let us define the *root mean square error*

$$\text{RMSE}_{i,s}^{\alpha} = \sqrt{\frac{1}{M}\left((\tilde{\mathbf{w}}_\mathbf{i})_s - \hat{\mathbf{y}}_{\mathbf{i},\mathbf{s}}(\alpha)\right)^2}. \tag{9}$$

In equation (9) the vector $(\tilde{\mathbf{w}}_\mathbf{i})_s \in \mathbb{R}^M$ contains the original normalized M readings of chemical species i in station s. The vector $\hat{\mathbf{y}}_{\mathbf{i},\mathbf{s}}(\alpha) \in \mathbb{R}^M$ contains the evaluations at those M points performed after numerically solving the system of ODEs (7) when solving Problem (8). This way, each vector $\hat{\mathbf{y}}_{\mathbf{i},\mathbf{s}}(\alpha)$ is compared to the original normalized observations. In order to find the smoothing parameter that performs best, a suitable approach is to find the solution to the following optimization

problem:

$$\min_{\alpha} \max_{1 \leq i \leq p} \quad \overline{\text{RMSE}}_i^{\alpha}, \tag{10}$$

where $\overline{\text{RMSE}}_i^{\alpha}$ is the average of $\text{RMSE}_{i,s}^{\alpha}$ over all the air quality stations. Therefore, Problem (10) aims to calibrate the smoothing parameter α such that it minimizes the value of the maximum forecasting error incurred, on average, by any chemical species.

The solution to this problem was tackled experimentally by considering v different values for α such that $0 < \alpha_1 < \alpha_2 < \cdots < \alpha_v < 1$. Figure 4 illustrates our procedure. For a given choice of α and for each station s, we solved problems $BS(i), i = 1, 2, \ldots, p$ and found a system of p (in our case $p = 2$) differential equations for each air quality station that read both NO_2 and O_3. Once this system of ODEs was integrated numerically, we extracted the resulting vector $\hat{\mathbf{y}}_{\mathbf{i,s}}(\alpha)$ of M points and a value of $\text{RMSE}_{i,s}^{\alpha}$ was obtained by (9). Then, we repeated these operations for all stations and averaged those RMSEs to obtain p different values of $\overline{\text{RMSE}}_i^{\alpha}$, whence the maximum was retrieved. After iterating over the v different values of α considered, we selected the minimum of those maximum average RMSEs as the solution to Problem (10). In our experiments we filtered our time series with $v = 21$ different values of α, from 0.05 to 0.95 in intervals of 0.05, plus 0.01 and 0.99. We did this for the 14 stations that could read both NO_2 and O_3, which were selected as our chemical species ($p = 2$).

Our results are consistent with the notion that larger time windows require a higher level of smoothing that can dampen the effect of noised data points. Consequently, it seems clear that the value of the optimal value of α in Problem (10) is nondecreasing as we enlarge our time window. The optimal value of the objective function in Problem (10) is nondecreasing as well (i.e., the minimum of the maximum average RMSE does not decrease as we use larger data samples). This is shown in Figure 5, where we compared our results for a time window centered at noon on April 1$^{\text{st}}$, 2018, with increasing window sizes between 3 hours and 19 hours.

As already discussed, the usefulness of the regression results depend on how closely they end up producing close estimations of the original data. For this reason, we can have excellent fits in the sense of the AIC criterion that are not very useful for fitting actual data because these original data have been damped excessively. As an example, consider Figure 6. Subfigures 6a and 6b show

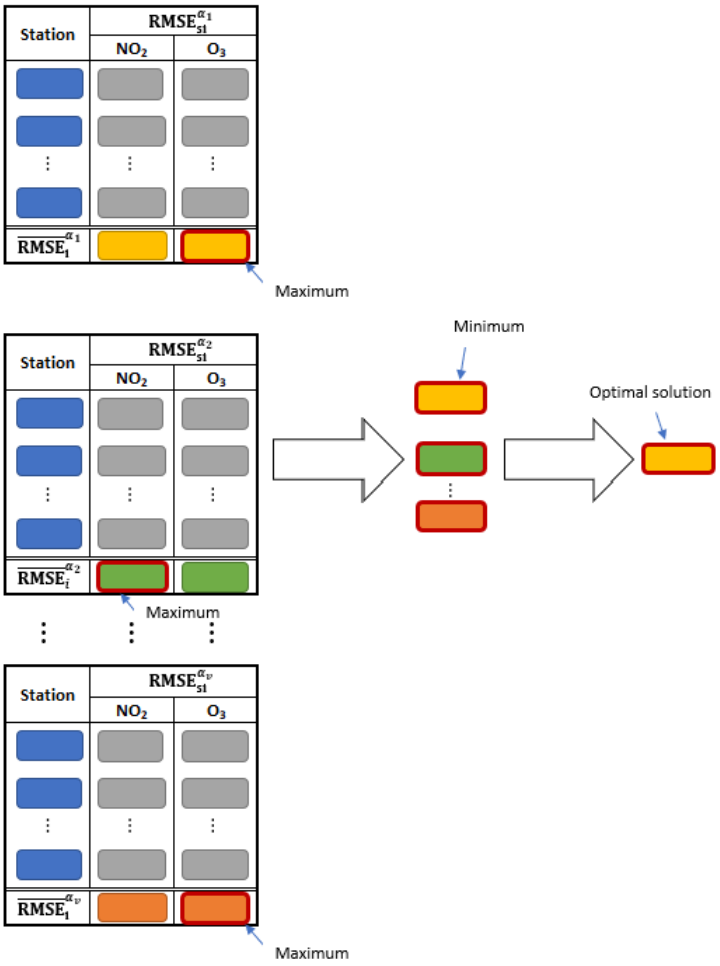


Figure 4: Process for finding the solution to Problem (10)

our regression results for two very disparate values of α (0.1 and 0.9). In Subfigure 6a the data filtered and splined is clearly wavier than in Subfigure 6b, a consequence of a softer denoising effort. These wavier curves are very close to the original noisy data points and produce derivatives that are more difficult to fit using second-order polynomials, thus yielding results that do not adjust very good to the solid lines. In Subfigure 6b the original time series has been modified much more significantly and the derivatives behave in a way that is more suitable for a second-order polynomial to be fitted. After integrating of the resulting system of ODEs, this results in dashed lines that effectively overlap the filtered and splined lines. However, our true measure of accuracy is given by the errors with respect to the original data, that is, the distance between the

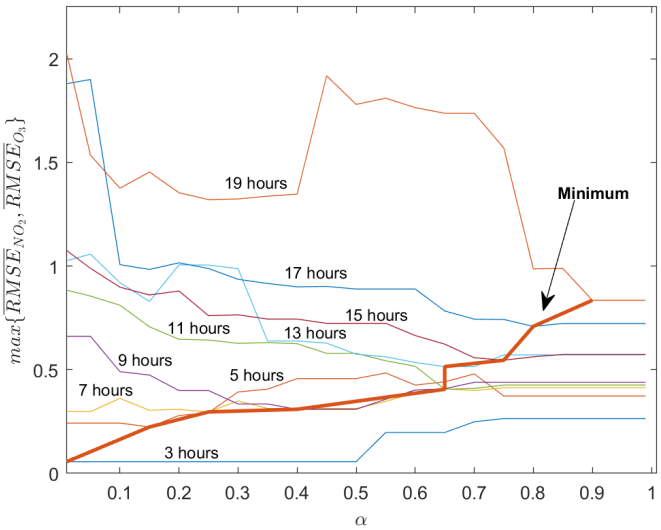
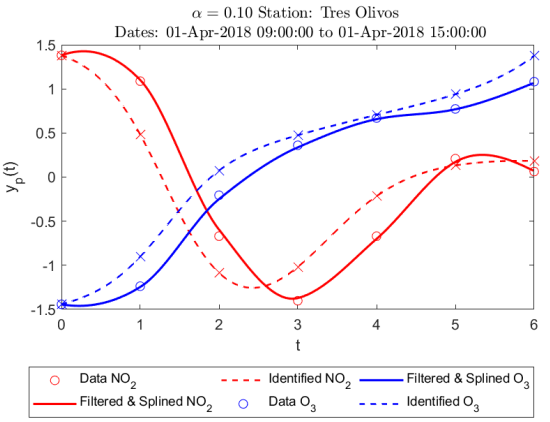


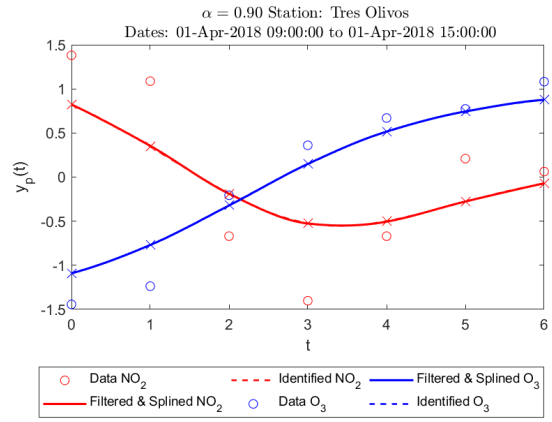
Figure 5: Worst average fit vs. smoothing factor

"×" and the "∘" and, in that sense the scenario with $\alpha = 0.1$ provides more accurate results for this station in this time window ($RMSE^{0.10}_{NO_2,Tres\ Olivos} = 0.3613$ vs $RMSE^{0.90}_{NO_2,Tres\ Olivos} = 0.5542$ and $RMSE^{0.10}_{O_3,Tres\ Olivos} = 0.2139$ vs $RMSE^{0.90}_{O_3,Tres\ Olivos} = 0.2588$). These differences are even more distinguishable in subfigures 6c and 6d, where the identified trajectories are clearly better with respect to the original data points in the case with $\alpha = 0.10$.

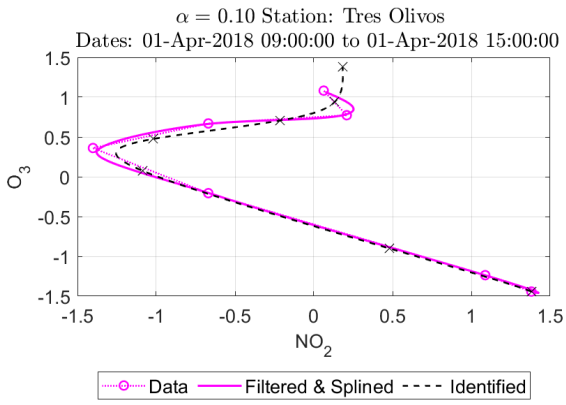
Another important result stems from the degree of sparsity found in the fitted regressions. The assumption that the time derivatives of the concentrations of chemical species in the troposphere can be modeled with second-order polynomials is, *per se*, an assumption that specifically targets a type of functionals. With two chemical species like NO_2 and O_3, this left us with at most 6 terms (5 variables and the intercept). From our results, it seemed that shorter time windows (and therefore most likely less noisy data sets) might produce sparser sets of ODEs, *ceteris paribus*. This phenomenon is shown on Figure 7. However, in most occasions the solutions to Problem (8) were full models or models without a large degree of sparsity. Therefore, we can say that while our intention was to use SINDy to further reduce the number of terms in these regressions, our results show that real-world data may not behave as well as generated data and hindered our ability to obtain sparser representations. In regard to the AIC, this means that the penalty



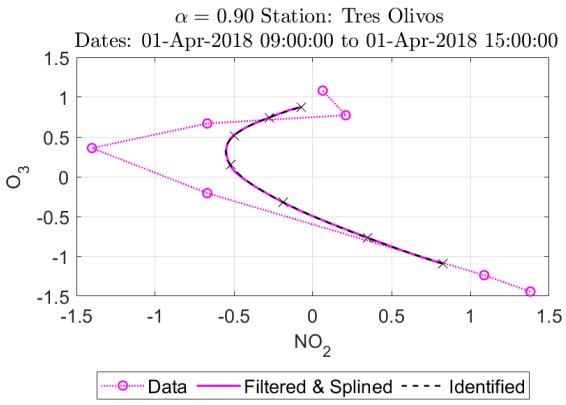
(a) Time series, $\alpha = 0.10$



(b) Time series, $\alpha = 0.90$

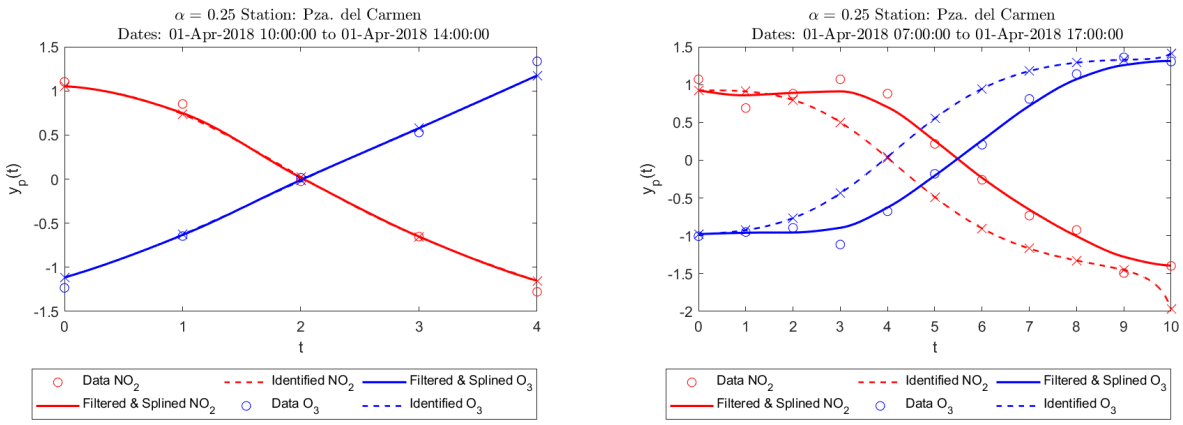


(c) State diagram, $\alpha = 0.10$



(d) State diagram, $\alpha = 0.90$

Figure 6: Comparison of results with two disparate values of the smoothing factor.



(a) 5-hour window

(b) 11-hour window

Window	$\dot{y}_i$	$\hat{\beta}_{i_0}$	$\hat{\beta}_{i_1}$	$\hat{\beta}_{i_2}$	$\hat{\beta}_{i_3}$	$\hat{\beta}_{i_4}$	$\hat{\beta}_{i_5}$
5 hours	$d[NO_2]/dt$	-0.7561	-1.2358	-1.3949	-	-1.0494	-0.6539
	$d[O_3]/dt$	-0.6142	0.6275	0.7181	-	0.2640	0.1814
11 hours	$d[NO_2]/dt$	-0.4279	-0.7495	-0.8854	-6.3679	-12.488	-5.8383
	$d[O_3]/dt$	0.4317	0.9800	1.0719	2.7402	5.2943	2.2776

(c) Fitted systems of differential equations for different time windows, $\alpha = 0.25$. General form: $\dot{y}_i = \hat{\beta}_{i_0} + \hat{\beta}_{i_1} y_{NO_2} + \hat{\beta}_{i_2} y_{O_3} + \hat{\beta}_{i_3} y_{NO_2}^2 + \hat{\beta}_{i_4} y_{NO_2} y_{O_3} + \hat{\beta}_{i_5} y_{O_3}^2$, $i = NO_2, O_3$.

Figure 7: Effect of the length of the time window on sparsity

imposed by the term $2\|\boldsymbol{\beta_i}\|_0$ is not sufficiently important to discard full models, even though in many instances we obtained very parsimonious models that performed almost as well as more complex regressions. In view of this, we also tried other criteria for selecting the best models such as the well-known $R^2_{adj} = 1 - (1 - R^2)m/(m - \|\boldsymbol{\beta_i}\|_0 - 1)$ or the *Bayesian Information Criterion* $BIC = m \log(\frac{\|\dot{\mathbf{y}}_\mathbf{i} - \tilde{\mathbf{F}}\boldsymbol{\beta_i}\|_2^2}{m}) + \log(m)(\|\boldsymbol{\beta_i}\|_0 + 1)$. We did not appreciate significant changes in the way these different methods ranked all the regressions.

3. Properties of Reconstructed ODEs

In this section, we analyse some basic mathematical properties of the ODEs obtained from the data, which offers some further analytical insight about the dynamics of $[NO_2]$ and $[O_3]$. The

general form of the equations we are fitting is

$$\frac{d[NO_2]}{dt} = \hat{\beta}_{1_0} + \hat{\beta}_{1_1}[NO_2] + \hat{\beta}_{1_2}[O_3] + \hat{\beta}_{1_3}[NO_2]^2 + \hat{\beta}_{1_4}[O_3][NO_2] + \hat{\beta}_{1_5}[O_3]^2 \tag{11}$$

$$\frac{d[O_3]}{dt} = \hat{\beta}_{2_0} + \hat{\beta}_{2_1}[NO_2] + \hat{\beta}_{2_2}[O_3] + \hat{\beta}_{2_3}[NO_2]^2 + \hat{\beta}_{2_4}[O_3][NO_2] + \hat{\beta}_{2_5}[O_3]^2, \tag{12}$$

where the coefficients $\{\hat{\beta}_{i_j}\}$ are obtained by solving the optimization Problem (8).

Such planar quadratic system can exhibit diverse behaviors and are well-studied in the mathematical literature. For example, $[NO_2]$ can blow-up in finite time if $\hat{\beta}_{1_0} > 0, \hat{\beta}_{1_3} > 0$ and all other coefficients are zero. Phase plane analysis, which characterizes topological properties of the two-dimensional trajectories of the system, reveals that this system, in general, can describe more than 700 different classes of phase portraits [55].

From the coefficients obtained in the time-window considered, $\hat{\beta}_{1_5}^2 + \hat{\beta}_{2_5}^2$ and $\hat{\beta}_{1_3}^2 + \hat{\beta}_{2_3}^2$ are both nonzero for all 14 stations. Furthermore, Theorems 2.1 and 2.2 of [55] assert that the sum of the multiplicities (called *finite multiplicity* m_f in [55]) of the critical points of (11)-(12) is 4 for all 14 stations. Below we summarize some results about real critical points[1] for all 14 stations.

The system (11)-(12) can be written in the compact form

$$\dot{y}_1 = P(y_1, y_2), \quad \dot{y}_2 = Q(y_1, y_2), \tag{13}$$

where $(y_1, y_2) = ([NO_2], [O_3])$ and P and Q are quadratic polynomials. If $P(y_1^*, y_2^*) = Q(y_1^*, y_2^*) = 0$, then (y_1^*, y_2^*) is called a critical point. Local stability of a critical point (y_1^*, y_2^*) is characterised by the eigenvalues of $J(y_1^*, y_2^*)$, where

$$J(y_1, y_2) = \begin{pmatrix} \hat{\beta}_{1_0} + 2\hat{\beta}_{1_3}y_1 + \hat{\beta}_{1_4}y_2 & \hat{\beta}_{1_2} + 2\hat{\beta}_{1_5}y_2 + \hat{\beta}_{1_4}y_1 \\ \hat{\beta}_{2_0} + 2\hat{\beta}_{2_3}y_1 + \hat{\beta}_{2_4}y_2 & \hat{\beta}_{2_2} + 2\hat{\beta}_{2_5}y_2 + \hat{\beta}_{2_4}y_1 \end{pmatrix}$$

is the Jacobian matrix; see [55, Section 2.3.1] or [45, Appendix A]. For example, suppose the eigenvalues $\{\lambda_1, \lambda_2\}$ are real and distinct (without loss of generality $\lambda_1 > \lambda_2$). Then the critical point is a stable node if $0 > \lambda_1 > \lambda_2$, an unstable node if $\lambda_1 > \lambda_2 > 0$, and a saddle point if $\lambda > 0 > \lambda_2$.

[1] A real (respectively, complex) critical point is a critical point whose coordinates are all real (respectively, complex) numbers

As mentioned in Section 2.4, the system (11)-(12) were standardized by subtracting the average concentration and dividing over the standard deviation. This normalization, being a linear transformation

$$(y_1, y_2) \mapsto (w_1, w_2) = \left(\frac{y_1 - \mu_1}{\sigma_1}, \frac{y_2 - \mu_2}{\sigma_2} \right),$$

where μ's are the means and σ's are the standard deviations, will not change the stability of the critical point. Furthermore, (w_1^*, w_2^*) is the critical point of the normalized system (11)-(12) if and only if $(\mu_1 + w_1^*\sigma_1,\ \mu_2 + w_2^*\sigma_2)$ is the critical point of the original (non-standardized) system.

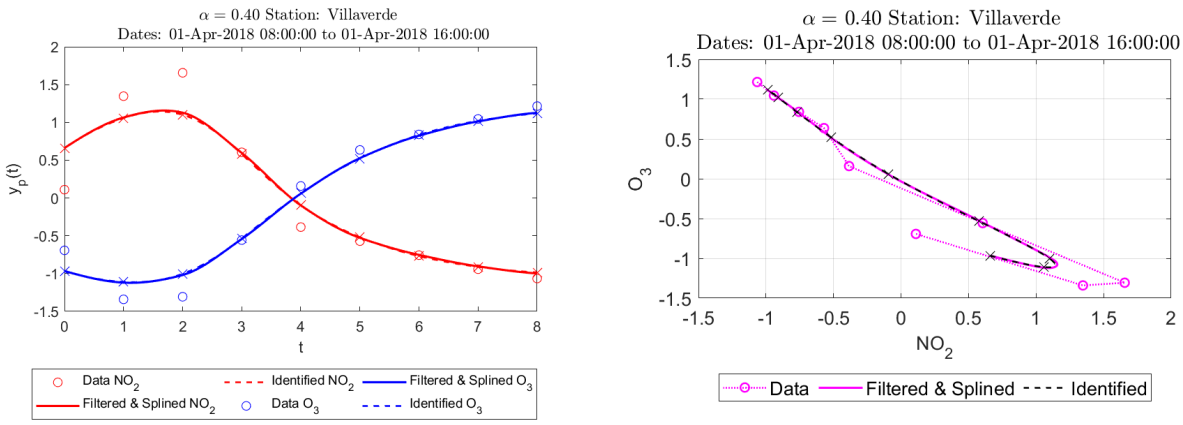
We can apply these analysis to the ODEs obtained for the 14 stations. In summary, 9 stations have 4 real critical points, 4 stations have a pair of complex critical points and two real critical points, and exactly one station has 4 complex critical points. Hence there are 44 real critical points and 12 complex critical points.

Among the 44 real critical points, 36 have positive coordinates and deserve special attention since they have physical interpretation as chemical concentrations. One critical point has negative coordinates and the remaining 7 have one positive coordinate and one negative coordinate (Arturo Soria Station is an example). Stability analysis are performed for the 36 real critical points that have positive coordinates, which reveals that more than half (20 out of 36) are saddle points (see Table 2 for a summary).

Type of critical point	**Eigenvalues** $\{\lambda_1, \lambda_2\}$	**Count**
Stable node	$0 > \lambda_1 > \lambda_2$	5
Unstable node	$\lambda_1 > \lambda_2 > 0$	4
Saddle point	$\lambda_1 > 0 > \lambda_2$	20
Stable spiral	$\{a + bi, a - bi\}$ where $a < 0$	5
Unstable spiral	$\{a + bi, a - bi\}$ where $a > 0$	2

Table 2: Classification of the 36 real critical points with positive coordinates in terms of local stability.

We now give some examples to describe the behavior of the trajectory of the solution curves of the ODEs near the critical points in the (y_1, y_2)-plane. For example, there are two critical points for Villaverde Station. The critical point $(3.1894, 91.8657)$, with standardized coordinates $(-1.0534,$



(a) Time series

(b) Two-dimensional trajectory

Figure 8: (Left panel) Time series plots for NO_2 and O_3 in Villaverde Station. (Right Panel) The two-dimensional trajectory on the right starts from bottom-right and move towards top-left, getting closer to the stable node with standardized coordinates $(-1.0534, 1.2109)$.

1.2109), is a stable node since the eigenvalues $\{-4.4819, -0.7735\}$ of its Jacobian matrices are distinct and negative. A solution starting nearby this critical point will move towards and converge to the critical point as time increases (see Figure 8b). The other critical point $(81.4092, 0.0714)$ in this station is a saddle point, since the eigenvalues $\{2.9546, -0.8807\}$ of its Jacobian matrices have opposite signs. A solution starting nearby this critical point will move towards the critical point in one direction and away from the critical point in a different direction. Standard phase portraits can be found in textbooks such as [45, Appendix A].

As a third and last example, the critical point $(19.5259, 76.2492)$ for El Pardo Station is a stable spiral since the eigenvalues $\{-0.9482 + 0.5442i, -0.9482 - 0.5442i\}$ of its Jacobian matrices are complex conjugate and have negative real parts. A solution starting nearby this critical point will spiral towards and converge to the critical point as time increases.

Our method, when applied to longer time-windows, would give a planar quadratic system with time-varying coefficients $\{\hat{\beta}_{i_j}(t)\}$. Due to their simple form and rich structure, such closed form equations are promising tools for various purposes including the study of long-time stochastic behavior [15, 47] and the enhancement to spatial-temporal models such as network models [30] and partial differential equations [49, 64, 34].

4. Reconstructing Missing Data

From our dataset, we were only able to fit 14 equations from 24 stations available. This is due to the fact that not all stations captured data for all the pollutants. As previously mentioned in Section 2.4, only 14 out of the 24 stations contained data for both NO_2 and O_3, but all 24 stations capture measurements for NO_2. Given our ability to fit the stations outlined in Section 2, we hypothesize that we can use Takens' delay embedding theorem [60] to recover the O_3 measurements from our data. In 1981, Floris Takens showed that global features of a trajectory in a dynamical system can be recovered using a single coordinate from the original data[2]. In practice the following map,

$$\hat{\mathbf{y}}(t) = \Phi_{\tau,d}(\mathbf{y}(t)) = (y_1(t), y_1(t+\tau),\ y_1(t+2\tau), \ldots,\ y_1(t+(d-1)\tau)) \tag{14}$$

yields an embedding that recovers the global properties of the original trajectory, for some lag τ and embedding dimension d. We represent y_1 as the first coordinate of the original data vector $\mathbf{y}$. If we assume that the data collected can be linked by an ODE using SINDy, as demonstrated in Section 2, we should be able to recover partial information from our missing O_3 measurements[3].

In practice, we need to estimate both τ and d, but in our case we can restrict ourselves to $d = 2$ since we are considering NO_2 and O_3. To estimate τ, we use the method of minimizing the average mutual information between the data and the first lag coordinated (or second coordinate of equation (14)) [26].

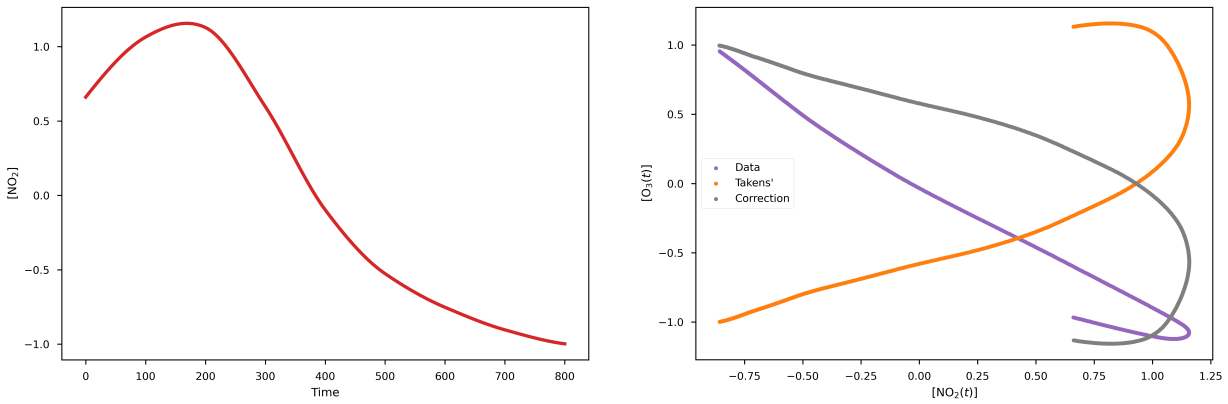
Define the average mutual information of a time series $\tilde{\mathbf{y}} = (y(t_1), y(t_2), \ldots, y(t_m))$ with lag τ by

$$AMI(\tau) = \sum_{i,j} q_{ij}(\tau) \log\left(\frac{q_{ij}(\tau)}{q_i q_j}\right),$$

where q_i is the probability that $y(t_l)$ is in bin i of the histogram constructed using samples from $\tilde{\mathbf{y}}$, and $q_{ij}(\tau)$ is the probability that $y(t_l)$ is in bin i and $y(t_l + \tau)$ is in bin j [63].

As mentioned previously, the embedding $\hat{\mathbf{y}}(t)$ will only recover some qualitative features of the trajectory, but for our problem we can exploit the fact that we only have two dimensions to recover

[2] See [57] for a detailed account of the necessary and sufficient conditions that need to be meet before the results from Takens' delay embedding theorem can be applied.



(a) Normalized $[NO_2]$ Data for Villaverde Station.

(b) Reconstructed Trajectories.

Figure 9: Our goal is to recover as much qualitative information as possible about our underlying system using a single data coordinate. (a) Normalized $[NO_2]$ data for April 01, 2018 from $8AM$ to $4PM$ at the Villaverde Station. (b) Comparison between the original trajectory (in purple), Takens' Reconstruction (in orange), and Takens' output after optimizing over $O(2)$ (in grey).

some information about the geometry of the trajectory e.g. the asymptotic behavior. Consider the following optimization problem

$$A = \underset{R \in \mathrm{O}(2)}{\arg\min} \sum_{t} |P_{y_1} R\hat{\mathbf{y}}(t) - y_1(t)|^2, \tag{15}$$

where $O(2)$ is the orthogonal group of all 2×2 matrices and P_{y_1} is the projection onto the first coordinate (our data coordinate). Exploring the local minima of this loss landscape yield matrices A (excluding the trivial solution $A = I$) that can be used to transform output obtained from the reconstruction by $A\hat{\mathbf{y}}(t)$. Figure 9 shows the results we obtained when performing this reconstruction for the Villaverde Station using NO_2 measurements to reconstruct O_3 samples (see Appendix C for more details). It is worth mentioning that the ODE SINDy recovered from the data shown in Figure 9 is near an attractor, in this case a stable point; a sufficient condition for the reconstruction procedure to apply [57].

This procedure has clear economic implications for developers since it allows for the ability to qualitatively reconstruct time-series in stations that lack the required infrastructure to recover all pollutants without the need to upgrade equipment.

5. Conclusion

We have validated and outlined a series of data-driven tools to deal with real-world atmospheric time series data. We showed how SINDy provides a framework to extract ODE models for real data collected across multiple stations distributed throughout the city of Madrid. In addition, we discussed some of the challenges we faced when extracting a parsimonious representation using LASSO and discussed how best subset regression, along with the Akaike information criterion, offered us a more stable way to consistently fit multiple stations. We also showed how performing the stability analysis of the reconstructed ODEs provides us with a way to interpret our fitted models by highlighting global features of the space of possible trajectories. Moreover, we outlined a reconstruction technique to recover unknown pollutants from correlated variables, which we believe can be useful to use at stations that are not equipped to measure all atmospheric pollutants.

We believe our methodology will provide researchers with the ability to construct interpretable, data-driven surrogate models from noisy chemical data sets. More importantly, it provides a more complete picture of the behavior of real world atmospheric chemical species (NO_2 and O_3 in our case, although our methodology can be extended to any others). We hope that the results obtain from the wide adoption of these tools allow pertinent authorities and policy makers make more informed decisions when designing future environmental policies.

Acknowledgment

The authors want to thank Michael S. Hughes and Paul Bruillard for helpful discussion on Takens' theorem and the theory of dynamical systems. W.T.Fan gratefully acknowledge the support of NSF grant DMS-1804492 and ONR grant TCRI N00014-19-S-B001. Support for C. Ortiz Marrero was provided by the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multi-program national laboratory operated by Battelle for the U.S. Department of Energy, Release No. PNNL-SA-157007.

References

1. Air quality in Madrid (2001-2018). https://www.kaggle.com/decide-soluciones/air-quality-madrid. Accessed: 2020-02-26.

2. Ayuntamiento de Madrid: Madrid 360 - estrategia de sostenibilidad ambiental, . Accessed: 2020-02-25.
3. Ayuntamiento de Madrid: Madrid Central - zona de bajas emisiones. https://tinyurl.com/y2jch2qb, . Accessed: 2020-02-25.
4. Portal de datos abiertos del Ayuntamiento de Madrid. https://www.datos.madrid.es, . Accessed: 2020-02-26.
5. Airborne nitrogen dioxide plummets over China. https://earthobservatory.nasa.gov/images/146362/airborne-nitrogen-dioxide-plummets-over-china. Accessed: 2020-03-07.
6. Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pages 199–213. Springer, 1998.
7. Hiroshi Akima. A new method of interpolation and smooth curve fitting based on local procedures. *Journal of the ACM (JACM)*, 17(4):589–602, 1970.
8. Jeremy Aw and Michael J Kleeman. Evaluating the first-order effect of intraannual temperature variability on urban air pollution. *Journal of Geophysical Research: Atmospheres*, 108(D12), 2003.
9. Nathan Baker, Frank Alexander, Timo Bremer, Aric Hagberg, Yannis Kevrekidis, Habib Najm, Manish Parashar, Abani Patra, James Sethian, Stefan Wild, et al. Workshop report on basic research needs for scientific machine learning: Core technologies for artificial intelligence. Technical report, USDOE Office of Science (SC), Washington, DC (United States), 2019.
10. JM Baldasano, MT Pay, O Jorba, S Gassó, and P Jiménez-Guerrero. An annual assessment of air quality with the CALIOPE modeling system over Spain. *Science of the Total Environment*, 409(11):2163–2178, 2011.
11. Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The annals of statistics*, pages 813–852, 2016.
12. Jenna A Bilbrey, Carlos Ortiz Marrero, Michel Sassi, Andrew M Ritzmann, Neil J Henson, and Malachi Schram. Tracking the chemical evolution of iodine species using recurrent neural networks. *ACS omega*, 5(9):4588–4594, 2020.
13. Robert J Blaszczak. Nitrogen oxides (NO_x): Why and how they are controlled; epa-456/f-99-006r. 1999.
14. Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the national academy of sciences*, 113(15):3932–3937, 2016.
15. Amarjit Budhiraja and Wai-Tong Louis Fan. Uniform in time interacting particle approximations for nonlinear equations of patlak-keller-segel type. *Electronic Journal of Probability*, 22, 2017.
16. Ricky TQ Chen, Yulia Rubanova, Jesse Bettencourt, and David K Duvenaud. Neural ordinary differential equations. In *Advances in neural information processing systems*, pages 6571–6583, 2018.
17. WW Cooper, H Hemphill, Z Huang, S Li, V Lelas, and DW Sullivan. Survey of mathematical programming models in air pollution management. *European Journal of Operational Research*, 96(1):1–35, 1997.

18. Giorgio Corani. Air quality prediction in Milan: feed-forward neural networks, pruned neural networks and lazy learning. *Ecological Modelling*, 185(2-4):513–529, 2005.
19. National Research Council et al. *Rethinking the ozone problem in urban and regional air pollution*. National Academies Press, 1992.
20. Paul J Crutzen. The influence of nitrogen oxides on the atmospheric ozone content. *Quarterly Journal of the Royal Meteorological Society*, 96(408):320–325, 1970.
21. Paul J Crutzen. The role of NO and NO_2 in the chemistry of the troposphere and stratosphere. *Annual review of earth and planetary sciences*, 7(1):443–472, 1979.
22. Aaron Daly and Paolo Zannetti. Air pollution modeling–an overview. *Ambient air pollution*, pages 15–28, 2007.
23. William B DeMore and E Tschuikow-Roux. Ultraviolet spectrum and chemical reactivity of the chlorine monoxide dimer. *Journal of Physical Chemistry*, 94(15):5856–5860, 1990.
24. KG Denbigh, Margaret Hicks, and FM Page. The kinetics of open reaction systems. *Transactions of the Faraday Society*, 44:479–494, 1948.
25. Péter Érdi and János Tóth. *Mathematical models of chemical reactions: theory and applications of deterministic and stochastic models*. Manchester University Press, 1989.
26. Andrew M Fraser and Harry L Swinney. Independent coordinates for strange attractors from mutual information. *Physical review A*, 33(2):1134, 1986.
27. James N Galloway, Zhao Dianwu, Xiong Jiling, and Gene E Likens. Acid rain: China, United States, and a remote area. *Science*, 236(4808):1559–1562, 1987.
28. Michael Guarnieri and John R Balmes. Outdoor air pollution and asthma. *The Lancet*, 383(9928):1581–1592, 2014.
29. Trevor Hastie, Robert Tibshirani, and Ryan J Tibshirani. Extended comparisons of best subset selection, forward stepwise selection, and the lasso. *arXiv preprint arXiv:1707.08692*, 2017.
30. Chunfeng Huang, Tailen Hsing, Noel Cressie, Auroop R Ganguly, Vladimir A Protopopescu, and Nageswara S Rao. Bayesian source detection and parameter estimation of a plume model based on sensor network measurements. *Applied Stochastic Models in Business and Industry*, 26(4):331–348, 2010.
31. IPCC. *Climate Change 2013: The Physical Science Basis. Contribution of Working Group I to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA, 2013. ISBN ISBN 978-1-107-66182-0. doi: 10.1017/CBO9781107415324. URL www.climatechange2013.org.
32. Eurika Kaiser, J Nathan Kutz, and Steven L Brunton. Sparse identification of nonlinear dynamics for model predictive control in the low-data limit. *Proceedings of the Royal Society A*, 474(2219):20180335, 2018.
33. Kim Knowlton, Joyce E Rosenthal, Christian Hogrefe, Barry Lynn, Stuart Gaffin, Richard Goldberg, Cynthia Rosenzweig, Kevin Civerolo, Jia-Yeong Ku, and Patrick L Kinney. Assessing ozone-related health impacts under

a changing climate. *Environmental health perspectives*, 112(15):1557–1563, 2004.

34. Debby Lanser and Jan G Verwer. Analysis of operator splitting for advection–diffusion–reaction problems from air pollution modelling. *Journal of computational and applied mathematics*, 111(1-2):201–216, 1999.
35. Thojorn Larssen and GR Carmichael. Acid rain and acidification in China: the importance of base cation deposition. *Environmental pollution*, 110(1):89–102, 2000.
36. Irene Lebrusán and Jamal Toutouh. Assessing the environmental impact of car restrictions policies: Madrid Central case. In *Ibero-American Congress on Information Management and Big Data*, pages 9–24. Springer, 2019.
37. Philip Leighton. *Photochemistry of air pollution*. Associated Press, 1961.
38. Xiang Li, Ling Peng, Yuan Hu, Jing Shao, and Tianhe Chi. Deep learning architecture for air quality predictions. *Environmental Science and Pollution Research*, 23(22):22408–22417, 2016.
39. Morton Lippmann. Health effects of ozone a critical review. *Japca*, 39(5):672–695, 1989.
40. Fred Lurmann, Ed Avol, and Frank Gilliland. Emissions reduction policies and recent trends in southern California's ambient air quality. *Journal of the Air & Waste Management Association*, 65(3):324–335, 2015.
41. Niall M Mangan, Steven L Brunton, Joshua L Proctor, and J Nathan Kutz. Inferring biological networks by sparse identification of nonlinear dynamics. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications*, 2(1):52–63, 2016.
42. Stefano Marsili-Libelli. Simplified kinetics of tropospheric ozone. *Ecological modelling*, 84(1-3):233–244, 1996.
43. Fredric C Menz and Hans M Seip. Acid rain in Europe and the United States: an update. *Environmental Science & Policy*, 7(4):253–265, 2004.
44. Konstantinos P Moustris, Ioannis C Ziomas, and Athanasios G Paliatsos. 3-day-ahead forecasting of regional pollution index for the pollutants no 2, co, so 2, and o 3 using artificial neural networks in Athens, Greece. *Water, Air, & Soil Pollution*, 209(1-4):29–43, 2010.
45. James D Murray. *Mathematical biology: I. An introduction*, volume 17. Springer Science & Business Media, 2007.
46. Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2): 227–234, 1995.
47. Phuong Nguyen, Chanh Kieu, and Wai-Tong Louis Fan. Stochastic variability of tropical cyclone intensity at the maximum potential intensity equilibrium. *Journal of the Atmospheric Sciences*, pages 1–41, 2020.
48. Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.
49. Yoshimitsu Ogura and Norman A Phillips. Scale analysis of deep and shallow convection in the atmosphere. *Journal of the atmospheric sciences*, 19(2):173–179, 1962.
50. Esteban Pardo and Norberto Malpica. Air quality forecasting in Madrid using long short-term memory networks. In *International Work-Conference on the Interplay Between Natural and Artificial Computation*, pages 232–239.

Springer, 2017.

51. David Popp. International innovation and diffusion of air pollution control technologies: the effects of NO_x and SO_2 regulation in the US, Japan, and Germany. *Journal of Environmental Economics and Management*, 51(1): 46–71, 2006.
52. Victor R Prybutok, Junsub Yi, and David Mitchell. Comparison of neural network models with arima and regression models for prediction of Houston's daily maximum ozone concentrations. *European Journal of Operational Research*, 122(1):31–40, 2000.
53. Christopher Rackauckas, Yingbo Ma, Julius Martensen, Collin Warner, Kirill Zubov, Rohit Supekar, Dominic Skinner, and Ali Ramadhan. Universal differential equations for scientific machine learning. *arXiv preprint arXiv:2001.04385*, 2020.
54. Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.
55. John Reyn. *Phase portraits of planar quadratic systems*, volume 583. Springer Science & Business Media, 2007.
56. SM Robeson and DG Steyn. Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. *Atmospheric Environment. Part B. Urban Atmosphere*, 24(2):303–312, 1990.
57. Tim Sauer, James A Yorke, and Martin Casdagli. Embedology. *Journal of statistical Physics*, 65(3-4):579–616, 1991.
58. John H Seinfeld and Spyros N Pandis. *Atmospheric chemistry and physics: from air pollution to climate change*. John Wiley & Sons, 2016.
59. AF Stein, Roland R Draxler, Glenn D Rolph, Barbara JB Stunder, MD Cohen, and Fong Ngan. Noaa's hysplit atmospheric transport and dispersion modeling system. *Bulletin of the American Meteorological Society*, 96(12): 2059–2077, 2015.
60. Floris Takens. Detecting strange attractors in turbulence. In *Dynamical systems and turbulence, Warwick 1980*, pages 366–381. Springer, 1981.
61. Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.
62. James Townsend, Niklas Koep, and Sebastian Weichwald. Pymanopt: A python toolbox for optimization on manifolds using automatic differentiation. *The Journal of Machine Learning Research*, 17(1):4755–4759, 2016.
63. Sebastian Wallot and Dan Mønster. Calculation of average mutual information (ami) and false-nearest neighbors (fnn) for the estimation of embedding parameters of multidimensional time series in matlab. *Frontiers in psychology*, 9:1679, 2018.
64. Robert Wilhelmson and Yoshimitsu Ogura. The pressure perturbation and the numerical modeling of a cloud. *Journal of the Atmospheric Sciences*, 29(7):1295–1307, 1972.

65. Na Zhu, Dingyu Zhang, Wenling Wang, Xingwang Li, Bo Yang, Jingdong Song, Xiang Zhao, Baoying Huang, Weifeng Shi, Roujian Lu, et al. A novel coronavirus from patients with pneumonia in china, 2019. *New England Journal of Medicine*, 2020.

Appendix A. Methods considered

1. *Best subset regression*: The very notion of sparse regression suggests the selection of a subset of $1 \leq c \leq n$ terms. This subproblem, called the *best subset* problem, can be cast generally as:

$$\begin{aligned} &\underset{\boldsymbol{\beta_i}}{\text{minimize}} \quad f(\boldsymbol{\beta_i}) \\ &\text{subject to} \\ &||\boldsymbol{\beta_i}||_0 \leq c, \end{aligned} \tag{A.1}$$

where the ℓ_0-norm $||\boldsymbol{\beta_i}||_0 = \sum_{j=1}^{n} 1\{\beta_{i_j} \neq 0\}$ is an indicator function that denotes the number of non-zero elements of the vector $\boldsymbol{\beta_i}$ (except for the intercept). Therefore, Problem (A.1) aims at finding a sparse representation of $\dot{y}_i(t)$ that has at most c terms and that minimizes $f(\boldsymbol{\beta_i})$ (very often the sum of errors squared, i.e., $f(\boldsymbol{\beta_i}) = \frac{1}{2}||\dot{\tilde{\mathbf{y}}}_\mathbf{i} - \tilde{\mathbf{F}}\boldsymbol{\beta_i}||_2^2$, also known as *least-squares regression*). However, its only constraint is combinatorial in nature and makes this optimization problem NP-hard [46]. Thus, despite of very promising and recent efforts with *mixed-integer optimization* reformulations [11], researches and practitioners alike usually resort to different alternatives to attain sparse regressions.

2. *LASSO regression*: One such option lies on a convex quadratic alternative to Problem (A.1) known as *LASSO* (**L**east **A**bsolute **S**hrinkage and **S**election **O**perator) regression [61]:

$$\begin{aligned} &\underset{\boldsymbol{\beta_i}}{\text{minimize}} \quad \frac{1}{2}||\dot{\tilde{\mathbf{y}}}_\mathbf{i} - \tilde{\mathbf{F}}\boldsymbol{\beta_i}||_2^2 \\ &\text{subject to} \\ &||\boldsymbol{\beta_i}||_1 \leq \phi, \end{aligned}$$

where $||\boldsymbol{\beta_i}||_1 = \sum_{j=1}^{n} |\beta_{i_j}|$ is the ℓ_1-norm of the vector $\boldsymbol{\beta_i}$ (except the intercept). If we denote by $\hat{\boldsymbol{\beta}}_\boldsymbol{i}^*$ the values of the regression coefficients of the full (i.e., the unconstrained) regression, then any value of ϕ such that $||\hat{\boldsymbol{\beta}}_\boldsymbol{i}^*||_1 > \phi$ will produce a *shrinkage*. Geometrical considerations in this model make this shrinkage such that some coefficients will be identical to zero as we decrease the upper bound on the ℓ_1-norm (see [61] for more details). This optimization

model is frequently expressed as an equivalent unconstrained problem with a regularization parameter λ:

$$LR(i): \underset{\boldsymbol{\beta_i}}{\text{minimize}} \quad \frac{1}{2}||\dot{\tilde{\mathbf{y}}}_\mathbf{i} - \tilde{\mathbf{F}}\boldsymbol{\beta}_i||_2^2 + \lambda||\boldsymbol{\beta}_i||_1 \tag{A.2}$$

Since the problems $LR(i), i = 1, 2, \ldots, p$ are quadratic and convex, they can be efficiently solved by some well-known optimization methods for finding the optimal solutions of a convex quadratic function over a polyhedron (see [48] for a reference of some usual optimization methods for this kind of problems). It is worth mentioning that many researchers have historically seen Problem (A.2) as a heuristic to solve Problem (A.1), which is widely regarded as the formulation that yields the most desired sparse solution with a subset of c variables. However, as noted in [29], in noisy settings both problems offer different *bias-variance tradeoffs* and, for this reason, the superiority of best subset regression over LASSO regression is not clear-cut.

Appendix B. Algorithm for SINDy with AIC for one station

Algorithm 1: SINDy with AIC for one station

Result: Find the best regression in the sense of AIC that did not present singularities.

Let p be the number of chemical species; n be the number of variables in the full regression model; ϵ be a threshold for the maximum value allowed for $\dot{y}_i(t)$;

$l_i = 1$ be a counter denoting which ranked model for chemical species i should be selected;

$l = 0$ be a counter for the total number of models discarded;

Find the $\mathcal{F} = \bigcup_{j=1}^{2^n} \mathcal{F}_j$ possible subsets of variables with size less or equal to n;

for $i \leftarrow 1$ **to** p **do**

  for $j \leftarrow 1$ **to** 2^n **do**

    Solve the system (7) in the sense of least-squares with the subset of variables $\mathcal{F}_j$;

  end

  Rank the 2^n different models for the i^{th} chemical species according to the AIC;

end

while $l \geq 0$ **do**

  for $i \leftarrow 1$ **to** p **do**

    Select the l_i^{th} best-ranked model for chemical species i;

  end

  With the selected models, solve numerically the system of ODEs defined in (11-12);

  if $\dot{y}_i(t) > \epsilon$ **then**

    $l_i = l_i + 1$;

    $l = l + 1$;

  else

    $l = -1$;

  end

end

Appendix C. Reconstruction Algorithm

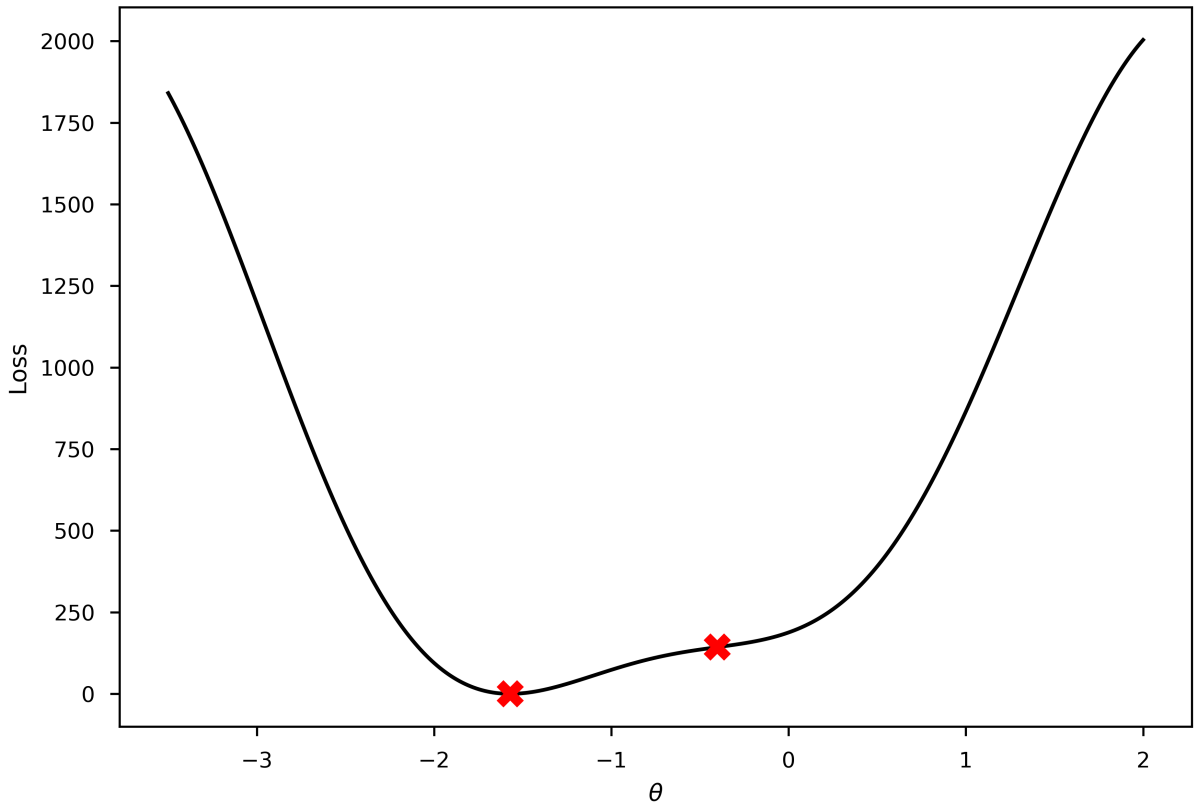


Figure C.10: This figure represents a portion of the loss landscape when minimizing equation 15 using the data outlined in Figure 9. In order to visualize the loss landscape, we use the fact that 2×2 rotations and reflections can be parametrized by one angle. Consider the following family of parametrized matrices in $O(2)$, $\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}\begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}$ for some angle θ. The marked values correspond to local minima of equation 15 over this family of matrices. The far left marked value corresponds to the angle we chose to perform the correction i.e. the grey curve show in Figure 9. For higher dimensional problems the loss landscape is more difficult to explore and visualize but packages such as Pymanopt [62] allow users to solve optimization problems over sets of $n \times n$ matrices.