Simple Versus Composed Temporal Lag Regression with Feature Selection, with an Application to Air Quality Modeling

Estrella Lucena-Sánchez
Dept. of Mathematics and Computer Science
University of Ferrara
Ferrara, Italy
Dept. of Physics, Informatics, and Mathematics
University of Modena and Reggio Emilia
Modena, Italy
estrellalucena.sanchez@unife.it

Fernando Jiménez
Dept. of Information and Communications Engineering
University of Murcia
Murcia, Spain
fernan@um.es

Guido Sciavicco
Dept. of Mathematics and Computer Science
University of Ferrara
Ferrara, Italy
guido.sciavicco@unife.it

Joanna Kamińska
Dept. of Mathematics
Wrocław University of Environmental and Life Sciences
Wrocław, Poland
joanna.kaminska@upwr.edu.pl

***Abstract*—Anthropogenic environmental pollution is a known and indisputable issue, and the need of ever more precise and reliable land use regression models is undeniable. In this paper we consider two years of hourly data taken in Wrocław (Poland), that contain the concentrations of NO_2 and NO_x in the atmosphere, and, along these, traffic flow, air pressure, humidity, solar duration, temperature, and wind speed. In the quest for an explanation model for the pollution concentrations, we improve and generalize the simple temporal lag regression model, and introduce a composed temporal regression model that entails a transformation of the data to improve the effectiveness of classical learning algorithms. We show that using the latter we obtain more accurate and better interpretable explanation models than using the former, and also than using the original, non-transformed data.**

***Index Terms*—temporal lag regression; land use regression model.**

I. Introduction

Anthropogenic environmental pollution is a known and indisputable issue. In everyday life, the human body is exposed to harmful substances in various ways, including consumption with food and drink or absorbtion with breathing. Every day, with every breath, chemicals that are dangerous to health or in amounts that make them dangerous, such as NO and NO_2, are absorbed into the lungs and body. The potential negative effects to the human health of such an exposure has been confirmed by several scientific studies (see [4], [18], [27], [30], [34], [36], among many others). Air quality is regularly monitored and alert systems inform residents about the forecasted or ensuing high concentration of air pollutants, and, generally, anthropogenic sources of air pollution are known; due to the development of civilization, it is impossible to completely eliminate them. Therefore, an effort is necessary to determine the impact of factors that modify the concentration of pollutants in the air such as transformation, retention or evacuation. For this purpose, mathematical models are created to describe phenomena and relationships occurring more or less in detail, for example autoregressive models [3], [32]. But more than predicting concentrations of pollutants, recognizing the factors that have the greatest impact on them gives the opportunity to take such actions that enable the fastest possible evacuation of contaminants, thereby shortening the time of exposure to its harmful effects and reducing the intensity of this action. *Land use regression models* (*LUR*s) are a cost-effective approach for predicting the variability in ambient air pollutant concentrations with high resolution; examples include [13], [16] (see also [17] for a survey). These models are often developed based on a series of previous observations; however, it is unclear how well such models perform when extrapolated over time [39]. Moreover, in addition to the identification of the factor itself, the time of its operation also plays a significant role; for example, a momentary gust of wind results in a much smaller evacuation of pollutants, and thus a decrease in concentration, than the wind with the same speed persisting for several hours. To better explain a contamination concentration and taking the temporal component into account, one may want to identify not only the factors into play

E. Lucena-Sánchez and G. Sciavicco acknowledge the support of the project *New Mathematical and Computer Science Methods for Water and Food Resources Exploitation Optimization*, founded by the Italian region Emilia-Romagna.

but also the moment in time for which their influence is the greatest, and build, in such a way, a *temporal land use regression model* (*TLUR*). By taking into account the temporal component, a TLUR is, at the same time, more accurate and more interpretable than a LUR.

In this paper we approach the problem of extracting a TULR with a novel learning methodology, at whose core there is an optimization problem. First, we define the concept of *optimum lag*, that is, the delay at which a influencing factor acts; for example, a certain pollution may be influenced by the amount of traffic in the area with one or two hours of delay. Then, we define the concept of *interval* (of influence), that is, the amount of delays, around the optimum one, at which an influencing factor still acts; for example, the amount of traffic may have an optimum delay of two hours, but still be important with a delay of one hour and with a delay of three hours, and in this case we would say that its interval is three hours. Finally, we introduce the idea of *combining function*, in the attempt to model in which way different delays have an effect; for example, with an optimum lag for the variable traffic of two hours and an interval of three hours, it may be that the amount of traffic with a delay of three hours (included in the interval) influences the amount of pollutant with a smaller factor than the amount of traffic at the optimal delay. We propose a methodology that searches for the optimal setting of such parameters, and returns a transformation of the original data, which can be thought of as a *convolution matrix*, and which can be applied to the original data so that a learning algorithm performs better. The matrix itself offers an angle for interpretation; in our example, this would entail devising an environmental theory that explains why, and in which way, the amount of traffic between three and one hour before a given moment influences the amount of pollution at that moment, and why the influence of two hours before is the greatest. Therefore, in the quest for an interpretable TLUR, one may even apply, after the transformation, a black-box learning algorithm (e.g., random forests), and still ending up with a (at least partially) interpretable model. We apply our methodology to a data containing NO_2 and NO_x concentrations measured hourly from 2015 to 2017 by a monitoring station located in Wrocław, Poland, along with meteorological and vehicle traffic data for each measure, to search for a TLUR of that specific area. We compare our results with those obtained with the same data using random forests but no transformation (as in [26]), and using only lag transformation and linear regression (as in [23]).

II. Background

A. Air Quality Modeling in Wrocław

Urban air pollution is a matter of growing concern for both public administrations and citizens. Road traffic is one of the main sources of air pollutants, though topography characteristics and meteorological conditions can make pollution levels increase or diminish dramatically. In this context an upsurge of research has been conducted towards functionally linking variables of such domains to measured pollution data, with studies dealing with up to one-hour resolution meteorological data. However, the majority of such reported contributions do not deal with traffic data or, at most, simulate traffic conditions jointly with the consideration of different topographical features.

In [26], the authors use random forests in an attempt to model the regression relationships between concentrations of NO_2, NO_x and $PM_{2.5}$, and nine variables describing meteorological conditions, temporal conditions and traffic flow. The study was based on hourly values of wind speed, wind direction, temperature, air pressure and relative humidity, temporal variables, and traffic flow, between 2015 and 2016. An air quality measurement station was selected on a main road, located at a short distance from a large intersection equipped with a traffic flow measurement system. Nine different time subsets were defined, based among other things on the climatic conditions in Wrocław. An analysis was made of the fit of models created for those subsets, and of the importance of the predictors. Both the fit and the importance of particular predictors were found to be dependent on season. The most important explanatory variable in the models of concentrations of nitrogen oxides was traffic flow, while in the case of $PM_{2.5}$ the most important factors were meteorological conditions, in particular temperature, wind speed and wind direction. Temporal variables (intended as taking into account the day of the week, or the fact that a particular measure was taken during a weekday or a holiday) were found to have no significant effect on the concentrations of the studied pollutants. The same data were analyzed in [23] using a simple lag regression method. The regression model used in that instance was linear, and, for each variable, the aim was to devise the *optimum lag*, that is, the optimum amount of delay for the variable to have an effect. The transformation, in that occasion, was obtained and applied on the same fraction of the data. The most important results were that, first, proving that non-trivial delays have a role in the pollution concentration (only NO_2 and NO_x were taken into account), and, second, that transforming the data using optimal delays does improve the ability of the experts to interpret the underlying phenomena.

B. Feature Selection and Evolutionary Algorithms

One of the most basic data transformations is *feature selection*, defined as the process of eliminating features from the data base that are irrelevant to the task to be performed [14]. Feature selection facilitates data understanding, reduces the storage requirements, and lowers the processing time, so that model learning becomes an easier process; *multivariate* feature selection [31], in particular, consists of assessing the quality of a subset of features, taking into account their mutual dependencies, for classification or regression. Feature selection strategies are usually categorized into filter, wrapper and embedded models. *Filters* are algorithms that perform the selection of features using an evaluation measure that classifies their ability to differentiate classes without making use of any machine learning algorithm. *Wrapper* methods select variables driven by the performances of an associated learning algorithm. Finally, *embedded* models perform the two operations (selecting variables and building a classifier) at the same time. Wrapper methods for feature selection are more common in the literature; often, they are implemented by defining the selection as a search problem, and solved using metaheuristics such as evolutionary computation (see, e.g., [6], [40], [43]).

The use of evolutionary algorithms for the selection of features in the design of automatic pattern classifiers was introduced in [38]. Since then, genetic algorithms have come to be considered as a powerful tool for feature selection [40], and have been proposed by numerous authors as a search strategy in filter, wrapper, and embedded models [1], [9], [20], as well as feature weighting algorithm and subset selection algorithms [43]. The first evolutionary approach involving multi-objective optimization for feature selection was proposed in [21]. A formulation of feature selection as a multi-objective optimization problem has been then presented in [10]. Wrapper approaches are also proposed in [33], [37]. In [12] two wrapper methods with three and two objectives, respectively, applied to cancer diagnosis are compared. The three-objectives version optimizes the sensitivity, the specificity and the number of genes, while the two-objectives one optimizes the accuracy and the number of genes. NSGA-II is used as search strategy, and a support vector machine is used for the classification task. Very recent examples of multi-objective feature selection systems can be found in [22], [24], [25]. Given that a *multi-objective optimization problem* [5] can be formally defined as the optimization problem of simultaneously minimizing (or maximizing) a set of k arbitrary functions:

$$\begin{cases} \min/\max \ f_1(\bar{x}) \\ \min/\max \ f_2(\bar{x}) \\ \dots \\ \min/\max \ f_k(\bar{x}), \end{cases} \tag{1}$$

where $\bar{x}$ is a vector of decision variables. A generic formalization of a wrapper methodology for feature selection in a data set with $n+1$ features as a two-objective optimization problem can be devised from the literature: it simply involves adapting (1) to simultaneously optimizing two functions:

$$\begin{cases} \min/\max \ f(\bar{x}) \\ \min \ CARD(\bar{x}), \end{cases} \tag{2}$$

where $\bar{x}$, that takes values in $\{0,1\}$, represents the set of chosen features (1 means that the feature is selected, and 0 that it is discarded), $CARD(\bar{x})$ represents its cardinality:

$$CARD(\bar{x}) = \sum_{t=1}^{n} \bar{x}(t) \tag{3}$$

and $f(\bar{x})$ is any measure of the goodness of the learning algorithm used when applied to the data set obtained by selecting only those features indicated by $\bar{x}$. If we choose to minimize such a function, then possible choices include defining it as the opposite of the accuracy, or the correlation coefficient, or the mean squared error, among many others, depending on the specific classification/regression problem to be solved. Observe that applying $\bar{x}$ entails transforming the original data set by selecting specific attributes while discarding the others; such a concept, as we shall see, can be easily generalized.

C. Multivariate Time Series and Lag Regression

A *time series* is a series of data points labelled with a temporal stamp. If each data point contains a single time-dependent value, then the time series is *univariate*; otherwise, it is called *multivariate*. Time series arise in multiple contexts, for example, medical patients, who can be considered as time series in which every interesting medical value varies over time (e.g., fever, pain level, blood pressure), or environmental monitoring stations, which can also be considered time series, in which atmospheric values change over time (e.g., pressure, concentration of chemicals). There are two main problems associated with single time series: *time series explanation* and *time series forecasting*. Explaining a time series aims to construct a (possibly interpretable) model that explains the present values; forecasting a time series implies testing and using the model to predict future values. In the univariate case, a model of a time series is based uniquely on the values of the series itself; for example, a forecasting model for the stock price of a certain company would allow one to predict the future price (e.g., in the next two days) based on the prices of the same company (e.g., the price in each day of the past week). Univariate forecasting models are very common in the literature [2], [19], [41]; however, they all belong to the *autoregressive integrated moving average* (ARIMA) family [35]. The multivariate version of this (family of) model(s) is known as ARIMAX.

In time series explanation, in the multivariate case, one identifies one dependent variable (time series), and aims to construct a model to explain (and possibly predict) its present values based on the past and present values of other, independent variables (which themselves are time series): this is usually done with *lagged* models. While ARIMA-type models emerge from computational statistics, lagged models belong to the machine learning domain, and, in general, they consist of creating *lagged* version of (a subset of) the independent variable to construct a larger data set that is then used to create a model of the dependent time series using classical, propositional algorithms (such as, for example, linear regression). Among the available packages to this purpose we mention Weka's *timeseriesForecasting* [15]. The main limitation of multivariate lagged models is precisely the choice of lag variables and lag amounts. In some cases, it is difficult to foresee the necessary lag amount. Moreover, uncontrolled lag variable creation may lead to very large data bases which, when treated with propositional algorithms, may lead to poorer results, as unnecessary lag variables become noise. Finally, even if lagged variables increase the quality of the result, the obtained function may not be easy to interpret.

In [23] the authors define the problem of devising the *optimal lag* for each variable in a multivariate lagged regression as the adaptation and generalization of the feature selection problem. The obtained methodology can be referred to as *simple lag regression*, and, in this paper, we further expand and generalize it.

III. Simple and Composed Lag Regression

A. Lag Regression

The underlying idea to lag regression is to devise a transformation of the original data to highlight the optimum lag for each independent variable. Variable lag optimization is, in a way, a generalization of feature selection; each variable is either lagged of a finite quantity (or zero), or eliminated. Therefore, it can be solved using the same schema, that is, via a wrapper defined as an optimization problem. By definition, a wrapper is based on a learning algorithm whose performance is used to guide the optimization. We choose to use linear regression to this purpose, but, as we have already mentioned, the optimal solution(s) is (are) transformation(s) than can be then applied to the original data, which can be subsequently dealt with any regression algorithm.

Given a data set A with n independent variables $A_1, \ldots, A_n$ and one observed variable B, solving a linear regression problem consists of finding $n+1$ *parameters* (or *coefficients*) $c_0, c_1, \ldots, c_n$ so that the equation:

$$B = c_0 + \sum_{i=1}^{n} c_i \cdot A_i + \epsilon, \tag{4}$$

where ϵ is a random value, is satisfied. Starting from a data set of observations:

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} & b_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} & b_m \end{bmatrix} \tag{5}$$

the regression problem is usually solved by suitably estimating the coefficients c_i so that, for each $1 \leq j \leq m$:

$$b_j \approx c_0 + \sum_{i=1}^{n} c_i \cdot a_{ij} + \epsilon. \tag{6}$$

There are several available, and well-known algorithms to solve such an inverse problem. The performance of such an estimation can be measured in several (standard) ways, such as *correlation, covariance, mean squared error*, among others. When we are dealing with a multivariate time series, composed by n independent and one dependent time series, then data are temporally ordered and associated to a time-stamp:

$$\begin{bmatrix} a_{11} & a_{12} & \ldots & a_{1n} & b_1 & t_1 \\ a_{21} & a_{22} & \ldots & a_{2n} & b_2 & t_2 \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots \\ a_{m1} & a_{m2} & \ldots & a_{mn} & b_m & t_m \end{bmatrix} \tag{7}$$

Using linear regression to explain B, then, entails that finding optimal coefficients for:

$$B(t) = c_0 + \sum_{i=1}^{n} c_i \cdot A_i(t) + \epsilon, \tag{8}$$

because we aim to explain B at a certain point in time t using the values $A_1(t), \ldots, A_n(t)$. Equations (4) and (8) model exactly the same problem, only in the latter the temporal component is made explicit. *Lag* (linear) regression consists of solving a more general equation, formulated as:

$$B(t) = c_0 + \sum_{i=1}^{n} \sum_{l=0}^{p_i} c_{i,l} \cdot A_i(t-l) + \epsilon. \tag{9}$$

In other words, we use the value of each independent variable A_i not only at time t, but also at time $t-1, t-2, \ldots, t-p_i$, to explain B at time t; each $A_i(t-l)$ is associated to a coefficient $c_{i,l}$, which must be estimated, along with each *maximum lag* p_i. There are available techniques, based on standard regression algorithms, that allow one to solve the inverse problem associated to (9); unfortunately, the resulting equation may result very difficult to interpret. Equation (9) is very similar to an ARIMAX model: this is due to the fact that we have chosen to focus, here, on linear models; as we have already observed, our approach is agnostic with respect to the regression model, allowing one to obtain, in fact, explanation models that are very different from (9).

Now, we work under the additional assumption that, for each variable index i there is precisely one lag l_i, such that $A_i(t-l_i)$ influences the output more than any other lag; this may be reasonable in some applications, and less so in others: as we shall see, it fits perfectly our case. Under such an assumption, the *simple lag regression* model is:

$$B(t) = c_0 + \sum_{i=1}^{n} c_i \cdot A_i(t-l_i) + \epsilon. \tag{10}$$

A multi-objective search strategy to approximate (10) has been presented in [23]. Simple lag regression can be further improved by allowing more than one lag for each predictor to have some influence on the result. This makes sense in certain applications in which a given predictor needs some time to act as a cause, but it may be the case that it contributes (lesser) even some time *before* and/or some time *after* the optimal time. So, we can take these potential contributions into account by first defining a function:

$$d_{q_i} : \mathbb{N} \to [0,1] \tag{11}$$

to formalize the amount of contribution of A_i at lags different from the optimal one (l_i), where $[0,1] \subset \mathbb{R}$ and q_i is the *amplitude* (or *interval*) of contribution (i.e., it measures how many time units before and after the optimum lag there is a nonzero contribution from A_i). Now, we can define the *combined lag regression model* as follows:

$$B(t) = c_0 + \sum_{i=1}^{n} c_i \cdot \frac{\sum_{s=0}^{q_i} d_{q_i}(s) \cdot A_i(t-l_i \pm s)}{2 \cdot q_i} + \epsilon. \tag{12}$$

There are several implicit assumption in (12). First, we assume that the contributions of A_i before and after the optimal lag are symmetric, while they may not be. Second, we assume that, for each l, s, the value of $A_i(t-l+s)$ is zero if $s > l$; in other words, the value of B at a given time cannot depend on the value of A_i at some future time. Third, we assume that the contribution of A_i over the interval (of range q_i) around the optimal lag l_i can be modelled with an analytic function ($d_{q_i}()$) and associated to the same constant c_i.

The purpose of this paper is to design a multi-objective optimization schema to solve the inverse problem associate to (12) by generalizing the one used for (10). The solution to it can be seen as a *convolution matrix*, that entails a transformation of the original data. The wrapper is trained with linear regression, and a fragment of the data; once the optimal solution(s) are chosen, they are applied to the entire data set, and then the problem of multivariate regression can be solved with any regression algorithm.

B. Optimization

Inspired by the classical multi-objective optimization-based solution for feature selection, we can instantiate (1) in order to solve the inverse problem associated to (12). For a multivariate time series $A_1(t), \ldots, A_n(t), B(t)$ with m distinct observations, let $\bar{x} = (x_1, \ldots, x_n)$ be a vector of decision variables; each x_i is, in turn, a triple (x_i^1, x_i^2, x_i^3). We set the following constants: K, to represent the *maximum lag* allowed, and I to represent the *maximum amplitude* (or *interval*) allowed; these are set so that the optimizer has the necessary limits for the search space. Given a decision variable $x_i = (x_i^1, x_i^2, x_i^3)$, we decide that x_i^1 represents the current lag of the i-th variable, so its domain is $[-1, \ldots, K] \subset \mathbb{N}$ (its actual value is denoted by l_i in (12), and the value -1 represents the fact that the i-th variable is discarded), that x_i^2 represents the current amplitude of the i-th variable (its actual value is represented by q_i in (12)), so its domain is $[0, \ldots, I] \subset \mathbb{N}$, and that x_i^3 represents the combining function for the lags of the i-th variable within its interval (it is the function d_{q_i} in (12)). We have chosen to implement three possible combining functions, denoted by $\{0, 1, 2\}$; in particular, 0 represents the *uniform* combining function (i.e., all lags in the vicinity of the optimum one contribute the same), 1 represents the *linear* combining function (i.e., the contribution decreases at longer lags in a linear way), and 2 represents the *hyperbolic* combining function (i.e., the contribution decreases at longer lags in a non-linear way):

$$d_I^{uni}(s) = \begin{cases} 1 & s \leq I \\ 0 & s > I \end{cases} \tag{13}$$

$$d_I^{lin}(s) = \begin{cases} -\frac{1}{I} + 1 & s \leq I \\ 0 & s > I \end{cases} \tag{14}$$

$$d_I^{hyp}(s) = \begin{cases} 1 & s = 0 \\ \frac{1}{I^2} & 0 < s \leq I \\ 0 & s > I \end{cases} \tag{15}$$

Therefore, every solution vector $\bar{x}$ entails a transformation of the original data set in which the i-th variable is either discarded, or lagged of a certain amount and combined with the same variable at different, neighboring, lags.

After applying a transformation, the resulting data set can be passed to any linear regression algorithm to solve the inverse problem associated to (12). Several functions can be designed to evaluate a candidate solution, for different goals, and the more complex is the transformation, the wider is the range of possible evaluation functions (objectives). In this particular experiment we have chosen to instantiate (1) in the simplest possible way:

$$\begin{cases} \max \ CORR(\bar{x}) \\ \min \ CARD(\bar{x}), \end{cases} \tag{16}$$

where $CARD$ is suitably modified to count the number of chosen variables:

$$CARD(\bar{x}) = \sum_{i=1}^{n} \begin{cases} 0 & \text{if } x_i \neq -1 \\ 1 & \text{otherwise}, \end{cases} \tag{17}$$

and $CORR$ is simply the correlation coefficient of the regressed function. *Multi-objective evolutionary algorithms* are known to be particularly suitable to perform multi-objective optimization such as (16), as they search for multiple optimal solutions in parallel. In this experiment we have chosen the well-known NSGA-II (Non-dominated Sorted Genetic Algorithm) [7] algorithm, which is available as open-source from the suite *jMetal* [8]. NGSA-II is an elitist Pareto-based multi-objective evolutionary algorithm that employs a strategy with a binary tournament selection and a rank-crowding better function, where the rank of an individual in a population is the non-domination level of the individual in the whole population. As black box linear regression algorithm, we used the class *linearRegression* from the open-source learning suite *Weka* [42], run in 10-fold *cross-validation* mode, with standard parameters and no embedded feature selection.

IV. Experiment and Results

A. Data and Setting

There is only one communication station for measuring the air quality in the city of Wrocław, and it is located within a wide street with two lanes in each direction (GPS coordinates: 51.086390 North, 17.012076 East). The center of one of the largest intersections in the city with 14 traffic lanes is located approximately 30 meters from the measuring station, and is covered by traffic monitoring. The measurement station is located on the outskirts of the city, at $9.6kms$ from the airport. Pollution data are collected by the Provincial Environment Protection Inspectorate and encompasses the hourly NO_2 and NO_x concentration values during three full years, from 2015 to 2017. The traffic data (denoted by t, in our results) are provided by the Traffic Public Transport Management Department of the Roads and City Maintenance Board in Wrocław, and include hourly count of all types vehicles passing the intersection. Public meteorological data are provided by the Institute of Meteorology and Water Management, and they include: air temperature (a), solar duration (d), wind speed (w), relative humidity (h), and air pressure (p). For the sake of uniformity, solar duration values have been renormalized in the real interval $[0, 1]$ (as standard). The full data set contains 26304 observations. In the pre-processing phase, the instances with at least one missing value (617 samples, 2.3%) have been excluded. To these data, we have: *(i)* applied linear regression (standard parameters) and random forest (standard parameters, 100 trees) in full training mode; *(ii)* applied linear regression (standard parameters) and random forest (standard parameters, 100 trees) in 10-fold cross-validation mode; *(iii)* applied simple and composed temporal lag regression to the first 30% of the data trained with linear regression, and, after choosing the best solution (one for the simple model and one for the composed model), used the corresponding transformation to the 100% of the data and run linear regression (standard parameters) and random forest (standard parameters, 100 trees) in 10-fold cross-validation mode. For each execution, NSGA-II has been run with an initial, randomly generated population of 100 individuals, for 100 generations. Mutation and crossover have been suitably modified to deal with our individuals, and their relative probabilities have been left as originally set in the *jMetal* library. At the end of each execution, the best individual in terms of correlation coefficient has been chosen. We have launched 10 independent executions, with random seeds from 1 to 10; K was set to 24 and I was set to 2 in each execution.

B. Results

Tab. I shows the results of the initial experiment with 30% of the data and with the wrapper trained with linear regression, following (16). Boldfaced values, despite not being the most performing in terms of correlation coefficient, were found to be those whose underlying theory is the most natural and interpretable; they were chosen for the rest of the experiment. The last column shows the correlation coefficient that emerges if data are transformed using lags only, and not using column combination (as in [23]). Tab. II shows the results of the second part of the experiment. The last five columns correspond to the following indicators: correlation coefficient (*coeff.*), mean absolute error (*mae*), relative absolute error (*rae*), root mean squared error (*rmse*), and root relative squared error (*rsse*); after applying Weka's random forest algorithm and Weka's linear regression algorithm firstly on the original data, with no transformation; and later on the data transformed as indicated by the chosen individual from Tab. I.

C. Discussion and interpretation

The above results can be analyzed along several directives. First, observe how both in average and the each single cross-validation execution the correlation coefficient of the models (both linear regression and random forest) trained on the transformed data improve those of the models trained on the original data. Such an improvement is more evident in the case of NO_2, but still very clear in the case of NO_x as well. This is an indication that the transformation has had a positive effect on the ability of the regression algorithms to extract the underlying rule that links the causes to the effect, as a convolution matrix is to be expected to do. Second, observe how using lags only (Tab. I), the correlation coefficients, even if only for the 30% of the data, are clearly inferior. This implies that simple temporal regression is not as performing as combined temporal regression, as one may expect. Third, and most importantly, our transformations are (at least partially) interpretable *per se*, as they highlight the role of the delay for each independent variable.

The NO_2 concentration is positively correlated with air temperature and traffic flow. This means that increase temperature and/or traffic flow results in increase of nitrogen dioxide concentration. Traffic flow, and, in particular, exhaust gasses emission, is the main source of nitrogen dioxide in the traffic corridor. In our model, the current value of NO_2 concentration is influenced by the amount of traffic between 2 hours before and the current moment, with a linear (descending) behaviour; this may be due to an accumulation phenomenon. The typical equations that model the decrement in concentration of pollutants propose a Gaussian behaviour (e.g., in Pasquilla's equations), and, in our model, this approximated by a linear decreasing behaviour. Air temperature has a positive effect on NO_2 concentrations; in our model, this effect is as wide as two hours, with a linear decreasing behaviour. As a matter of fact, as the temperature increases, the rate of chemical reactions in the atmosphere increases; in addition, during the day when the temperature is higher, the concentrations of pollutants are also higher, while at night, when the emission is lower and chemical reactions reducing the concentration of NO_2 occur, these concentrations are lower. In fact, the most important variable that influences chemical reactions with nitrogen oxides is sunshine duration: under the influence of sunlight, a Leighton relationship occurs [11], [29], as an effect of which NO_2 disintegrates into NO and ozone; from our model, it emerges that the influence of sunlight is maximal at moment of the measure and one hour before that, with a uniform behaviour. Wind speed has an impact on the evacuation of pollution: the stronger its speed, the more intense is the evacuation, and the lower is pollution concentration. Due to the distance between the meteorological and the communication station, the effect of wind speed is delayed by 1 to 3 hours, which is the amount of time needed to cover the $9.6kms$ at the average speed of $3.1m/s$, and taking into account the porosity of urban buildings. This, again, emerges from our model, that gives us a two hours delay and two hours of interval, with a uniform behaviour. The influence of humidity on NO_2 is negative. This is due to two main factors: first, an increase in air humidity causes a decrease in nitrogen oxides in exhaust gases [28], and, second, it is related with low solar duration, which in turn decreases chemical reactions intensity. The change in air humidity is characterized by a large inertia compared to the change in cloud cover, which explains the delay between 4 and 6 hours, with a uniform behaviour. Finally, air pressure is another meteorological condition that indirectly influences NO_2 concentration: the higher the pressure, the more elevated sunshine duration, with the effects that we have already

TABLE I

TRAINING RESULTS, ONE LINE PER EXECUTION. BOLDFACED EXECUTIONS ARE THOSE THAT HAVE BEEN CHOSEN. *: COEFFICIENT OBTAINED WITHOUT USING COLUMN COMBINATION.

individual							
a	d	w	h	p	t	coeff.	coeff*.
NO_2							
(8,2,1)	(0,1,0)	(0,2,0)	(6,1,0)	(0,2,1)	(0,1,0)	0.730	0.683
(0,0,1)	(7,0,1)	(0,2,0)	(7,2,0)	(2,2,1)	(0,2,0)	0.728	0.683
(21,1,1)	(0,2,0)	(3,2,0)	(5,2,0)	(0,0,1)	(0,2,0)	0.732	0.695
(0,2,0)	(10,2,0)	(0,2,0)	(9,2,0)	(0,1,1)	(0,2,0)	0.729	0.676
(0,1,1)	(13,1,0)	(0,2,0)	(6,2,0)	(0,0,0)	(0,2,0)	0.729	0.679
(4,2,0)	(1,1,1)	(2,1,0)	(8,2,1)	(19,2,0)	(0,2,1)	0.733	0.711
(21,2,1)	(1,1,1)	(2,2,1)	(6,2,1)	(3,0,0)	(1,2,1)	0.737	0.723
(0,2,1)	**(0,1,0)**	**(2,1,0)**	**(5,1,0)**	**(0,2,0)**	**(0,2,1)**	**0.732**	0.707
(0,2,1)	(0,2,1)	(1,2,0)	(8,1,0)	(9,2,0)	(0,2,0)	0.733	0.699
(21,1,1)	(0,2,1)	(2,1,0)	(7,1,1)	(0,1,1)	(0,2,0)	0.737	0.712
NO_x							
(0,1,0)	**(9,2,0)**	**(0,2,0)**	**(0,1,0)**	**(0,2,0)**	**(0,1,0)**	**0.629**	0.585
(3,1,0)	(6,2,1)	(2,2,0)	(0,1,1)	(19,1,0)	(0,1,0)	0.628	0.611
(0,1,1)	(8,2,1)	(2,1,0)	(24,0,1)	(0,1,1)	(0,2,1)	0.629	0.611
(2,2,1)	(11,1,0)	(0,2,0)	(0,2,1)	(0,0,1)	(0,1,0)	0.628	0.587
(0,2,1)	(9,1,0)	(2,2,0)	(0,2,0)	(22,1,1)	(0,1,0)	0.638	0.619
(0,0,1)	(10,2,1)	(2,1,0)	(0,1,1)	(5,2,1)	(0,1,0)	0.632	0.615
(0,1,0)	(7,2,0)	(2,2,0)	(0,1,1)	(0,2,0)	(0,2,1)	0.632	0.615
(0,2,1)	(11,2,0)	(2,0,1)	(0,1,0)	(0,1,1)	(0,2,1)	0.630	0.612
(0,1,0)	(8,2,0)	(0,2,0)	(1,1,0)	(2,2,0)	(0,2,1)	0.631	0.594
(0,1,0)	(8,1,0)	(2,0,0)	(0,2,1)	(12,2,1)	(0,1,0)	0.630	0.619

TABLE II

RESULTS, IN 10-FOLD CROSS-VALIDATION MODE AFTER APPLYING LINEAR REGRESSION AND RANDOM FOREST TO THE EXECUTIONS CHOSEN IN TAB. I .

		coeff.	mae	rae	rmse	rrse
NO_2						
RF	original	0.737	11.598	44.851	35.091	67.592
	transf.	**0.798**	**10.482**	**58.105**	**13.975**	**60.326**
LR	original	0.629	13.591	18.017	67.803	77.773
	transf.	**0.713**	**12.349**	**68.452**	**16.240**	**70.104**
NO_X						
RF	original	0.699	45.801	62.221	74.172	71.539
	transf.	**0.730**	**44.928**	**61.034**	**70.862**	**68.342**
LR	original	0.608	53.484	72.658	82.287	79.365
	transf.	**0.626**	**52.996**	**71.995**	**80.883**	**78.007**

discussed above. Its modeled influence starts two hours before the current moment, with a linearly decreasing behaviour.

The NO_x concentration is, in fact, the sum of the concentrations of NO and NO_2. Consequently, the transformation of nitric oxide into nitrogen dioxide, and vice versa, does not affect NO_x concentrations. Therefore, the increase in air temperature affects NO_x concentrations only by reducing emissions due to reduced traffic volumes (nice weather encourages public transport or alternative means of transport) and reduced emissions as a result of combustion with higher temperature inlet air. The traffic volume influence on NO_x is the similar to that on NO_2, and in our model the difference in delay (one versus two hours) is compensated by the difference in behaviour (no decreasing versus linear decreasing). Because of the lack of influence of the transformation of nitric oxide into nitrogen dioxide and vice versa for NO_x concentration, the optimal delay for sunshine duration increases to 7 to 11 hours delay, with uniform behaviour: for example, after a sunny day, we observe relatively high NO_x concentrations at night. The wind speed influence is almost the same as for NO_2; in this case the interval is between 0 and 2 hours delay. Finally, the current relative humidity and air pressure influence positively on NO_x concentrations, with a maximal delay of 2 hours. In fact, higher humidity related with a cloudy sky and even with precipitation prevents the occurrence of chemical reactions between NO_x and VOC. Thus, oxides of nitrogen float in the air, increasing the pollution concentration.

V. CONCLUSIONS

The temporal component in multivariate temporal series regression is very important; yet, available methods tend to ignore it, or to take it into account in a primitive way. Simple temporal lag regression is a first systematic attempt to define the problem of finding the optimal lag for each independent variable in a temporal series, introducing, *de facto*, the concept of convolution matrix for temporal series; it is a transformation of the original data, defined as an optimization problem and solved via an evolutionary algorithm, aimed to improve the performances of any classical regression algorithm. In this paper we generalize and improve the simple temporal lag regression, introducing what we called composed temporal lag regression. We considered atmospheric pollution data from the city of Wrocław (Poland), and we computed optimal convolution matrices for two pollutants: NO_2 and NO_x. We proved that not only the regression algorithms run over the transformed data perform better than those run over non-transformed ones, but that composed lags offer a previously unexpected angle for interpretation. As a matter of fact, our convolution matrices alone are so expressive that a solid, theoretically justified interpretation is possible even for typically non-interpretable model extraction algorithms such as random forests.

References

[1] R. Anirudha, R. Kannan, and N. Patil, "Genetic algorithm based wrapper feature selection on hybrid prediction model for analysis of high dimensional data," in *Proc. of the 9th International Conference on Industrial and Information Systems*, 2014, pp. 1–6.

[2] G. Box, G. Jenkins, G. Reinsel, and G. Ljung, *Time Series Analysis: Forecasting and Control*. Wiley, 2016.

[3] E. Chianese, F. Camastra, A. Ciaramella, T. C. Landi, A. Staiano, and A. Riccio, "Spatio-temporal learning in predicting ambient particulate matter concentration by multi-layer perceptron," *Ecological informatics*, vol. 49, pp. 54–61, 2019.

[4] L. Cifuentes, J. Vega, K. Köpfer, and L. Lave, "Effect of the fine fraction of particulate matter versus the coarse mass and other pollutants on daily mortality in Santiago, Chile," *Journal of the Air and Waste Management Association*, vol. 50, no. 8, pp. 1287–1298, 2000.

[5] Y. Collette and P. Siarry, *Multiobjective Optimization: Principles and Case Studies*. Springer Berlin Heidelberg, 2004.

[6] M. Dash and H. Liu, "Feature selection for classification," *Intelligent Data Analysis*, vol. 1, pp. 131–156, 1997.

[7] K. Deb, *Multi-objective optimization using evolutionary algorithms*. Wiley, 2001.

[8] J. Durillo and A. Nebro, "Jmetal: a Java framework for multi-objective optimization," *Avances in Engineering Software*, vol. 42, pp. 760–771, 2011.

[9] M. ElAlamil, "A filter model for feature subset selection based on genetic algorithm," *Knowledge-Based Systems*, vol. 22, no. 5, pp. 356–362, 2009.

[10] C. Emmanouilidis, A. Hunter, J. Macintyre, and C. Cox, "A multi-objective genetic algorithm approach to feature selection in neural and fuzzy modeling," *Evolutionary Optimization*, vol. 3, no. 1, pp. 1–26, 2001.

[11] J. Galloway, F. Dentener, D. Capone, E. Boyer, R. Howarth, S. Seitzinger, G. Asner, C. Cleveland, P. Green, E. Holland, D. Karl, A. Michaels, J. Porter, A. Townsend, and C. J. Vöosmarty, "Nitrogen cycles: Past, present, and future," *Biogeochemistry*, vol. 70, pp. 153–226, 2004.

[12] J. García-Nieto, E. Alba, L. Jourdan, and E. Talbi, "Sensitivity and specificity based multiobjective approach for feature selection: Application to cancer diagnosis," *Information Processing Letters*, vol. 109, no. 16, pp. 887–896, 2009.

[13] N. Gilbert, M. Goldberg, B. Beckerman, J. Brook, and M. Jerrett, "Assessing spatial variability of ambient nitrogen dioxide in montreal, canada, with a land-use regression model," *Journal of the Air & Waste Management Association*, vol. 55, no. 8, pp. 1059–1063, 2005.

[14] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of Machine Learning Research*, no. 3, pp. 1157–1182, 2003.

[15] M. Hall, "Time series analysis and forecasting with WEKA," 2014, last accessed: May, 2019. [Online]. Available: *https://wiki.pentaho.com*

[16] S. Henderson, B. Beckerman, M. Jerrett, and M. Brauer, "Application of land use regression to estimate long-term concentrations of traffic-related nitrogen oxides and fine particulate matter," *Environmental science & technology*, vol. 41, no. 7, pp. 2422–2428, 2007.

[17] G. Hoek, R. Beelen, K. D. Hoogh, D. Vienneau, J. Gulliver, P. Fischer, and D. Briggs, "A review of land-use regression models to assess spatial variation of outdoor air pollution," *Atmospheric environment*, vol. 42, no. 33, pp. 7561–7578, 2008.

[18] P. Holnicki, M. Tainio, A. Kałuszko, and Z. Nahorski, "Burden of mortality and disease attributable to multiple air pollutants in Warsaw, Poland," *International Journal of Environmental Research and Public Health*, vol. 14, no. 11, 2017.

[19] C. Holt, "Forecasting seasonals and trends by exponentially weighted moving averages," *International Journal of Forecasting*, vol. 20, no. 1, pp. 5–10, 2004.

[20] J. Huang, Y. Cai, and X. Xu, "A hybrid genetic algorithm for feature selection wrapper based on mutual information," *Pattern Recognition Letters*, vol. 28, no. 13, pp. 1825–1844, 2007.

[21] H. Ishibuchi and T. Nakashima, "Multi-objective pattern and feature selection by a genetic algorithm," in *Proc. of the Genetic and Evolutionary Computation Conference*, 2000, pp. 1069–1076.

[22] F. Jiménez, R. Jódar, M. Martín, G. Sánchez, and G. Sciavicco, "Unsupervised feature selection for interpretable classification in behavioral assessment of children," *Expert Systems*, vol. 34, no. 4, pp. 1–15, 2017.

[23] F. Jiménez, J. Kamínska, E. Lucena-Sánchez, J. Palma, and G. Sciavicco, "Multi-objective evolutionary optimization for time series lag regression," in *Proc. of the 6th International Conference on Time Series and Forecasting (ITISE 2019)*, 2019, pp. 373 – 384.

[24] F. Jiménez, C. Martínez, E. Marzano, J. Palma, G. Sánchez, and G. Sciavicco, "Multiobjective evolutionary feature selection for fuzzy classification," *IEEE Transactions on Fuzzy Systems*, vol. 27, no. 5, pp. 1085–1099, 2019.

[25] F. Jiménez, G. Sánchez, J. García, G. Sciavicco, and L. Miralles, "Multi-objective evolutionary feature selection for online sales forecasting," *Neurocomputing*, vol. 234, pp. 75–92, 2017.

[26] J. Kamińska, "A random forest partition model for predicting NO_2 concentrations from traffic flow and meteorological conditions," *Science of the Total Environment*, vol. 651, pp. 475–483, 2019.

[27] L. Knibbs, A. Cortés, B. Toelle, Y. Guo, L. Denison, B. Jalaludin, G. Marks, and G. Williams, "The australian child health and air pollution study (ACHAPS): A national population-based cross-sectional study of long-term exposure to outdoor air pollution, asthma, and lung function," *Environment International*, vol. 120, pp. 394 – 403, 2018.

[28] S. Krause, D. Merrion, and G. Green, "Effect of inlet air humidity and temperature on diesel exhaust emissions," *SAE Technical Papers*, 1973.

[29] P. Leighton, *Photochemistry of air pollution*. Academic Press, 1961.

[30] T. Mar, G. Norris, J. Koenig, and T. Larson, "Associations between air pollution and mortality in Phoenix, 1995-1997," *Environmental health perspectives*, vol. 108, no. 4, pp. 347–353, 2000.

[31] W. Ni, "A review and comparative study on univariate feature selection techniques," Ph.D. dissertation, University of Cincinnati, 2012.

[32] P. G. Nieto, F. S. Lasheras, E. García-Gonzalo, and F. de Cos Juez, "Pm10 concentration forecasting in the metropolitan area of oviedo (Northern Spain) using models based on SVM, MLP, VARMA and ARIMA: A case study," *Science of The Total Environment*, vol. 621, pp. 753–761, 2018.

[33] G. Pappa, A. Freitas, and C. Kaestner, "Attribute selection with a multi-objective genetic algorithm," in *Proc. of the 16th Brazilian Symposium on Artificial Intelligence*, 2002, pp. 280–290.

[34] R. D. Peng, F. Dominici, and T. A. Louis, "Model choice in time series studies of air pollution and mortality," *Journal of the Royal Statistical Society. Series A: Statistics in Society*, vol. 169, no. 2, pp. 179–203, 2006.

[35] L. Poulos, A. Kvanli, and R. Pavur, "A comparison of the accuracy of the box-jenkins method with that of automated forecasting methods," *International Journal of Forecasting*, vol. 3, pp. 261–267, 1987.

[36] J. Schwartz, "Lung function and chronic exposure to air pollution: A cross-sectional analysis of NHANES II," *Environmental research*, vol. 50, no. 2, pp. 309–321, 1989.

[37] S. Shi, P. Suganthan, and K. Deb, "Multiclass protein fold recognition using multiobjective evolutionary algorithms," in *Proc. of the 2004 Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, 2004, pp. 61–66.

[38] W. Siedlecki and J. Sklansky, "A note on genetic algorithms for large-scale feature selection," in *Handbook of Pattern Recognition and Computer Vision*, C. Chen, Ed. World Scientific, 1993, pp. 88–107.

[39] Y. Son, A. Osornio-Vargas, M. O'Neill, P. Hystad, J. Texcalac-Sangrador, P. Ohman-Strickland, Q. Meng, and S. Schwander, "Land use regression models to assess air pollution exposure in mexico city using finer spatial and temporal input parameters," *Science of the Total Environment*, vol. 639, pp. 40–48, 2018.

[40] H. Vafaie and K. D. Jong, "Genetic algorithms as a tool for feature selection in machine learning," in *Proc. of the 4th Conference on Tools with Artificial Intelligence*, 1992, pp. 200–203.

[41] P. Winters, "Forecasting sales by exponentially weighted moving averages," *Management Science*, vol. 3, no. 6, pp. 324–342, 1960.

[42] I. Witten, E. Frank, and M. Hall, *Data mining: practical machine learning tools and techniques, 3rd Edition*. Morgan Kaufmann, Elsevier, 2011.

[43] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 2, pp. 44–49, 1998.