Environmental Modelling & Software 83 (2016) 47–57

Contents lists available at ScienceDirect

Environmental Modelling & Software

journal homepage: www.elsevier.com/locate/envsoft

Lazy Learning based surrogate models for air quality planning

Claudio Carnevale, Giovanna Finzi, Anna Pederzoli, Enrico Turrini*, Marialuisa Volta

DIMI, University of Brescia, Via Branze 38, 25123 Brescia, Italy

ARTICLE INFO

Article history:
Received 25 May 2015
Received in revised form
22 January 2016
Accepted 25 April 2016

Keywords:
Air quality
Surrogate models
Lazy Learning
Design of experiment

ABSTRACT

Air pollution in atmosphere derives from complex non-linear relationships, involving anthropogenic and biogenic precursor emissions. Due to this complexity, Decision Support Systems (DSSs) are important tools to help Environmental Authorities to control/improve air quality, reducing human and ecosystems pollution impacts. DSSs implementing cost-effective or multi-objective methodologies require fast air quality models, able to properly describe the relations between emissions and air quality indexes. These, namely surrogate models (SM), are identified processing deterministic model simulation data. In this work, the Lazy Learning technique has been applied to reproduce the relations linking precursor emissions and pollutant concentrations. Since computational time has to be minimized without losing precision and accuracy, tests aimed at reducing the amount of input data have been performed on a case study over Lombardia Region in Northern Italy.

1. Introduction

According to World Health Organization (WHO, 2008), particulate matter (PM) is the pollutant that most affects population's health. PM pollution can cause acute and chronic diseases to respiratory (asthma, bronchitis, allergies, tumors) and cardiovascular systems (worsening of cardiac symptoms) (Seaton et al., 1995). In addition to the effects on human health, PM can also affect ecosystems by interfering with photosynthesis and inhibiting the exchange of CO_2 with atmosphere. Anthropogenic emissions of various precursors interact with biogenic (natural) emissions, between each other, with local meteorology and on emissions and meteorology of surrounding areas, resulting in time and space varying PM concentration levels. The complexity of the problem increases further due to the nonlinearity of the physical and chemical reactions involving precursors such as volatile organic compounds (VOC), nitrogen oxides (NOx), ammonia (NH3), primary particulate matter (PM) and sulfur dioxide (SO2), leading to the formation of so-called 'secondary particulate matter'. These extremely intricate dynamics are extensively represented by the complexity of physically distributed multiphase models (Zannetti (2003)).

To deal with the complexity of the issue, decision makers need tools to help them planning the decisions required to improve air quality. That is why decision support systems (DSS) have been developed to guide towards political emission reduction strategies, in order to minimize, at the same time, atmospheric particulate matter concentrations and the costs related to reduction policies.

One of the main challenges in the formalization and solution of an air quality planning problem is the description of the link between precursor emissions and PM concentrations. This relationship can be simulated by means of multiphase deterministic 3D modeling systems, describing chemical and physical phenomena involved in pollutant formation and accumulation (Sokhi et al. (2006), Cuvelier et al. (2007); Carnevale et al. (2008a); Finzi et al. (2000)). These models are usually applied to evaluate the effects of given emission reduction measures (Carnevale et al., 2008b). They are however unsuitable to solve the inverse problems such as determining a set of measures to reduce an indicator below a prescribed level at minimum cost or where to invest more effectively to achieve the maximum air quality improvement within a given budget. Finding a solution to these problems requires the application of cost-effectiveness (Mediavilla-Sahagún and ApSimon (2003); Carlson et al. (2004)) or multi-objective approaches (Pisoni et al. (2008)). The first of these approaches is based on the minimization of an air pollution index, considering the cost as a constraint, while, in the second one, an objective function, composed by an air pollution index and a cost index, must be minimized by varying the application rates of defined emission reduction measures. These approaches require thousands of model runs to iteratively evaluate the impact of decision changes on air

* Corresponding author.
E-mail address: enrico.turrini@unibs.it (E. Turrini).

http://dx.doi.org/10.1016/j.envsoft.2016.04.022
1364-8152/

quality and thus, using the above mentioned 3D air quality models would require an unacceptable amount of time. That is why the identification of surrogate models, synthesizing the precursor emission-PM concentration relationship, is needed (Castelletti et al. (2012)). In the literature, source-receptor functions have been described by means of ozone isopleths (Shih et al., 1998), or with reduced form models such as (a) simplified photochemical models, applying semi-empirical relations calibrated with experimental data (Venkatram et al., 1994), and (b) statistical models, identified through the results of 3D Chemical Transport Models (Friedrich and Reis (2000), Ryoke et al. (2000); Guariso et al. (2004)).

The most common approach, on a continental and national scale, is to describe the air quality indexes using linear models (e.g. Clappier et al. (2015); Fruergaard et al. (2010); Schöpp et al. (1998)). At a regional and local level, because of the impact of the non-linearities involved, also non-linear models have been applied (Guariso et al. (2004)), among which are included models based on Artificial Neural Networks (e.g. in Corani (2005); Carnevale et al. (2012); Pisoni et al. (2008)). Since LL has the ability to reproduce both linear and non-linear relations, it can be applied to the problem as a suitable and flexible technique. To emphasize the difference between this approach and traditional methods, with regard to the type of data representation of the phenomenon under study, it is used to refer to Lazy Learning with the expression "memory-based learning", opposed to "model-based methods". Another criterion by which it is possible to distinguish the traditional methods from Lazy Learning, concerns the concept of the model. If traditional methods reconstruct, from a series of data, a global model of the given function, the Lazy Learning approach goes beyond the concept of locality and provides no explicit representation of the function, but it takes the form of an algorithm designed to extract a prediction from an example database. In the first case, we could say that the goal is the estimation of a function, while, in the second case, the estimation is limited to the result of a function in a very specific point.

The main objective of this work is therefore to identify surrogate models for air quality control, that are able to correctly describe the relationship between emissions and air quality indexes on a regional scale. Among the models that can be applied in this context, there are artificial neural networks (ANNs), used to describe the non-linear relations between the control variables and the indexes of pollution (Carnevale et al., 2012), and Lazy Learning (LL), which is a learning technique involving the use of polynomial models whose parameters vary on the basis of the input values for each air quality index evaluation request. This technique has already been applied in Corani (2005) to the Lombardy Region domain, next to artificial neural networks, for air quality prediction. Another work dealing with the identification of surrogate models for Lombardy Region, for air pollution forecasting, can be found in Pisoni et al. (2009). This paper is structured as follows: in the first section, Materials and Methods adopted are described and Lazy Learning technique is introduced; then input and output datasets are presented and, finally, a description of the tests performed to evaluate the performances of Lazy Learning models is given for a case study and the results obtained are shown. Finally some conclusions are drawn.

2. Materials and methods

2.1. The decision support system

The decision problem for which LL surrogate models have been developed, can be formalized as a multi-objective problem in which, in a given domain, an **Air Quality Index (AQI)**, representing the impacts of emission reduction measures and their implementation **costs (C)** should be minimized, while satisfying a set of **constraints** (Carnevale et al. (2008c)). Due to the non-linear relationships linking the precursors emissions to the pollutant concentrations (and related AQI), the problem is both non-linear and bi-objective. So, this can be formalized as a non-linear multi-objective optimization problem as follows:

$$\min_{\theta\in\Theta} J(\theta) = \min_{\theta\in\Theta}[AQI(E(\theta)), ICI(\theta)] \quad (1)$$

where:

- AQI is an Air Quality Index, depending on an emission scenario ($E(\theta)$);
- ICI is the Internal Cost Index, namely the policy cost;
- θ are the decision variables of the problem. θ values represent the application rates of the different measures considered;
- Θ is the set of applicable θ values.

In this context, a fast and accurate computation of the AQI with respect to emission changes is a key problem. Thus, the focus of this work is to apply LL to compute the relationship linking, in particular, PM_{10} concentrations and PM_{10} precursors (NO_x, VOC, NH_3, SO_2, primary PM_{10}) emissions.

2.2. Lazy Learning

Lazy Learning indicates the name of a family of learning methods that differ from traditional ones because they miss a real model identification phase distinguished from a validation phase (Birattari et al. (1999); Corani (2005)). A lazy method retards the parameter estimation until an output value computation is required. The demand is met by locally interpolating a set of examples according to a measure of distance. Each evaluation, therefore, requires a local procedure in the data space, which is composed of a structural identification phase and the parameter identification itself. The structural identification includes, among other things, the selection of a local approximation family, of a metric, which is a criterion used to evaluate the most significant examples, and the bandwidth, which indicates the region size in which the data is correctly modeled by members of the chosen approximation family. This phase can be performed statically at the beginning of the work or dynamically during the computation. The parameter identification consist in the optimization of the local approximation parameters.

The size of the region (centered in the value of interest), which is considered for the estimation of the local model, is called bandwidth. The larger is the bandwidth, the greater is the number of considered examples. The bandwidth choice has to be carried out together with the function family choice aiming to produce an estimate as uniform as possible, avoiding the introduction of an excessive distortion, and that is therefore able to grasp the relationship between the input and the output.

2.2.1. The family of local approximations

To implement Lazy Learning, as stated before, a function belonging to a parametric family is required to approximate the regression function. The local approximation must be linear in the parameters, allowing a faster identification and enabling the use of instruments applicable only on linear models. Generally, an approximation family of p-degree polynomials is considered, thus the choice of the family is reduced to the choice of the polynomial degree. This choice, is affected by the trade-off between bias and variance. With the same number of samples available for identification, polynomials of high degree are able to adapt better to the