Sequential Feature selection in a multi-objective optimization problem

Claudio Carnevale, Giovanna Finzi, Enrico Pisoni, Marialuisa Volta

Dipartimento di Ingegneria dell’Informazione, Università degli Studi di Brescia, via Branze 38, 25123 Brescia, Italy

Abstract: The design of air quality plans requires the development and the application of Decision Support System (DSS) to assess both the impact of emission reduction strategies on pollution indexes and the costs of such emission reductions. The problem can be formalized as a multi-objective mathematical program, integrating local pollutant-precursor models and the estimate of emission reduction costs. Both aspects present several complex elements. In particular the exposure index can be assessed by means of deterministic complex modelling systems with high computational costs that can’t be used during the optimization procedure. The most valuable solution to this issue consists in the identification of neural network models linking precursor emissions and the air quality index starting from the results of a limited number of deterministic model simulations. One of the main points in the identification procedure is the selection of the input and consequently the neural network architecture. In this paper an interactive input selection technique is applied to identify the neural network linking the PM10 (particulate matter with size smaller than 10 $\mu$m) concentration to the emission of its precursor. The methodology has been applied to Northern Italy for PM10 exposure control. The results stress that high performance can be obtained with a very limited set of input data both in terms of neural network performances and optimization robustness.

Keywords: Air pollution, Nonlinear system, Environmental engineering, Neural networks.

1. INTRODUCTION

In recent years tropospheric secondary pollution (namely PM10 and ozone) episodes have become more and more critical all over Europe. The cause-effect relations between secondary pollution and its precursors (VOC and NOx for ozone; primary PM10, VOC, NOx, NH3, SOx for PM10) are very numerous, complex and non-linear. The European Union suggests regulatory Agencies to investigate the air quality control problem at the mesoscale in order to better interpret the local air pollution situation and to define plans assessing population and ecosystem exposure and emission reduction costs.

The reduction measures selection can be formalized as a multi-objective problem where the air quality and the emission abatement costs are considered (Finzi and Guariso [1992], Schöpp et al. [1999], Shi et al. [1998], Friedrich and Reis [2000], Guariso et al. [2004], Carnevale et al. [2007]). The multiobjective approach requires to include the complex non linear dynamics of secondary pollution formation within the optimization problem formulation.

The emission-exposure relationship can be simulated by deterministic 3D modelling systems, describing transport and chemical phenomena. Such models have so high computational costs that they are not of practical application in a multi-objective mathematical program. The identification of simplified models synthesizing the relationship between the precursor emissions and secondary pollutant concentrations is required.

In recent years, this relationship has been computed using model reduction/mete-modelling techniques (Kleijnen [2008] Queipo et al. [2005] Simpson et al. [2001]), and in particular by means of statistical models identified by processing the results of complex 3D transport-chemical models (Schöpp et al. [1999]; Friedrich and Reis [2000], Barazzetta et al. [2002], Volta [2003]). Due to heavy non-linearity in the relationship, artificial neural network has been confirmed as one the most valuable approach to this model reduction problem [Carnevale et al., 2009]. Therefore, one of the main issue to be addressed in this frame is the selection of the input variable to considered in order to achieve a suitable estimation of the function linking emission and concentration without data redundancy. In this work, a Feature Selection Technique (Hejazi and Cai [2009], Tikka [2009], Fernando et al. [2009]) has been applied to assess the simplified model input and the impact on the estimation capability of the model and on the results of the PM10 control problem has been evaluated on a case study in Northern Italy.

2. PROBLEM FORMULATION

The PM10 decision model is formalized as a two-objective optimization problem, including the effectiveness of emission reduction policies on an Air Quality Index (AQI) and their costs (RC). The problem can be formalized as follows:
\[
\min_{\theta} J(E(\theta)) = \min_{\theta}[AQI(E(\theta)) \cdot RC(E(\theta))] \\
\theta \in \Theta
\]

where \( E \) represents the precursor \((NH_3, NOx, PM_{10}, PM_{2.5}, SOx, \text{and } VOC)\) emissions, \( \theta \) are the decision variables, namely the emission reductions, constrained to assume values in \( \Theta \), \( AQI(E(\theta)) \) is the air quality objective and \( RC(E(\theta)) \) are the reduction costs, both depending on precursor emissions and emission reductions. In this section the formalization of the the control variables, the air quality and the cost objectives are presented.

The emissions \( E \) are expressed, for each considered precursor, with respect to a reference situation and split into 11 source sectors according to the CORINAIR classification [EMEP/CORINAIR, 1999].

The next sections, will focus on the computation of \( AQI \) and in particular in the selection of the minimum subset of input needed to perform this computation.

### 2.1 Air quality objective

The full description of the relationship between PM10 and its precursors should be given by the application of deterministic 3D modeling systems; however these models are not of practical use in an optimization problem due to their high computational requirements. For this reason simplified source-receptor models have been identified through the processing of a limited number of simulations performed by a deterministic modeling system.

#### Air quality index

The air quality index is defined as the PM\(_{10}\) exposure index over a grid domain and it is a function of emissions. The emissions are expressed with respect to a reference scenario and split into the CORINAIR macrosectors [EMEP/CORINAIR, 1999]. Since a regional Authority can impose different reduction to different emission macrosectors, the air quality index \((AQI)\) can be expressed stressing the emission dependence of the exposure index function \((\Psi)\) for cell \((i, j)\), as follows:

\[
AQI(E(\theta)) = \Psi \left( E_{i,j}^p(\theta^{p,s}) \right)
\]

where \( E_{i,j}^p \) is the emission of the \( p \) precursor species for macrosector \( s \) in the cell \((i, j)\);

#### Deterministic approach

PM10 concentrations are typically simulated by three-dimensional deterministic modeling systems. In this work the Gas Aerosol Modeling Evaluation System (GAMES) [Volta and Finzi, 2006] has been used. It consists of three main modules as shown in Figure 1: (a) the multi-phase Eulerian 3D model TCAM [Carnevale et al., 2008]; (b) the meteorological pre-processor PROMETEO; (c) the emission processor POEM-PM [Carnevale et al., 2006]. The general mass balance equation for a generic pollutant \( h \), whose concentration is \( C_h \) \([\mu g \ m^{-3}]\) (in spatial-temporal coordinates \( x, y, z, t \)), is then given by:

\[
\frac{\partial C_h(x, y, z, t)}{\partial t} = T_h + R_h + D_h + S_h \\
\quad h = 1, 2, \ldots, n_{species}
\]

![Fig. 1. The GAMES modeling system.](image)

where:

- \( T_h \) is the transport term \([\mu g \ m^{-3} \cdot s^{-1}]\);
- \( R_h \) is the reaction term \([\mu g \ m^{-3} \cdot s^{-1}]\);
- \( D_h \) is the deposition term \([\mu g \ m^{-3} \cdot s^{-1}]\);
- \( S_h \) is the source (emissions) term \([\mu g \ m^{-3} \cdot s^{-1}]\).

Equation 3 written for all considered species is the basis for the development of the air quality models. More details about the model can be found in [Carnevale et al., 2008]. Input and output of the deterministic model simulations are then used to identify source-receptor model, that implements the Air Quality Index in the optimization procedure.

#### Source-receptor approach

As previously stated, the function linking precursor emission levels to PM10 concentration has been here estimated through stochastic models formalized by means of neural networks, identified using deterministic model simulation scenarios (see Section 3.1 for more details). In particular, the feed-forward neural network (Figure 2) has been used in this study. This network computes a vector function \( f_{NN} : \mathbb{R}^Q \rightarrow \mathbb{R}^L \) where \( Q \) and \( L \) are the dimensions of the net input and output vectors of the net respectively; the \( l \)-th element of the vector function \( f_{NN} \) is defined as \((M\) is the number of the neurons in the hidden layer):
Fig. 2. Feed-forward neural network scheme.

dimensionality of the neural network input data. This procedure aims at the definition of a subset of the initial collection of input data, able to provide the identification procedure non-redundant data by selecting only (first) the features that give the model the majority of information content in the input set. In this way an a high performance neural network can be efficiently identify. In particular, a stepwise forward selection method has been applied [Bowden et al., 2005]. More in details, the algorithm has 2 main components:

- a criterion function, which the algorithm have to minimize;
- a minimization algorithm, that starting from the null set selects step by step a features or a group of feature that can be added to the input set. Since each evaluation of the criterion function needs the identification of at least a different neural network, an exhaustive exploration of all the input set is unfeasible. For this reason, the minimization algorithms move only in one direction, selecting each time the best input from the remaining elements of the starting dataset.

2.3 Cost objective

The cost objective of PM$_{10}$ control can be formulated as follows:

$$ RC(\theta) = \sum_p \sum_s RC^{p,s}(E^{p,s}(\theta^{p,s}), u^{p,s}(\theta^{p,s})) $$

(6)

where:

- $RC^{p,s}$ represents the total cost associated to reduction of precursor $p$ in macrosector $s$;
- $E^{p,s}$ is the total annual emission of the $p$ precursor species for macrosector $s$ in the reference case;
- $u^{p,s}$ represent the cost functions, that link emission reductions and unit cost, for each $p$ precursor species and macrosector $s$.

The cost functions are polynomial cost curves, identified starting from the RAINS dataset [Amann et al., 2004] and encompassing all the technologies available to reduce pollutant emissions in every European country. More details about the methodology to derive cost objective are provided in Carnevale et al. [2007].

3. CASE STUDY

The methodology has been applied to Northern Italy (Figure 3), a region affected by PM$_{10}$ levels often exceeding the EU limit value established for health protection. In particular, high concentrations characterize the central part of the domain, where the most important industrial and residential areas are located. The winter months of 2004 have been selected as basecase for the simulation with the deterministic model.

Fig. 3. Study domain (with orography and road network).

Table 1. Reductions applied to PM$_{10}$ precursors emissions, with respect to Quitsat base-case, for the 10 reduced emissions scenarios.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>NO$_2$</th>
<th>VOC</th>
<th>PM</th>
<th>SO$_2$</th>
<th>NH$_3$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>30.89%</td>
<td>27.26%</td>
<td>21.45%</td>
<td>26.70%</td>
<td>35.85%</td>
</tr>
<tr>
<td>2</td>
<td>61.78%</td>
<td>54.52%</td>
<td>42.90%</td>
<td>53.40%</td>
<td>71.70%</td>
</tr>
<tr>
<td>3</td>
<td>61.78%</td>
<td>27.26%</td>
<td>21.45%</td>
<td>26.70%</td>
<td>35.85%</td>
</tr>
<tr>
<td>4</td>
<td>30.89%</td>
<td>54.52%</td>
<td>21.45%</td>
<td>26.70%</td>
<td>35.85%</td>
</tr>
<tr>
<td>5</td>
<td>30.89%</td>
<td>27.26%</td>
<td>42.90%</td>
<td>26.70%</td>
<td>35.85%</td>
</tr>
<tr>
<td>6</td>
<td>30.89%</td>
<td>27.26%</td>
<td>21.45%</td>
<td>53.40%</td>
<td>35.85%</td>
</tr>
<tr>
<td>7</td>
<td>30.89%</td>
<td>27.26%</td>
<td>21.45%</td>
<td>26.70%</td>
<td>71.70%</td>
</tr>
<tr>
<td>8</td>
<td>30.89%</td>
<td>54.52%</td>
<td>21.45%</td>
<td>53.40%</td>
<td>35.85%</td>
</tr>
<tr>
<td>9</td>
<td>61.78%</td>
<td>54.52%</td>
<td>21.45%</td>
<td>53.40%</td>
<td>71.70%</td>
</tr>
<tr>
<td>10</td>
<td>61.78%</td>
<td>27.26%</td>
<td>42.90%</td>
<td>26.70%</td>
<td>35.85%</td>
</tr>
</tbody>
</table>

3.1 Air quality objective

The Air Quality Index, representing the value of PM$_{10}$ concentration over the grid domain, has been evaluated through a source-receptor model based on a feed-forward neural network identified and validated by processing the results of the deterministic model TCAM, performed in the frame of Quitsat project [Di Nicolantonio et al., 2009] over a 640x410km$^2$ domain in Northern Italy. The TCAM simulations have been performed with an horizontal resolution of 10x10km$^2$. The datasets are referred to 11 scenarios: the basecase, assumed as the reference scenario, and 10 scenarios computed by reducing the precursors emissions. Table 1 presents the percentage of emissions reductions with respect to the base case.

Starting from these data, the target of the source-receptor models is to accurately estimate the domain average of PM$_{10}$ concentration given the precursors total emissions over the same period. One single neural network has been identified for the whole domain.

In more details, the output of the neural network is, for each domain cell, the PM$_{10}$ concentration computed by TCAM model. In order to take into account the contribution of the emission in the different direction, the starting dataset for the feature selection procedure includes the total emission for all the PM$_{10}$ precursors over 4 surrounding triangular-shaped zones, defined as described by Figure 4. The dimension of the triangle is selected in order to better reproduce the emission-PM$_{10}$ link taking into account that the selection of a larger triangle could improve the performance during the training phase, losing in generality. Consequently, 4 input for each
of the 5 main PM10 precursor for a total of 20 input. The validation has been performed using for each of the simulated scenario the 20% of the domain cells (Figure 5). The main feature of the considered neural network is summarized in Table 2.

3.2 Feature Selection results

Once defined the identification and validation dataset, the SFS procedure performs the selection of the neural network input using as criterion function the normalized mean squared error of the network computed on the validation dataset. In order to select a network that can be actually used for the solution of the air quality control, a constraint forcing the insertion of at least one direction for each of the PM10 precursors in the input set is required.

At each step, a neural network fed by input obtained adding to the previous selected set the new candidate feature, is identified and evaluated. In order to limit the impact of local minima reachable by the backpropagation algorithm during the identification of the network, for each candidate the identification is performed 100 times starting from different initial conditions, and then the best network is considered.

Tables 3 and 4 show the number of input selected for different test cases with respect the value of normalised mean square error (Table 3) and correlation (Table 4). It can be seen that selecting only 9 from the total set of 20 initial features, the neural network achieve performances comparable to that can be obtained with 20 input. In particular, the 9 selected input in this case are, in order of selection: NH₃W, PM10₅, VOC₅, SO₂₅, NOₓ₅, NH₃E, NH₃S, PM10₅, VOC₅. It is interesting to note that PM10 and NH₃ accounts for 5 of the 9 input, and that NH₃ is picked for 3 of the 4 possible direction considered.

More in details, Table 5 presents performance comparison between the 9 and 20 input networks on the validation set. The only significant difference concerns the 95th percentile, with a slightly overestimation reproduced by the 9 input net.

![Fig. 4. Scheme representing, for a generic domain cell, the 4 areas (N, S, W, E) on which the emissions (ton/year) are considered as input of the source-receptor model.](image)

3.3 Optimization Results

In this section, the impact of the feature selection procedure over the results of the optimization problem has been

![Fig. 5. Selected cells (red square) for the validation of the PM10 source-receptor model.](image)

Table 2. Neural network structure of the PM10 source-receptor model.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. neurons in the hidden layer (M)</td>
<td>20</td>
</tr>
<tr>
<td>Activation function hidden layer (a₁)</td>
<td>logsig</td>
</tr>
<tr>
<td>Activation function output layer (a₂)</td>
<td>purelin</td>
</tr>
</tbody>
</table>

Table 3. Number of input selected as a function of normalised mean square error computed on the validation dataset for the different tests.

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
<th>Test 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.90</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>0.90 ÷ 0.92</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>0.92 ÷ 0.94</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>0.94 ÷ 0.96</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>0.96 ÷ 0.98</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>&gt; 0.98</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 4. Number of input selected as a function of correlation coefficient computed on the validation dataset for the different tests.

<table>
<thead>
<tr>
<th>Correlation coefficient</th>
<th>Test 1</th>
<th>Test 2</th>
<th>Test 3</th>
<th>Test 4</th>
<th>Test 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt; 0.90</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>0.90 ÷ 0.92</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>0.92 ÷ 0.94</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>0.94 ÷ 0.96</td>
<td>7</td>
<td>8</td>
<td>8</td>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>0.96 ÷ 0.98</td>
<td>9</td>
<td>10</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>&gt; 0.98</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 5. Performance comparison between 9 and 20 input network on the validation dataset.

<table>
<thead>
<tr>
<th>Index</th>
<th>Net9</th>
<th>Net20</th>
<th>TCAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation</td>
<td>0.975</td>
<td>0.978</td>
<td>-</td>
</tr>
<tr>
<td>NMAE</td>
<td>0.014</td>
<td>0.012</td>
<td>-</td>
</tr>
<tr>
<td>Mean (µg/m³)</td>
<td>17.28</td>
<td>17.48</td>
<td>17.63</td>
</tr>
<tr>
<td>95th Percentile (µg/m³)</td>
<td>33.16</td>
<td>33.36</td>
<td>33.19</td>
</tr>
</tbody>
</table>

Table 6. Computational time need to solve the optimization problem.

<table>
<thead>
<tr>
<th>Network</th>
<th>Comp. Time (sec)</th>
</tr>
</thead>
<tbody>
<tr>
<td>20 Input</td>
<td>129022</td>
</tr>
<tr>
<td>9 Input</td>
<td>62774</td>
</tr>
</tbody>
</table>
assessed. Two optimization problems have been solved with the two selected network structures by means of a weighted sum method [Pisoni et al., 2009], and the results are presented in Figure 6 in terms of Pareto curves. The curve computed using the 9 input network (dashed line) is overlapped to that computed with 20 input (continuous line), with differences in the PM10 level lower than 0.5 $\mu g/m^3$ for the same levels of costs. Table 6 presents the computational time needed by the algorithm in the two cases, highlighting the gain of about 35% obtained using the 9 input network. Figures 7-11 present the comparison of the results of the two different optimizations in terms of emission reductions for the 5 precursors with respect to the AQI reduction. The comparison shows that the curves are consistent, in particular for NH3 and PM10. The discrepancy is higher for NOx, VOC and SO2, but this hasn’t significant impact on the Pareto curve (Figure 6), confirming that PM10 and NH3 are the two most important input for the estimation of air quality index.

4. CONCLUSION

One of the complexities of the air quality control problem is related to the nonlinearities of atmospheric processes and to the high number of input used by model describing the link between precursor emissions and air quality concentrations. In this paper, after having presented the multi-objective optimization approach to control air quality, a sequential feature selection procedure has been applied to reduce the number of input used to feed neural network describing the air quality objective in the presented approach. The validation of ANNs obtained with different number of input show that with a limited number of input it is possible to obtain very high performances. Moreover, the optimization problem results computed with two different configuration shows that the use of a limited number of input don’t lead to significant differences in terms of optimal solutions of the decision problem with respect to the case with the full input network, allowing a large gain in terms of computational time.

ACKNOWLEDGEMENTS

The work has been developed in the frame of EU NoE ACCENT (Atmospheric Sustainability). Part of this work has been also funded in the frame of Pilot Project QUITSAT.
Fig. 11. Comparison of the VOC emission reduction computed using the 20 input (continuous line) and 9 input (dashed line) networks

(QUalita’ dell’aria mediante l’Integrazione di misure da Terra, da Satelliti e di modellistica chimica multifase e di Trasporto - contract I/035/06/0 - http://www.quitsat.it), sponsored by the Italian Space Agency (ASI). We also acknowledge APD-IIASA staff for suggestions and data sharing.

REFERENCES


