UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

DOCTORAL THESIS

Deep neural architectures and time series: a scalable system for air quality prediction and its application

Author:
Ricardo NAVARES

Supervisor:
José Luis AZNARTE, PhD

A thesis submitted in fulfilment of the requirements for the PhD. in Intelligent Systems

June 10, 2020

UNIVERSIDAD NACIONAL DE EDUCACIÓN A DISTANCIA

Abstract

Computer Science Engineering

PhD. in Intelligent Systems

Deep neural architectures and time series: a scalable system for air quality prediction and its application

by Ricardo NAVARES

[ESP]

En los últimos años se está produciendo un aumento en el interés sobre los niveles de contaminación debido a su demostrado impacto medio-ambiental y sobre la salud. Existe una relación directa entre las concentraciones de contaminantes y las afecciones respiratorias, circulatorias y cardiovasculares. El conocimiento anticipado de los niveles de polución en el aire resulta de gran interés para un amplio espectro de campos de estudio: desde ámbitos sanitarios hasta ámbitos de gestión de políticas medio-ambientales pasando por pacientes particulares que sufren algún tipo de alergia o problema respiratorio relacionado. Sumado a esto, las aplicaciones de estas predicciones son diversas: desde aplicar medidas preventivas para paliar posibles efectos dañinos, hasta la planificación y optimización de recursos para centros clínicos.

Tradicionalmente, el problema de predicción ha sido tratado a partir de modelos estocásticos para determinar las relaciones entre las medidas de una o más variables, normalmente meteorológicas, y la variable independiente a predecir. Esto conlleva el estudio y conocimiento de diversos campos como la meteorología, ciencias ambientales o biología entre otros. Dichos conocimientos no siempre están disponibles en las distintas instituciones interesadas en el uso de las predicciones. Por este motivo, una solución unificada capaz de extraer y filtrar automáticamente la información contenida en los datos necesaria para establecer predicciones de manera precisa sería de gran interés para la comunidad científica.

La línea argumental de esta tesis parte del estudio individual de los problemas de predicción de contaminantes, ya sean de origen antrópico o biológico. Este estudio conlleva la investigación de factores influyentes en dichos contaminantes así como la idoneidad de los distintos modelos de aprendizaje automático disponibles. Esta primera parte de la tesis provee el conocimiento en profundidad de los desafíos característicos de cada problema para, posteriormente, avanzar hacia el desarrollo de un sistema unificado basado en redes neuronales profundas. Una vez desarrollado dicho sistema, las predicciones resultantes son utilizadas para determinar los efectos de los contaminantes sobre la salud. Esta última tarea no es tratada como un problema independiente, sino como una extensión de la funcionalidad del sistema ya existente.

A pesar de que las técnicas de aprendizaje automático están ganando popularidad en los últimos años, en los ámbitos científicos sobre los que se aplica esta tesis hay un

predominio del uso de técnicas estadísticas clásicas. Para validar la aceptación de las propuestas incluidas por parte de la comunidad científica la línea de investigación contiene hitos basados en artículos publicados en revistas correspondientes a los campos de estudio. Este compendio de publicaciones sirve por un lado como validación de los resultados obtenidos y por otro como difusión de las posibilidades de las técnicas presentadas en este estudio a los diferentes campos científcos abarcados.

$[ENG]$

During the last decades, air quality has been in the spotlight due to its proven direct impact on the environment and human health. There is a direct relationship between the quality of the air and respiratory, circulatory and cardiovascular disorders. To be able to forecast the concentration of pollution is of great interest to a wide range of academic and practical fields such as health institutions, environmental policies management and individuals with related respiratory diseases. Moreover, predicted pollution levels enable a wide range of applications including the management of traffic and environmental factors in urban areas, the alerting of potential peaks in the admissions in clinical institutions to enable resource planning and optimization or minimizing the exposure for patients in order to prevent adverse effects.

Traditionally, the problem of predicting environmental series using observation–based models is based on a number of different methods to relate records of pollutants to one or more variables that can be measured or predicted, usually meteorological data. This implies expert field knowledge in meteorology, environmental sciences or biology among others. All this knowledge is rarely available at the same time in most research departments, neither is it the time to deepen into each specific model driver. For this reason, a unified solution able to filter and extract relevant information in order to establish precise predictions is of great interest for the scientific community.

The subject matter of this thesis starts with the study of the prediction of individual pollutants, both chemical and biological. This first step implies the research of relevant factors which drive pollutants air concentrations, as well as the availability and suitability of computational intelligence methods to perform such tasks. This first part provides a thorough knowledge of the specific challenges of each problem and builds up the foundations for implementing the aforementioned unified solution based on deep neural networks. Once the unified air quality prediction system is established, its predictions can be applied to determine pollution effects on human health by extending the system functionality to predict the number of admission due to pollution related disorders in populations.

Even though computational intelligence methods are gaining popularity, traditional statistical techniques predominate in the application fields of this thesis. In order to validate the proposals included in this thesis, the research line is compounded by a number of milestones based on published articles in journals which encompass all related scientific areas. The objective of this collection of publications is twofold: validating the obtained results and increasing the awareness of the scientific community about the methods presented, which also implies the adaptation of the concepts to domain specific language.

Acknowledgements

Me gustaría que estas líneas sirvieran para expresar mi más sincero agradecimiento al Dr. José Luis Aznarte por aceptarme para realizar este trabajo bajo su dirección, por su apoyo y confianza. Su aporte ha sido invaluable no solamente en el desarrollo del trabajo, sino también en mi formación. Las ideas propias, siempre enmarcadas en su orientación y rigurosidad, han sido claves para el desarrollo de las actividades. Le agradezco también el haberme facilitado los medios suficientes para llevar a cabo todas las actividades propuestas.

También querría agradecer a Dr. Julio Díaz y Dra. Cristina Linares por sus contribuciones y colaboraciones que, no solamente me han hecho crecer en conocimientos científicos, sino también la forma de afrontar y comunicar publicaciones específicas adaptadas al contexto del campo de estudio. A Damir Valput por la ayuda y colaboración a la hora de extender ideas para hacer este trabajo más completo.

Por último agradecer a mi madre Julia y mi padre Mariano que desafortunadamente no pudieron asistir al final de esta tesis.

Contents

Chapter 1

Introduction

1.1 Research context

The World Health Organization (WHO) estimates 7 million pollution-related deaths every year worldwide [1]. Air pollution is considered a major environmental risk to health and it is directly related to respiratory, cardiovascular and circulatory disorders according to the European Environmental Agency [40] and the review of evidence on health aspects of air pollution project released by the WHO [41].

The WHO Environment and Health Information System (ENHIS) shows that 83% of the population in European cities are exposed to particulate matter (PM) increasing the risk of cardiovascular and respiratory diseases. Similarly, exposure to nitrous oxides NO_x and sulfur oxides SO_x have both short-term and chronic health consequences for people with respiratory diseases and are considered hazardous substances by the United States Center for Disease Control and Prevention (CDC) [2]. Furthermore, climate change is becoming more evident as an indirect effect of air pollution, being carbon dioxide (CO_2) and NO_x two of the main greenhouse gases which are produced by human activity and target for reduction in the United Nations Kyoto protocol [136]. Breathing difficulties and asthma symptoms are triggered by high concentrations of ozone pollution [41]. Moreover, this situation is aggravated by the presence of atmospheric pollen concentrations, with some genus as Plantain and Grasses being two of the most common and aggressive in terms of allergic and respiratory disorders [133].

Air Quality Guidelines published by the WHO [102] establish the recommended air pollutant exposure limits to preserve public health and encourage air quality. Avoiding exposure to air pollutant levels is important for the health of individuals, especially for those who are susceptible because of their preexisting cardiovascular and respiratory diseases. Even though air quality has improved significantly since the Clean Air Act in 1970 [116], air quality is still problematic in many cities, which requires local authorities to implement environmental measures to reduce pollution and educate people about its health effects, according to the US Environmental Protection Agency (EPA). Furthermore, the Organization for Economic Cooperation and Development (OECD) estimates that the economic impact of pollution-related health effects is estimated to be 3 trillion USD in 2015 [100].

Forecasting air pollution has been of paramount importance as the basis for implementing effective pollution control measures. Air quality forecasting is an effective way to provide an early warning of pollutants in order to protect public health [135].

[1] https://www.who.int/health-topics/air-pollution

[2] https://www.cdc.gov/

This can be achieved through implementing emission control mechanisms like traffic restrictions [21, 45], and apply tighter emission controls in areas where vulnerable populations are affected. These are the reasons why extensive literature can be found about methods of air pollution forecasting which can be broadly divided into three categories: numerical forecasting methods, statistical methods and artificial intelligence.

Numerical models are based in the idea of determining pollutant dispersion processes in the atmosphere [63]. Generally, these processes are driven by convection or advection which are numerically simulated according to the conditions surrounding the pollution source which can be physical, chemical or biological. On the contrary, statistical methods are not dependent on the mechanism of the change of the process (source) in order to analyse the events [144]. However, the main limitation of statistical methods is that the characteristics of the pollutant must be specified according to its type, emission sources, series patterns, and influencing factors. Moreover, spatial variability of pollutant concentrations must be assessed in comparison of the same pollutant recorded in various locations. Former statistical methods provide scientifically sound information, statistical assessment and uncertainty of estimations. However, artificial intelligence generally performs better in terms of accuracy in most applications [59, 86, 90, 126]. Artificial intelligence methods focus on the relationships between input and output signals without the need of any prior information such as the type of pollutant or the dynamics of the inputs.

1.2 Research problem gap and hypothesis

In the framework of air quality forecasting, time series analysis is tailored to the diagnosis and prediction of the concentration of pollutants in the air, taking into account its precursors and influential meteorological variables. Despite the extensive literature and resources available, air quality forecasting methods are not always successful due to the specific characteristics of each pollutant, area of interest and the pollutants relationship to the meteorology and topography. These innate conditions of the air quality forecasting problem lead to a limitation of the methods applied and a higher uncertainty of forecasts when compared, for instance, to the problem of weather prediction [120].

Governmental agencies such as the Environmental Protection Agency (EPA) in the US and European Environment Agency (EEA) report the Air Quality Index daily (AQI) to communicate how polluted the air in certain regions is. The AQI is a scale which ranges from *Good* to *Hazardous* based on the main chemical pollutants measured by the air quality monitors. It is noticeable that this index excludes airborne pollen in its computation. However, the interaction between chemical pollutants and pollen has been demonstrated [146]. Furthermore, there is evidence [99] that shows that pollen grains found in polluted areas are not only found in greater concentrations but also have higher allergenicity.

Air quality is directly linked to air pollution which encompasses measured concentrations of several pollutants that are hazardous to health and it occurs when the natural characteristics of the atmosphere are modified by chemical, physical or biological agents. Consequently, the air quality forecasting problem requires the analysis and diagnosis of several types of pollutants, each of them with different relations to their influential factors. Therefore, the aforementioned limitations increase proportionately

to the number of pollutant types taken into account. As an example, predicting chemical pollutants and airborne pollen concentrations based on meteorological conditions are inherently two different mathematical problems: atmospheric pollen concentrations depend on plant development during previous seasons which, at the same time, depend on the climatological conditions during plant evolution [17, 129]. This implies that there is both a long and mid-term relationship between past atmospheric conditions and current plant status. Contrarily, chemical air pollutant levels are related to recent past atmospheric conditions, emission sources and meteorological conditions [93].

In addition to temporal dependencies, air pollution levels demonstrate a high degree of spatial dynamics which are closely linked to the spatial distribution of economic activities such as areas with intense traffic, as well as topography which, for instance, defines air currents. Consequently, a single monitoring station measurement can only be considered representative of a limited surrounding area. This is an important topic which is linked to the assessment of risk exposure, monitoring networks and data evaluation. Moreover, European regulations lack a clear definition about the spatial representativeness of air quality [42]. In the literature, there is no unified agreement to address this complex problem. As a consequence, cost-effective fixed monitors are not available for all pollutants and, as a result, samples of the same pollutant depend on location and show different probability distributions as they are affected by proximity to the source.

The intricacies of this problem as described lead to the vast majority of scientific studies to focus either on one type of pollutant or one specific location or, sometimes, both. In addition to these spatio-temporal dynamics, the study of air quality involves several scientific disciplines depending on the source of the pollutant, these include meteorology, environmental sciences or biology which means studies require different resources and expertise. As a result of this focused approach and lack of available resources, current solutions generate many model and pollutant-specific systems that need to be combined in order to obtain the full representation of future environmental conditions. Therefore, initiatives are being launched globally to facilitate the understanding of the multiple relationships between the physical and natural environments such as the European Commission Digital Earth [3]. This initiative not only includes conventional environmental data sources, it also contains information provided by national and local authorities as well as sources from the private sector including social networks.

In summary, apart from the temporal and spatial dimensions implicit in the problem of forecasting a single source of pollution, it is important to take into account the relations between all factors that contribute to the quality of the air. The hypothesis of this thesis states: a successful forecasting method in order to take effective control measures and, consequently protect public health, is achieved when all these problem characteristics are considered. If mutual influential relations exist between the variables as well as shared implicit information (either temporal or spatial), the information and the relations can be automatically extracted in order to obtain an optimal solution of the prediction model that can be used to forecast. If this information and the relations can be extracted, there must be a method able to cope with the increment in complexity every time a new data source, such as a new location or variable, is added to the system.

[3] `https://ec.europa.eu/jrc/en/research-topic/digital-earth`

1.3 Objectives

The objective of this research is to develop a comprehensive scalable system able to efficiently predict air quality: a system based on deep neural architectures which allows to automate tasks with minimal or no assumptions on the underlying data and easily applicable to air quality related forecasting problems. The specific objectives are detailed as follows:

1. Build computational intelligence models that succeed in capturing pollen time series dynamics, being able to forecast future concentrations.
2. Use automatic feature selection to gain insight on the inner complexity of the series and prove the ability of such models to capture it.
3. Build neural architectures that succeed in capturing the dynamics of atmospheric pollutant concentrations, being able to forecast future values.
4. Create a holistic deep neural architecture to estimate future air quality based on spatio-temporal relations.
5. Extend the deep neural architecture to forecast hospital admissions related to air quality.

1.4 Research plan, contributions and structure of the document

Air quality forecasting challenges the expertise in several scientific fields such as environmental sciences, aerobiology, meteorology, artificial intelligence, statistics and probability. This expertise is not always available in practical situations and, as shown above, it is detrimental to the scalability of current solutions with the consequent misrepresentation of actual quality of the air. The postulate of this thesis is to propose, from a purely engineering and data point of view, a system able to provide accurate future environmental scenarios which is easily scalable when a new observation station or a new type of pollutant is requested. In order to do so, the research plan is divided in 4 main phases as shown in Figure 1.1 representing the contributions to each individual problem (biometeorological and chemical pollutant time series) and the proposal of the final system and its application.

In the first research phase, we centered our attention on pollen time series dynamics which are characterised by the presence of sudden high peaks during pollination season outside of which airborne concentrations are not considered harmful for human health. The identification of these sudden peaks represents a challenge given the limited amount of observations during the study periods. These peaks also define the main pollination season in order to apply preventive measures. This phase examines the application of soft computing methods to forecast such peaks.

Through turning the problem into a classification problem, we contribute to a better definition of the pollination periods of which there is currently no consensus within the scientific community. With a proper definition of the pollination season, different forecast horizons were proposed to anticipate high levels of pollen concentrations in the air and consequently prevent risky exposures. These contributions are presented in Chapter 3 and 4 of this thesis.

FIGURE 1.1: Research plan phases and corresponding publications with their imparct factor and quartiles.

Biometeorological time series dynamics are mainly driven, among other factors, by current atmospheric conditions and the weather conditions during plant development. This implies long and short-term dependencies in order to estimate future airborne concentrations. In the second phase of the research plan we analysed the capabilities of soft computing methods to extract, from the data point of view, these dependencies in order to provide accurate estimations.

This problem is approached through the examination of potential influential meteorological and biological conditions to forecast pollen concentrations of both of the pollen genus considered in this thesis. The demonstration of this assumption establishes statistical inference tests to align the results with previous scientific contributions based on biology, phenology and meteorology. Chapter 5 and 6 prove with statistical soundness that this problem, which is commonly approached from the point of view of the aforementioned scientific fields, can be solved through computational intelligence methods and automatic feature selection.

Chemical air pollutants concentrations are related to recent emission sources either anthropogenic such as traffic or natural sources such as forest fires. This implies short-term dependencies within time series observations as they represent emission source status when no further information is available. In recent years, neural networks have achieved tremendous success in diverse application domains. In this phase of the research plan, we pose the suitability of several neural network models and their extension to solve, in one unified model, biometeorological time series as well.

Firstly, a comparison of the predictive power of a set of neural network models is performed over a chemical pollutant in Chapter 7. Subsequently, we extend the problem by including all available locations where several chemical pollutants and pollen concentrations are observed. Chapter 8 provides a comprehensive set of deep network configurations to identify which are able to better extract relevant information out of the set of time series in order to predict air quality.

We introduced in Section 1.1 the risks to health associated with poor air quality. One implication of the original hypothesis of this thesis is that an improved family of clinical models can be applied to protect public health through a successful air quality forecasting system. In this last phase we extend the functionality of the previously implemented deep neural architecture to predict air quality related health disorders represented by hospital admissions.

Chapter 9 proposes covering four model families: ensemble methods, boosting methods, artificial neural networks and ARIMA to establish a benchmark and, to prove that a dynamic combination of the predictions improves the performance with respect to the individual models. This concept is used in Chapter 10 to propose the final deep neural scalable system to extract temporal-spatial relations out of the set of time series in order to predict hospital admissions.

As a consequence of the interest captured by the results published, two extra side projects, corresponding to Chapters 11 and 12, derived from the core research as part of collaboration with the scientific community.

Chapter 2

Methods

2.1 Traditional time series approach

George E. P. Box y Gwilym Jenkins proposed the autoregressive integrated moving average models (ARIMA) to predict and analyze time series related to economic variables [12]. Since then, this sort of stochastic models has been applied to forecast the evolution of time series in different fields such as environmental atmospheric [36, 73, 101] or biological pollution [115]. Not only are ARIMA models used in forecasting, but also in determining through statistical significance the influence of independent variables on the behavior of certain dependent variable.

The ARIMA forecasting equation for a stationary time series is a linear (i.e., regression-type) equation in which the predictors consist of lags of the dependent variable and/or lags of the forecast errors. That is:

Predicted value of Y equals a constant (μ) and/or a weighted sum of one or more recent values of Y $(Y_{t-1}, \cdots, Y_{t-p})$ and a weighted sum of recent values of the errors $(e_{t-1}, \cdots, e_{t-q})$. In terms of y, the general forecasting equation is:

$$\hat{y_t} = \mu + \phi_1 y_{t-1} + ... + \phi_p y_{t-p} - \theta_1 e_{t-1} - ... - \theta_p e_{t-p}, \tag{2.1}$$

where ϕ_i is the coefficient of the autoregressive (AR) term i and θ_i represents the coefficient of the moving average (MA) term i.

Apart from the corresponding lags of the series p (Y_{t-p}), the errors (e) and its lags (e_{t-q}), exogenous variables $(X, \ldots, Z)$, which represent the environmental independent variables, were included along with their corresponding lags up to $t-s$ and $t-m$, resulting in:

$$\begin{aligned} \hat{y_t} = \mu &+ \phi_1 y_{t-1} + ... + \phi_p y_{t-p} \\ &- \theta_1 e_{t-1} - ... - \theta_p e_{t-p} \\ &+ \beta_0 X_t + \beta_1 X_{t-1} + ... + \beta_s X_{t-s} + \ldots \\ &+ \gamma_0 Z_t + \gamma_1 Z_{t-1} + ... + \gamma_m Z_{t-m} \end{aligned} \tag{2.2}$$

The value of the estimator $\beta_0, \beta_1 \ldots \gamma_0, \gamma_1 \ldots$ of the variables that are significant at $p < 0.05$ indicating increased Y to increment by one unit of each independent variable $(X, \ldots, Z)$ respectively.

2.2 Computational intelligence methods

Logistic Regression. Logistic regression is part of a broader family of generalized linear models where the conditional distribution of the response falls in some parametric family, and the parameters are set by a linear predictor. In binary logistic regression the response represents the absence or presence of a specific event, which is in this case whether the data point is over a predefined threshold or not.

The stability of the estimation of the parameters suffers when those covariate in a similar fashion. As several features were derived from others, it is likely to find dependencies between them, thus it is intended to avoid the misbehavior of the maximum likelihood parameter estimation. Thus, a ridge estimator [22] was introduced to add penalty on weights learned to avoid over-fitting.

Support Vector Machines. The current Support Vector Machines (SVM) standard algorithm, proposed by [27] in 1995, is a learning method used for binary classification which finds a hyper-plane which separates the d-dimensional data perfectly into its two classes. However, since sample data is often not linearly separable, SVM's introduces the notion of a *kernel induced feature space* which casts the data into a higher dimensional space where the data is separable. A good classifier is achieved when the hyperplane has maximum distance to the closest point of each class.

The radial basis function kernel (RBF) was used for the experiment in order to handle the nonlinear relations between the class and the features and to ease the numerical difficulties.

Random Forest. In the last years, there has been a growing interest in ensemble learning which aggregates the results of several independent models selected to boost their predictive performance. A well-known method is called *bagging* or bootstrap aggregating, proposed by [13]. Subsequently, [15] presented a model called random forests (RF) which adds an additional layer of randomness to *bagging* providing robustness against overfitting with a limited number of parameters. These two characteristics favor RF against other computational intelligence methods such as neural networks.

The procedure combines several randomized regression trees generated over sample fractions of the data, and aggregates their prediction by averaging. This averaging process mitigates the influence of outlier data points giving RF advantage over other common methods as support vector regression, which are highly sensitive in presence of outliers. As opposed to classification trees, the optimal split condition is the variance, which at the same time, is used to compute a measure of the importance of the independent variables.

The importance of a variable is estimated by measuring the increases in prediction error or variance when data from that variable is randomly permuted while the rest are left unchanged. The underlying idea is that if the variable is not important, then rearranging the values of that variable will not degrade prediction accuracy. For each tree, the prediction error is recorded before and after the permutation, the difference between both errors is averaged over all trees and normalized, thus providing the relative importance. The bigger the difference, the higher the importance of the permuted variable.

RF and Logistic Regression (LR) make different assumptions about the data and have different rates of convergence. On the one hand, RF assumes that the decision boundaries are parallel to the axes based on whether a feature is $\geq, \leq, <$ or $>$ to certain value so the feature space is chopped into hyper-rectangles. On the other hand, LR finds a linear decision boundary in any direction by making assumptions on $P(C|X_n)$ applied to weighted features so non-parallel to the axes decision boundaries are picked out. This trade off motivates to take into account SVM as an alternative.

Gradient Boosting Machines. In addition to average the combination of multiple learners, another popular ensemble technique is *boosting*. The principle behind is starting with a weak learner and turning it into a strong learner. This process is also known as *additive training*. Proposed by [46] Gradient Boosting Machines (GBM) adds sequentially new models (trees) to the ensemble, which is represented by an error function of the previous iteration fitted model. Thus, each new tree is trained with respect to the error of the whole ensemble so far. In the regression problem the error function is the classic square error (SE) which conforms the objective function to optimize.

An important concern in computational intelligence is the generalization capabilities of the models which might suffer from a non proper learning scheme and, resulting in overfitting. In order to mitigate the effects of *overfitting*, [46] proposes a technique known as *shrinkage* to control the complexity of the model. *Shrinkage* is a common regularization approach which shrinks regression coefficients to zero and consequently, reduces the impact of unstable coefficients. In the context of GBM, *shrinkage* penalizes (reduces) the importance of each tree at each consecutive step. Hence, the final objective consists of two terms, a training loss function represented by SE and the regularization which measures the complexity of the model.

Artificial Neural Networks. ANNs are a tool for modelling nonlinear processes based on the information collected by a vector named input layer, through which the information is propagated layer by layer establishing the relations between the inputs and the final layer called output layer. Intermediate or hidden layers consist of one or more units called neurons which are interconnected to the neurons of the previous and subsequent layers. The number of hidden layers and the number of neurons of each one define the *topology* of the network.

Each neuron generates an excitatory response to signals received through an activation function which can be selected among the different functions available but, following recommendations from the literature the sigmoidal activation function was chosen [10, 58].

The learning of the network is based on obtaining the relationship between the input and the output layer by comparing, via root square mean error (RMSE), network outputs with the actual values through the well-known *backpropagation* algorithm [119].

The aim is to find the network *topology* which minimizes the error. This procedure is based on a trial and error approach which, starting from a simple network of one hidden layer with few neurons, consists of increasing the capacity of the network (sequentially incrementing the number of neurons in a hidden layer as well as the number of layers) to optimize the results.

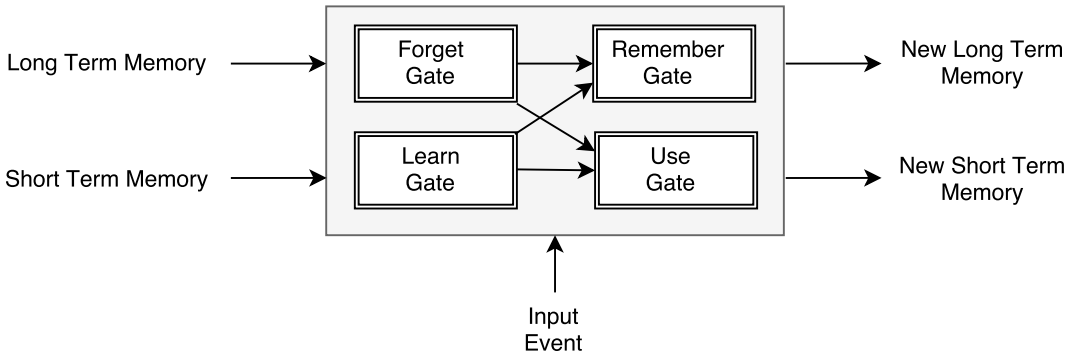


FIGURE 2.1: LSTM unit architecture.

Stacked Generalization. Proposed by [142], stacked generalization or stacking is an ensemble technique that uses a new model to learn how best to combine the predictions from two or more models with the aim of reducing error generalization. Opposed to more traditional approaches of ensemble learning such as voting or averaging, as in the case of RF, which are winner-takes-all ways of combining [142], using a meta-learner to ensemble allows to identify the circumstances under which the pooled predictions shall gain or lose weight in the final forecast.

The idea behind stacking is splitting the training set into two subsets train_a and train_b. A first stage trains the pool of selected models on train_a to create predictions for train_b and repeat using train_b for training to generate train_a predictions. As a final step of this first stage the pool of models are trained over the full training set to create predictions for the test set. The second stage consists of training the meta-learner using the training set, which contains the predictions of the models from stage one, and creating the final predictions for the test set.

2.3 Long short-term memory networks

Compared to traditional neural networks, recurrent neural networks (RNN) are implemented with loops or connections between units allowing information persistence from one step of the network to the next. The ability to map input sequences to output sequences by incorporating past context into their internal state makes them especially promising for tasks that require to learn how to use past information such as time series analysis. RNNs can be thought of as multiple copies of the same neural network, each transferring information to its successor and forming a chain-like architecture which is naturally related to sequences.

RNNs might be able to look at recent information to perform a present task which makes them suitable for time series predictions. However, relevant information might appear further in the past and, as the time gap grows, RNNs are unable to connect the information.

Long short-term memory networks (LSTM) were first introduced in 1997 by [60] and improved in 2000 by [53]. They are a variation of RNNs capable of learning long-term dependencies by including in the architecture special units called memory blocks. In addition, multiplicative units called gates (Figure 2.1) control the flow of information from a LSTM unit to another.

The learn gate. The learn gate takes the short term memory (STM) and the input event and combines them. Actually, after combining the event and the STM it ignores redundant information. Mathematically, the learn gate obtains as an input the short term memory STM_{t-1} and the event E_t and puts them into a linear function which consists on joining the vectors, multiplying it by the weight matrix W_n, adding a bias b_n and squeeze the result with a $tanh$ activation function:

$$N_t = \tanh\left(W_n \cdot [\mathrm{STM}_{t-1}, E_t] + b_n\right). \tag{2.3}$$

The new information N_t passes through the gate but still needs to ignore the information which is not relevant. In order to do so, N_t is multiplied by the ignore vector i_t. This ignore vector is calculated via a simple small neural network whose inputs are again the STM and the event and uses the sigmoid (σ) activation function to squeeze the information:

$$i_t = \sigma\left(W_i \cdot [\mathrm{STM}_{t-1}, E_t] + b_i\right), \tag{2.4}$$

being the learn gate represented as $N_t \times i_t$.

The forget gate. Takes the long term memory (LTM) and decides which parts to keep and to forget. The LTM at $t-1$ is multiplied by a forget factor f_t which is calculated through a one layer neural network with a linear function, which uses the STM at $t-1$ and the event E_t, and combines it with a sigmoid activation:

$$f_t = \sigma\left(W_f \cdot [\mathrm{STM}_{t-1}, E_t] + b_f\right) \tag{2.5}$$

being b_f the bias and W_f the weight matrix. The forget gate can be expressed as $\mathrm{LTM}_{t-1} \times f_t$.

The remember gate. Takes the output from the forget gate and the output from the learn gate and adds them to obtain the new LTM:

$$\mathrm{LTM}_t = \mathrm{LTM}_{t-1} \cdot f_t + N_t \cdot i_t \tag{2.6}$$

The use gate. It combines the LTM that just came out from the forget gate and the STM that came out from the learn gate to come out with a new STM and an output. In order to do so, it applies a small neural network on the output of the forget gate using the $\tanh$ activation function (2.7) and another neural network on the STM and the events using the sigmoid function (2.8):

$$U_t = \tanh\left(W_u \cdot \mathrm{LTM}_{t-1} \cdot f_t + b_u\right) \tag{2.7}$$

$$V_t = \sigma\left(W_v[\mathrm{STM}_{t-1}, E_t] + b_v\right) \tag{2.8}$$

As a final step, the network multiplies (2.7) and (2.8) to obtain the new output $STM_t = U_t \times V_t$ which also works as a new STM.

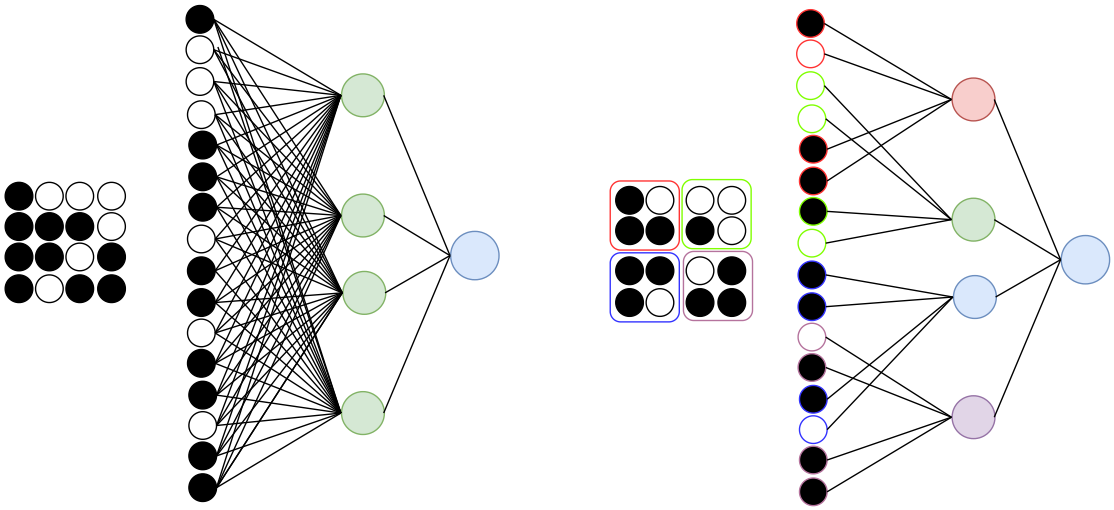


FIGURE 2.2: Fully-connected MLP (left hand) vs. Locally connected (right hand).

2.4 Convolutional neural networks

Convolutional Neural Networks (CNN) [75] have been successfully applied in several domains such as image recognition [72] or linguistics [52]. Based on their success, researches have started to use them for time series analysis [49]. CNNs differ from feed-forwards neural networks mainly by the existence of convolutional layers, which are hidden layers that utilise the power of mathematical convolution to transform inputs. Convolution allows for the encoding of the local properties of the input in such a way that propagates the information in a more efficient manner since fewer paramenters are needed (Figure 2.2).

CNN filters or kernels, obtained by the convolution of inputs and weights, are local in input space and are thus to exploit the strong, spatially local correlation present in the time series. That means they work well for identifying simple patterns within local regions of the data (subset of features) which then will be used by subsequent layers to form more complex patterns. One-dimensional CNNs share the same characteristics with the most commonly used 2-dimensional ones differing only in the dimensionality of the input and how the filter slides across the data. However, they overcome the limitation of being computationally expensive when compared to the 2-dimensional analogues.

A traditional convolution layer has too many parameters, for instance, a 3x3 convolution filter has 9 parameters which at the same time, increases by a power of 2 when filter size increases. Too many parameters not only does it take a long time to learn, but it also takes equally long to make predictions while performing inference. Since convolution consists of vector (matrix) multiplication, factorization can be used to reduce the number of parameters [25]. Thus, the 3x3 convolution kernel can be factorized to a 1x3 and 3x1 vectors which multiplied achieve the same effect and, at the same time, reduces the number of parameters to 6.

Another technique to decrease the computational power required to process the data through dimensionality reduction consists of including pooling layers to reduce the spatial size of the convolved feature. However, the pooling layer loses positional

information about the different objects [122]. This is why many new architectures have stopped using the pooling layer altogether.

2.5 Non-parametric Friedman test

The Friedman test [47] is a non-parametric analogue of the parametric two-way ANOVA. The objective of the application of the test is to determine if there is a difference among model performances over different data sets and consequently, whether one (or more) is consistently better than the others.

Parametric tests have been commonly used in the analysis of experiments. When comparing the differences between more than two related sample means, the common statistical method is the repeated-measures ANOVA. However, parametric tests require some conditions, such as normality and symmetry in data distributions, which are not fulfilled in this case study.

Given that non-parametric hypothesis tests are applied to nominal or ordinal data, the original computed root mean squared error (RMSE) of each model over each data set is converted to its correspondent rank within the set and combined by averaging: $R_j = \frac{1}{n}\sum_i r_i^j$ (where j denotes the model, i refers to each data set and n is the total number of pairs $\{\text{model}, \text{dataset}\}$). Since the error is being used to compare the models, the highest rank 1 will be assigned to the highest error, thus the worst performer. The null hypothesis of equality of medians is tested by the F-statistic

$$F = \frac{12n}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right] \tag{2.9}$$

where k is the number of algorithms and $F \sim \chi^2_{k-1}$. Still, this test is not sufficient as it only indicates the presence of significant differences in the whole model performance space. A ranking conversion is computed to obtain the p-value of each pair [26]. The former is a valid procedure to compare two models but is not suitable for multiple comparison as there is no control of error propagation (Type I errors) when making more than one comparison.

Thus, once the existence of significant differences in the group of models is evidenced, a post-hoc test adjusts the value of the significance level α at each pairwise comparison to allow multiple comparisons. [62] proposed the adjustment by selecting the p-values of each test, starting with the most significant p_i, and test the hypothesis of $H_i : p_i > \alpha/(k-i)$, being k the total number of models in our proposal. If H_i is rejected then allows to test H_{i+1}, being p_{i+1} the next most significant p-value and so on. An extension of this step-down method was proposed by [125], which uses a logical relation between the combination of the hypotheses of all pairwise comparisons. For instance, if a model a_1 is better/worse than a_2, it is not possible that a_1 is as good/bad as a_3 and a_2 has the same performance as a_3. Based on this argument and following Holm's method, instead of rejecting $H_i : p_i \leq \alpha/(k-i)$, rejects $H_i \leq \alpha/t_i$, being t_i the maximum number of hypotheses which can be true given the number of false hypotheses in $j \in \{1, ..., i\}$.

Chapter 3

Forecasting the Start and End of Pollen Season in Madrid

Type:	Book Chapter
Title:	*Forecasting the Start and End of Pollen Season in Madrid*
Volume:	Advances in Time Series Analysis and Forecasting
Collection:	Contributions to Statistics
Authors:	Ricardo Navares & José Luis Aznarte
Published:	August 2017
Editorial:	Springer International Publishing AG
ICEE:	670.000
SPI Rank:	4
DOI	10.1007/978-3-319-55789-2_27

Forecasting the Start and End of Pollen Season in Madrid

Ricardo Navares and José Luis Aznarte

Abstract In this paper we approach the problem of predicting the start and the end dates for the pollen season of grasses (family Poaceae) and plantains (family Plantago) in the city of Madrid. A classification-based approach is introduced to forecast the main pollination season, and the proposed method is applied to a range of parameters such as the threshold level, which defines the pollen season, and several forecasting horizons. Different computational intelligence approaches are tested including Random Forests, Logistic Regression and Support Vector Machines. The model allows to predict risk exposures for patients and thus anticipate the activation of preventive measures for clinical institutions.

Keywords Forecasting · Time series · Pollen · Poaceae · Plantago · Support vector machines · Logistic regression · Random forests

1 Introduction

Airborne pollen levels have been associated to allergic rhinoconjunctivitis, asthma and the oral allergy-symptom in about 15 million people in Europe. Allergies have been continuously increasing in developed countries, not only in the number of affected patients but also in the severity of allergic reactions [20]. The establishment and the prediction of a pollen calendar is essential to reduce the exposure of allergic patients to pollen during the days of higher pollen concentration. It is also important to enable the development of other preventive measures.

There is no consensus on how to define the pollination season [9] which is the period where airborne concentrations of pollen are measured. Some authors define it based on the cumulative daily pollen counts [1, 7, 13] and other authors define it based on predefined threshold levels over which the season is considered to be started

R. Navares · J.L. Aznarte (✉)
Department of Artificial Intelligence, Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain
e-mail: jlaznarte@dia.uned.es

I. Rojas et al. (eds.), *Advances in Time Series Analysis and Forecasting*, Contributions to Statistics, DOI 10.1007/978-3-319-55789-2_27

and ended [18]. This study visits both approaches in order to define the season which is going to be forecast.

Climate directly or indirectly defines the vegetation and acts on two levels: (1) during the stages prior to flowering [4, 14], and (2) during the pollen season [12, 17]. In this study, we characterize different features of the pollen season in order to determine the effect of meteorological parameters on the incidence of Poaceae and Plantago pollen in Madrid, Spain. Once the features are defined, several computational intelligence techniques are applied and compared according to their performance on this problem. We cast the season predicting problem into a binary classification one, in order to obtain the most accurate estimates for the start and end of the pollination season with special attention to the threshold at which allergy reactions might appear.

The rest of this paper is as follows. Section 2 deals with data preprocessing, including its cleansing, formatting and set up. Then in Sect. 3, we summarize the different approaches of what is considered a peak season and present its definition in order to identify the data points which belong to it. The computational intelligence models considered are described in Sect. 4, which walks through the system design and the definition of the features which will be tested according to its forecasting relevance. Section 5 contains the results and analysis of the different experiments. Finally, Sect. 6 draws the conclusions and the future lines of work in this line.

2 Data Description

The study uses observations of Poaceae and Plantago pollen from the Faculty of Pharmacy of Complutense University of Madrid, Spain (40°26′52.1″ N, 3°43′41.1″ W) from 1994 to 2013, provided by Red Palinológica de la Comunidad de Madrid. Meteorological data is provided by weather stations located in Barajas, Cuatro Vientos, Getafe and Colmenar and consists of hours of sunlight per day, the speed of wind in km/h, rainfall in mm/h and daily maximum, minimum and average temperature in degrees Celsius.

A first look at the pollen observations reveals the presence of missing data points. Bearing in mind the season start problem and the minimization of the loss of information, it is clear that those missing data points which appear around the months of February, March and April have a more severe impact, as they are the months in which usually the pollen season start is usually recorded. Thus, a long sequence of consecutive missing data points might multiply the forecasting errors as it may artificially delay the predicted season start. If we find a sequence of missing data around the season start date, the use of the traditional 'last observation carried forward' (LOCF) method may lead to an incorrect prediction of the season start. These reasons support the initial hypothesis that interpolation within each year is not enough.

Consequently, we propose to redistribute the data into a matrix of dimensions $N \times 365$, where N denotes the year. As there are leap years in the data sample, a first check has been done to verify whether there is any data point on the 29th of February which is missing. As it is not the case, each data point which lays on that

date is not taken into account to interpolate. Later on, the data point will be plugged into the correspondent year. With this format, missing data points can be regressed using data within the year and between years.

From this matrix, two new matrices are generated, one with the missing data estimated using regression by rows (within the year) and another with the data regressed by columns (by years). Given the different years' conditions due to factors which directly influence pollen concentrations, it is important to avoid over-influence of data from previous or subsequent years when estimating a data point. High concentration of grains the same day in other years as the one to be estimated does not imply high concentration on day that day. In order to give more importance to most recent data, it is within the year, the final estimation is weighted.

Meteorological data, on the other hand, presented very few missing data points, so they were directly linearly interpolated.

3 Definition of Season Start and End

There is no consensus on the definition of the main pollination season, but the different proposals lie in two main categories: those based on cumulative daily pollen counts, which define the period with respect to a percentage of yearly total sum of daily concentrations, and those which rely upon a consistent pollen threshold breach [9].

Table 1 shows how different the effective computed dates are for our data depending on the season definition. It is noticeable that the definitions which use thresholds, such as [6, 18], instead of cumulative concentrations, such as [1, 7, 13], tend to limit the season to the period where the peak concentrations appear. They are also sensitive to out of period isolated peak concentrations.

Fig. 1 shows the pollen concentrations, both for Plantago and for Poaceae, for the same years considered in Table 1, as well as the limits of the season defined according to [13, 18]. In the case of Poaceae (bottom row), it is interesting to see how the latter approach (based on a threshold of 30 grains/m^3) restricts the pollination season to a few days around the main peak in 2006. The same applies to Plantago pollen (top row). It is noticeable how concentrations differ for each species, being Plantago less prolific compared to Poaceae, which motivates a threshold level adjustment based on the pollen class as proposed by [21].

In general, the proposal of [18] seems much more restrictive than [13]. This can be mitigated by relaxing the threshold condition by reducing the threshold to 15 grains/m^3, which produces a more realistic result in the cases studied (shown in the graph as a shaded rectangle).

Cumulative approaches imply forecasting, before the season start, the expected total yearly accumulation, which is an entirely different problem. Henceforth, we will limit this work to threshold-based definitions. In order to establish a systematic approach which allows for a more informed decision about the threshold, we will study a set of thresholds allowing the experts to choose the most influential defini-

Table 1 Considered definitions for the start and end of the Poaceae pollination season, with examples for some years

Approach	Definition	Year	Start	End
Nilsson et al. [13]	The day in which the sum of daily pollen concentration reaches a value over 5% (start) and 95% (end) of the total yearly sum	1999	26 Feb	06 Aug
		2001	17 Mar	09 Jul
		2006	16 Feb	05 Jul
		2010	12 Apr	27 Jul
Galán et al. [7]	The day in which the sum of daily pollen concentration reaches a value over 1% (start) and 99% (end) of the total yearly sum	1999	24 Jan	19 Oct
		2001	13 Feb	24 Sep
		2006	02 Feb	02 Sep
		2010	18 Feb	16 Sep
Andersen et al. [1]	The day in which the sum of daily pollen concentration reaches a value over 2.5% (start) and 97.5% (end) of the total yearly sum	1999	08 Feb	13 Sep
		2001	22 Feb	08 Aug
		2006	09 Feb	29 Jul
		2010	20 Mar	17 Aug
Sánchez-Mesa et al. [18]	The first day in which the daily pollen concentration reaches values over (start) and below (end) 30 grains/m^3	1999	16 May	20 Jun
		2001	17 May	03 Jun
		2006	01 May	02 Jun
		2010	23 May	24 Jun
Feher et al. [6]	The first day in which the daily pollen concentration reaches values over (start) and below (end) 3 grains/m^3 for 4 consecutive days	1999	09 Feb	11 Jul
		2001	02 Feb	05 Sep
		2006	03 Feb	17 Jul
		2010	29 Mar	08 Sep

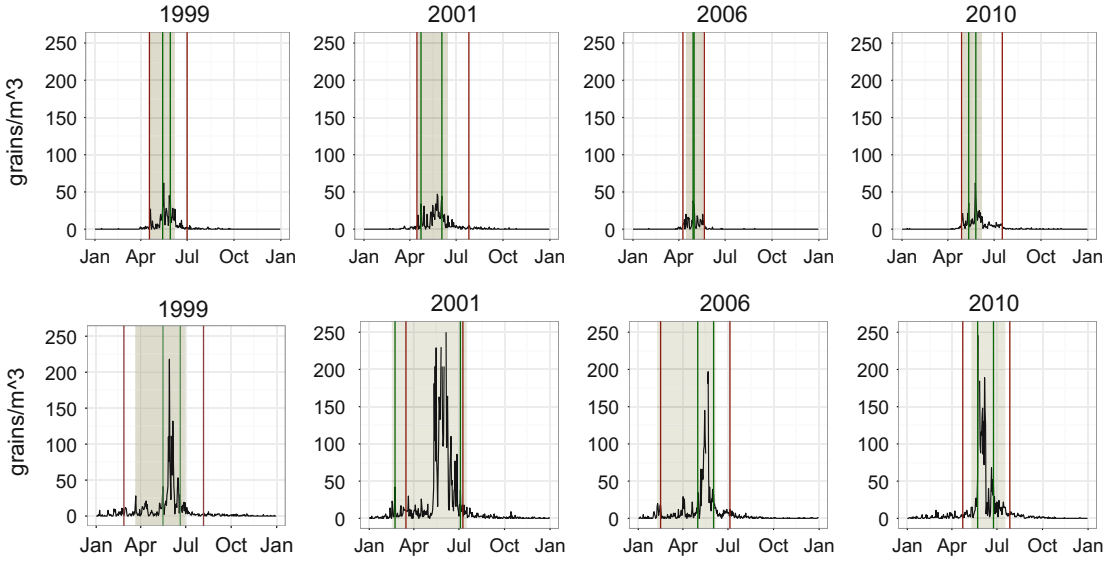


Fig. 1 Plantago (*top row*) and Poaceae (*bottom row*) pollen concentrations for years 1999, 2001, 2006 and 2010 and definition of the season according to [13] (*vertical red line*) and [18] (*vertical green line*). The shaded rectangle represents the latter approach relaxing the threshold to 15 grains/m^3

tion according to the relevance on their field. In what follows, let u be a fixed daily pollen concentration threshold, then the pollen season start as the first (last) day that surpasses u.

4 Methods

The final aim of this work is to help allergy patients in knowing in advance between which dates the pollen concentrations will be at risk levels. Given the above definitions of pollination season start and end, we aim at developing a model which forecasts these dates.

As seen in Sect. 3, there is no consensus as to which are the pollen concentrations considered as risk levels. Hence, several thresholds, ranging from 5 to 50 grains/m^3 for Poaceae and 5–15 grains/m^3 for Plantago [21], will be used in this work in order to provide a variety of options and to compare them.

Another important element that needs to be fixed is the forecasting horizon, which corresponds to the number of days in advance pollen concentrations will be forecast. There is always a trade off between precision and anticipation, and in the literature we can find predictions of the pollen season which range from 1 to 10 days in advance. In order to test its predictive capacities, the model will produce forecasts for several forecasting horizons ranging from 1 to 15 days.

Finally, for each combination of thresholds and horizons, different derived meteorological and pollen features are computed to set up the instances on which different machine learning algorithms will be trained.

Our approach is based on the idea that one can cast the forecasting problem into a binary classification problem where the featured instances represent influential factors for the predictions. Hence, daily pollen concentrations are mapped to $\{0, 1\}$ depending on whether they are above the threshold (1) or not (0). Given the definition of season start, the first data point classified as 1 will indicate the start of the season.

4.1 Feature Generation

In order to build such a classification system, the instances of each class should contain the most relevant data for that class. This relevant data can be meteorological conditions or pollen levels themselves, either for the day in which the prediction is to be made or for previous days, weeks or months, as it is generally assumed that those are the values that play a role in the development of the pollination process. At the same time, we need to avoid data which might not be related with the problem, for example we can assume that the average maximum temperatures of 5 years ago might not carry much information for the pollination period of the actual year.

According to previous works [16, 19], it is important to include the influence of most recent data, and hence cumulative pollen observations until the forecast day is defined as a synthetic variable. A 10 and 30 days cumulative sums of pollen daily concentrations prior to the forecast date are defined as features along with the prior 7 daily concentrations and the total sum of the pollen concentrations within the year.

Some authors assume that there is a linear relationship between the energy a plant receives and the growth state of buds [4]. This energy is represented in several ways, for example it is usual to consider that the sum of temperatures up to some point can be of help to forecast the state of the flowers [1, 4, 17].

On the other hand, some authors use the concepts of *chilling temperatures* and *forcing temperatures*, which are the weighted sum of the temperatures below and above certain thresholds given a period. Instead of using a predefined period, our approach is intended to capture all possible relevant periods which might influence the state of buds. In [14] the start of the chilling period is defined as the 1st of October and the start of the forcing period as of the 1st of February, which consequently it is also called the end of the chilling period. The forcing period ends when the pollination season start, and according to this approach is when the pollen concentration surpasses certain threshold. Instead, our approach calculates for each month the forcing and chilling temperatures and generate for each instance features that represent previous forcing and chilling temperatures, computed for previous months, quarters and so on.

Given the non-fixed definition of the chilling and forcing period, we decided not to apply any weight to the temperatures so the calculation of the forcing temperature sum is as follows:

$$F_{\text{sum}}(d) = \sum_{i=d-n}^{d} R_{\text{forc}}(i), \tag{1}$$

where

$$R_{\text{forc}}(i) = \begin{cases} 0 & \text{if } T(i) < T_{\text{forc}} \\ T(i) - T_{\text{forc}} & \text{if } T(i) \geq T_{\text{forc}} \end{cases} \tag{2}$$

being d the forecast date, n the number of days which define the calculation period for the sum of forcing temperatures, $T(i)$ the temperature for day i, and T_{forc} the base temperature for forcing (all temperatures are in degrees Celsius). The same applies for the chilling. In order to determine the base temperatures for T_{forc} and T_{chill} the levels proposed by [14] are used as a reference. The authors proposed a base temperature for the forcing period of 1 °C and 16 °C for the pollen thresholds of 10 grains/m^3 and 50 grains/m^3, respectively, and −6 °C and 8 °C for the chilling period. As this study uses different threshold, it is fair to approximate the values using simple geometrical relations setting the new values accordingly. Given a definition of the threshold of 30 grains/m^3 the corresponding base temperatures are 8 °C for the forcing and 6 °C for the chilling. Cumulative temperature parameterization is widely used to capture the energy induced to the plant during the early stage of bud states.

Table 2 Number of features generated by variable

	i	10	y	m	q	Q	std	MA_5	MA_{10}
Pollen	7	1	1	1	1	1	–	–	–
T[a]	21	–	–	–	–	–	–	3	3
T_{forc}	–	–	–	1	1	1	–	–	–
T_{chill}	–	–	–	1	1	1	–	–	–
Wind	7	1	1	1	1	1	1	1	1
Rain	7	1	1	1	1	1	1	1	1
Sun	7	1	1	1	1	1	1	1	1

i previous $i \in [1, 7]$ day observation; *10* previous 10 day cummulative sum; *y* year to date cummulative sum; *m* previous month cummulative sum; *q* previous 90 day cummulative sum; *Q* previous 180 day cummulative sum; *std* previous 15 days standard deviation; MA_5: previous 5 days moving average; MA_{10}: previous 10 days moving average
[a]accounts for 3 variables ($T_{min,max,avg}$)

Finally, it is known that pollen release is more prolific during dry weather rather than in rainy periods, even during cooler weather. Thus, it makes sense to take the cumulative approach introduced for temperatures in order to capture the prolonged rain periods. We need to capture heavy rains as well so the approach is based on the standard deviation of the last 15 days rainfall before the forecasting day. On the other hand, there are long term issues with heavy rains which need to be captured. For example, heavy spring rains are known to cause grass species to become more abundant as they grow more rapidly. Heavy rains during fall and winter cause pollen level increases in spring. Having said that, it is logical to include the accumulation of previous meteorological seasons. The same applies to the daily sun hours which is used as a proxy for dryness. For all climate data, similar as for the pollen counts, the prior 7 daily data observations are included. All variables are summarized in Table 2.

4.2 Setting up the Data

The aforementioned feature generation process leaves us with a total of 90 features. Depending on the desired threshold and forecast horizon, the data is set up according to the parameters in order to transform it into a classification problem. The first step consists on discretising the class and then assigning the class to the correspondent instance based on the forecast horizon defined.

$$\left[\begin{array}{cccc|c} x_{1,1} & x_{1,2} & \dots & x_{1,90} & p_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,90} & p_n \end{array}\right] \rightarrow \left[\begin{array}{cccc|c} x_{1,1} & x_{1,2} & \dots & x_{1,90} & c_{1+t} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{n-t,1} & x_{n-t,2} & \dots & x_{n-t,90} & c_n \end{array}\right] \tag{3}$$

$$c_i = \begin{cases} 0 & if \;\; p_i < u \\ 1 & if \;\; p_i \geq u \end{cases}, \tag{4}$$

where p_i is the daily pollen observation at time i, t the forecast horizon in number of days and u is the threshold as defined in Sect. 3. The observations are split into two subsets which will be used to train the correspondent algorithm and test its prediction accuracy. The test set consist on the observations which belong to the years 2011, 2012 and 2013 and the rest of the available years belong to the training set. As well, to avoid the over-fitting phenomenon, common to many machine learning models, a cross-validation procedure is performed on each year of the training set using as an error measure the absolute value of number of days between the estimated and the observed season start and end.

4.3 Feature Selection

Some models are highly sensitive to collinearity in the variables. In order to provide equal competitiveness to the algorithms, we need to reduce the number of features to those which are relevant for the class.

Hence, a filter algorithm based on [8] and on the definition of feature relevance by [10] is applied to rank subsets of features according to a correlation based evaluation function. This algorithm will select subsets that contain features highly correlated with the class and uncorrelated with each other. A feature is accepted when it predicts the class in areas of the instance space not already predicted by other features. The features are treated uniformly by discretisation in a pre-processing step, and then a correlation based heuristic is repeatedly applied to test the merit of a subset, defined as

$$M_s = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \tag{5}$$

where M_s the merit of a subset S containing k features and $\overline{r_{cf}}$ is the mean feature-class correlation and $\overline{r_{ff}}$ the average feature-feature correlation.

4.4 Computational Intelligence Models

Different classification approaches are trained using the training set in order to forecast the start and end of the season for test set. Concretely we compare Random Forests (RF) [3], Logistic Regression (LR) [11] and Support Vector Machines (SVM) [15].

Proposed in 2001 by Leo Breiman [3], a Random Forest is the name for an ensemble approach which leverages the performance of many decision trees to produce predictive models. It is a supervised learning procedure which combines several randomized decision trees and aggregates their predictions by averaging. The procedure operates over sample fractions of the data, grows a randomized tree predictor on each one and aggregate these predictors together.

With respect to logistic regression, is a widely used regression model used in Statistics where the dependent variable is categorical. The model predicts the probability that a given example belongs to one class via the sigmoid function. A ridge estimator [11] was introduced to add penalty on weights learned to avoid over-fitting.

RF and LR make different assumptions about the data and has different rates of convergence. On the one hand, RF assumes that the decision boundaries are parallel to the axes based on whether a feature is $\geq, \leq, <$ or $>$ to certain value so the feature space is chopped into hyper-rectangles. On the other hand, LR finds a linear decision boundary in any direction by making assumptions on $P(C|X_n)$ applied to weighted features so non-parallel to the axes decision boundaries are picked out. This trade off motivates to take into account SVM as an alternative.

The current SVM standard algorithm, proposed by Cortes and Vapnik [5] in 1995, is a learning method used for binary classification which finds a hyper-plane which separates the d-dimensional data perfectly into its two classes. However, since sample data is often not linearly separable, SVM's introduces the notion of a *kernel induced feature space* which casts the data into a higher dimensional space where the data is separable.

In sum, the experiments are tailored to compare the models and compute their forecasts for each threshold and time horizon previously defined. Both parameters, threshold and horizon, define a set up of the data presented to the models according to Eqs. (3) and (4). Then a three step process applies consisting on feature selection and evaluation of the learning algorithm over the training set and prediction on the test set.

5 Results

One of the objectives of the experiments was to evaluate, in terms of their predictive ability in the framework of forecasting the pollen season in Madrid, different general purpose machine learning or statistic methods based on different paradigms. Hence, in order to select the best suited model of those described in Sect. 4.4, we tried them against the data described in Sect. 2.

For a set of thresholds and for a set of forecast horizons $h = \{1, 2, 5, 7, 10, 15\}$, we trained the three methods using the training set and checked their performance against the test data set.

Fig. 2a shows the start and end of the season for Poaceae pollen along with the predicted values for each combination of algorithm, threshold and forecast horizon.

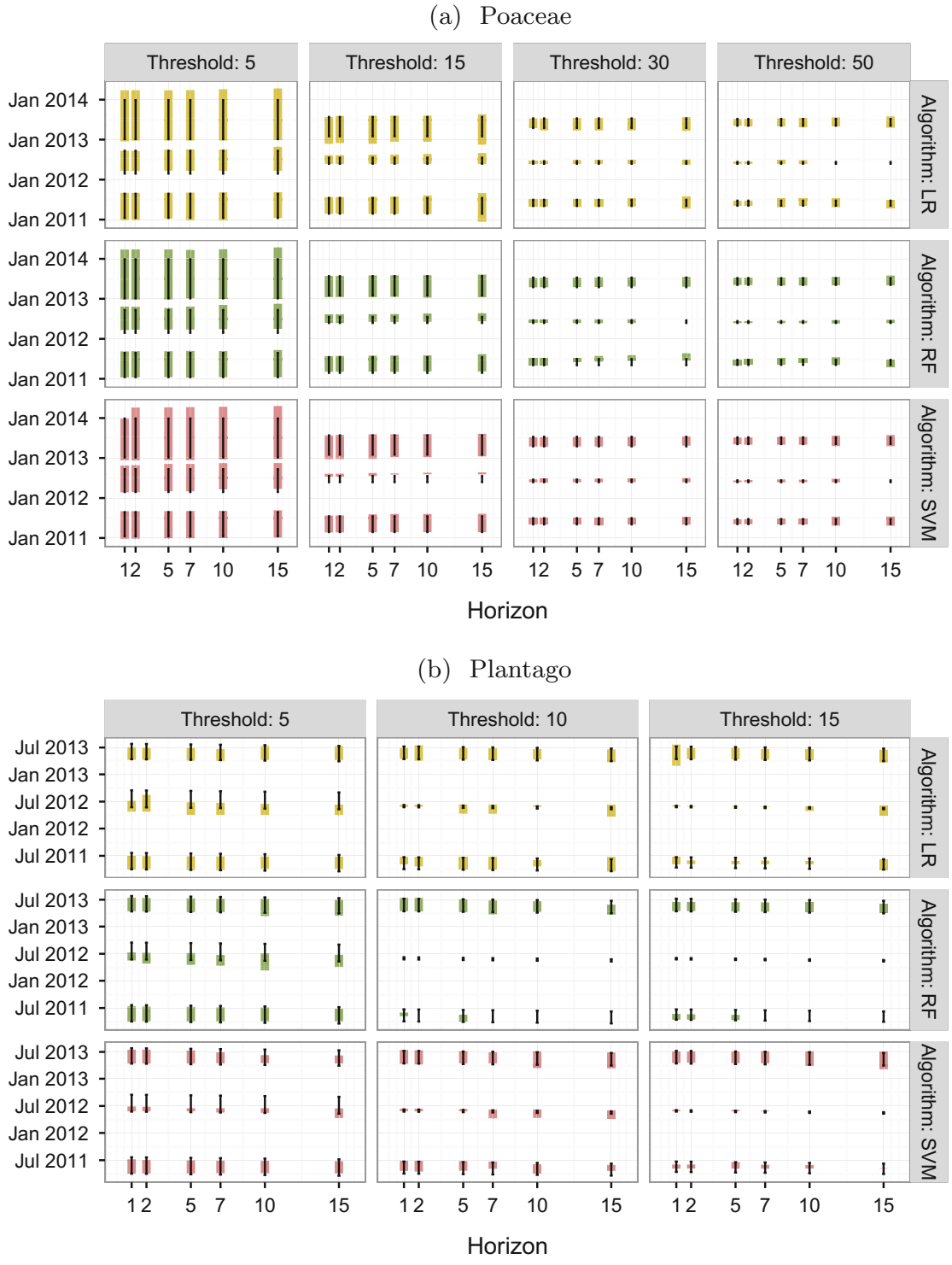


Fig. 2 Predicted (*coloured rectangles*) and observed (*black vertical lines*) season start and end dates for 2011, 2012 and 2013, by algorithm and threshold

Table 3 Test data set average errors for $u = 30$ and $u = 15$ for Poaceae and Plantago respectively, in number of days, of the predictions for the start of the season

Poaceae	Horizon					
Algorithm	1	2	5	7	10	15
LR	1.00	8.00	8.67	9.00	9.00	19.33
RF	0.33	1.00	10.67	12.33	15.67	23.50
SVM	1.33	1.67	1.33	1.67	1.33	3.33
Plantago	Horizon					
Algorithm	1	2	5	7	10	15
LR	1.00	1.67	0.67	0.78	5.18	12.47
RF	12.22	12.72	15.50	15.83	12.50	16.06
SVM	10.39	9.61	10.17	11.89	12.28	18.53

It can be clearly seen the highly dependence between the season duration and the definition of the threshold.

The results with the test data set might derive from the fact that the models do not have enough data to properly generalize, as we only have 20 years, which means only 20 season starts and ends. However, it is clear that high threshold levels lead to more satisfactory results, enabling the classifier to identify the patterns which influences the season start and end even for long forecasting periods.

On the other hand, Fig. 2b shows comparatively very short pollination seasons for Plantago pollen. This is due to the fact that this species is not as common in metropolitan areas as in rural regions [21]. For this reason, the threshold levels were relaxed according to the findings in [21]. However, the proposed models are in this case tested with a small set of data that are effectively classified as main pollination season, and this plays against the computational intelligence models as they need a high number of training observations. For instance, RF strives to identify the main pollination season in 2012 for thresholds over 5 grains/m^3.

From a clinical point of view, predicting the moment in which most of the patients will start having symptoms is of a greater interest than predicting the moment when they will experience relief. Hence, Table 3 shows the error obtained by each model for all the horizons considered at predicting the start of the season. Only the threshold $u = 30$ is considered for Poaceae, following [2] (all patients experience moderate or severe symptoms) and $u = 15$ for Plantago [21]. It is clear that SVM outperforms the other algorithms for horizons over 5 days, while RF is the best for 1 or 2 days ahead forecasts of Poaceae pollen season. Conversely, LR is shown as the best performer for Plantago given the limited amount of training samples in this case. This situation leads to an increase in robustness for LR compared to the other proposals, which need a higher number of observations over the threshold in order to obtain the inner information from the data.

6 Conclusions and Future Works

This study introduces a new approach to foresee the start and end of the pollination season, which might help allergic patients as well as public health institutions. It is shown that tackling the problem from a purely data-driven point of view produces good results and gets accurate forecasts of the pollination season even in years with particularly odd characteristics as it is 2012, which shows a specially short main pollination period with a sudden start.

We have seen SVM as the most general model for prediction on this problem having accurate results for horizons within a week. The definition of the threshold, which dictates the start and end of the pollination season, takes an important role on the performance of the models. This study shows that levels above 20 grains/m^3 allow an accurate prediction in the case of Poaceae. It is to note that previous works set the threshold at 30 grains/m^3 or above [9].

Regarding Plantago pollen, the season definition produced a limited number of observations over the threshold above 15 grains/m^3, and LR was the most robust approach.

The proposed approach provide forecasts based on the data and making no assumptions on the phenology of the plant. Thus, it can be applied to any kind of pollen regardless its origin. The results are presented in a way to be easily interpreted either by experts from other fields or patients.

The results are promising but some ideas are worth deeper exploration. For example, the generation and selection of features could be improved by using bio-inspired algorithms. As well, the introduction of numerical weather predictions should enhance the prediction results. As well, predictions which account for uncertainty for the start date, like probabilistic predictions, could also be of interest.

Acknowledgements This work has been partially funded by Ministerio de Economía y Competitividad, Gobierno de España, through a *Ramón y Cajal* grant awarded to Dr Aznarte (reference: RYC-2012-11984).

References

1. Andersen, T.B.: A model to predict the beginning of the pollen season. Grana **30**, 269–275 (1991)
2. Antépara, I., Fernández, J.C., Gamboa, P., Jauregui, I., Miguel, F.: Pollen allergy in the Bilbao area (European Atlantic seaboard climate): pollination forecasting methods. Clin. Exp. Allergy **25**(2), 133–140 (1995)
3. Breiman, L.: Random forest. Mach. Learn. **45**, 5–32 (2001)
4. Cannell, M.G.R., Smith, R.I.: Thermal time, chill days and prediction of budburst in Picea sitchensis. J. Appl. Ecol. **20**, 269–275 (1983)
5. Cortes, C., Vapnik, V.N.: Support-vector networks. Mach. Learn. **20**, 273–276 (1995)
6. Feher, Z., Jarai-Komlodi, M.: An examination of the main characteristics of the pollen seasons in Budapest, Hungary (1991–1996). Grana **36**, 169–174 (1997)

7. Galan, C., Emberlin, J., Dominguez, E., Bryant, R.H., Villamandos, F.: A comparative analysis of daily variations in the gramineae pollen counts at Cordoba, Spain and London, UK. Grana **34**, 189–198 (1995)
8. Hall, M.A.: Correlation-based feature selection for machine learning. Ph.D. thesis. University of Waikato (1999)
9. Jato, V., Rodriguez-Rajo, F.J., Alcazar, P., De Nuntiis, P., Galan, C., Mandrioli, P.: May the definition of pollen season influence aerobiological results? Aerobiologia **22**, 13–25 (2006)
10. Kohavi, R., John, G.H.: Wrappers for feature subset selection. Artif. Intell. **97**, 273–324 (1997)
11. le Cessie, S., van Howelingen, J.C.: Ridge estimators in logistic regression. Appl. Stat. **41**, 191–201 (1992)
12. Myszkowska, D.: Predicting tree pollen season start dates using thermal conditions. Aerobiologia **30**, 307–321 (2014)
13. Nilsson, S., Persson, S.: Tree pollen spectra in the stockholm region (Sweden), 1973–1980. Grana **20**, 179–182 (1981)
14. Pauling, A., Gehrig, R., Clot, B.: Toward optimized temperature sum parametrizations for forecasting the start of the pollen season. Aerobiologia **30**, 45–57 (2014)
15. Rakotomamonjy, A.: Variable selection using SVM-based criteria. J. Mach. Learn. **3**, 1357–1370 (2003)
16. Ribeiro, H., Cunha, M., Abreu, I.: Definition of main pollen season using logistic model. Ann. Agric. Environ. Med. **14**, 259–264 (2007)
17. Rodriguez-Rajo, F.J., Frenguelli, G., Jato, M.V.: Effect of air temperature on forecasting the start of the Betula pollen season at two contrasting sites in the south of Europe (1995–2001). Int. J. Biometeorol. **47**, 117–125 (1983)
18. Sanchez-Mesa, J.A., Smith, M., Emberlin, J., Allitt, U., Caulton, E., Galan, C.: Characteristics of grass pollen seasons in areas of Southern Spain and the United Kingdom. Aerobiologia **19**, 243–250 (2003)
19. Smith, M., Emberlin, J.: A 30-day-ahead forecast model for grass pollen in North London, UK. Int. J. Biometeorol. **50**, 233–242 (2006)
20. Sofiev, M., Bergmann, K.C.: Allergenic Pollen: A Review of the Production, Release, Distribution and Health Impacts. Springer Science and Business Media (2012)
21. Tobías, A., Sáez, M., Galán, I., Benegas, R.: Point-wise estimation of non-linear effects of airborne pollen levels on asthma emergency room admissions. Allegy **64**, 961–962 (2009)

Chapter 4

Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features

Type:	Published Article
Title:	*Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features*
Journal:	International Journal of Biometeorology
Authors:	Ricardo Navares & José Luis Aznarte
Published:	September 2016
Impact Factor:	2.377
Quartile:	Q2
DOI	10.1007/s00484-016-1242-8

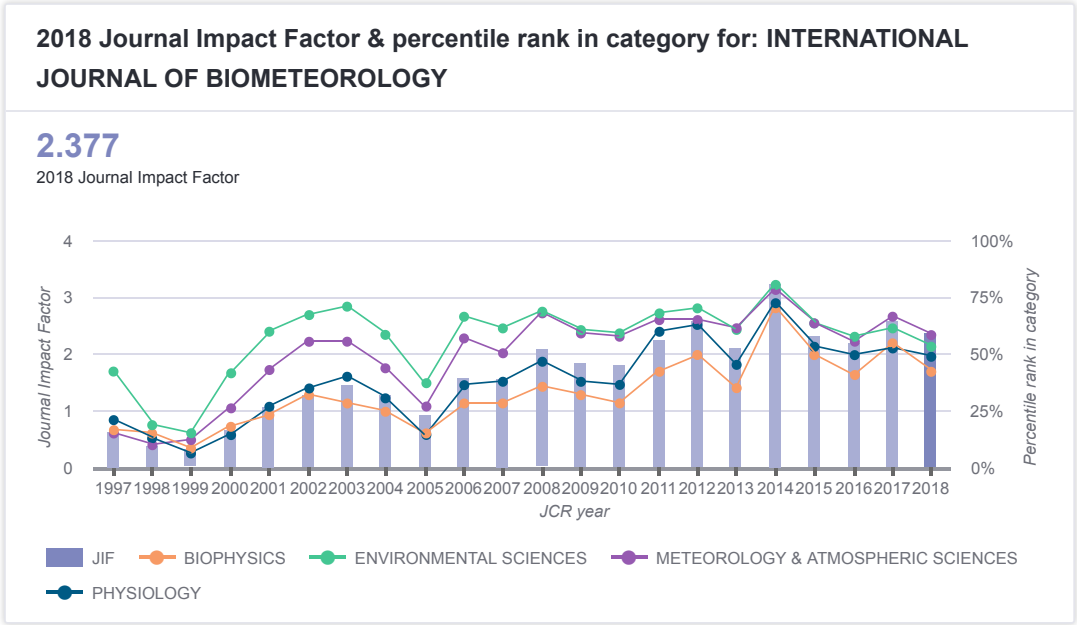


FIGURE 4.1: Impact factor International Journal of Biometeorology

Int J Biometeorol
DOI 10.1007/s00484-016-1242-8

ORIGINAL PAPER

Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features

Ricardo Navares[1] · **José Luis Aznarte**[2]

Received: 24 May 2016 / Revised: 24 August 2016 / Accepted: 25 August 2016

Abstract In this paper, we approach the problem of predicting the concentrations of Poaceae pollen which define the main pollination season in the city of Madrid. A classification-based approach, based on a computational intelligence model (random forests), is applied to forecast the dates in which risk concentration levels are to be observed. Unlike previous works, the proposal extends the range of forecasting horizons up to 6 months ahead. Furthermore, the proposed model allows to determine the most influential factors for each horizon, making no assumptions about the significance of the weather features. The performace of the proposed model proves it as a successful tool for allergy patients in preventing and minimizing the exposure to risky pollen concentrations and for researchers to gain a deeper insight on the factors driving the pollination season.

Keywords Poaceae · Pollen · Random forest · Forecasting · Time series

Introduction

Continuously increasing allergy symptoms in developed countries, and the clinic and socioeconomic relevance of this problem, have boosted recent research around some of the issues dealt with by aerobiology, especially concerning predictive models. The fact is that not only has the number of cases increased, but also the severity and the prevalence of the reactions (de Weger et al. 2013). In order to enable preventive measures and reduce the exposure for patients, this study focuses on the prediction of pollen concentration levels which imply high risk for allergic population.

The main pollination season is defined as the period where high pollen concentrations are measured. In the literature, several definitions of what it is considered a pollen season have been established (Jato et al. 2006). It is possible to classify them into two main approaches, those based on the cumulative daily atmospheric concentrations (Andersen 1991; Galán et al. 1995; Nilsson and Persson 1981) and those based on a predefined threshold level over which the season is defined to start and end Sánchez-Mesa et al. (2003).

Weather plays a major role in the severity and length of the pollination season, as it is the cause for increases and decreases of the pollen concentration levels through its effect on the plants. For example, a mild winter usually implies an early pollen season, as it influences the plant development stages prior to the flowering (Cannell and Smith 1983; Pauling et al. 2014). On the other hand, a dry and windy weather spreads the airborne quickly, leading to higher distributions (Myszkowska 2014; Rodríguez-Rajo et al. 1983). In this study, we investigate the meteorological effects which determine the season of Poaceae pollen in Madrid, Spain. In our approach, the forecasting problem is cast to a binary classification problem with attention to the thresholds considered risk levels for the appearance of allergy reactions.

Several research teams have established models to predict the pollination season based on assumptions about

✉ José Luis Aznarte
jlaznarte@dia.uned.es

[1] Superior Technical School of Computer Engineering, UNED, Juan del Rosal, 16, 28040, Madrid, Spain

[2] Department of Artificial Intelligence, UNED, Juan del Rosal, 16, 28040, Madrid, Spain

Published online: 16 September 2016

the influence of meteorological conditions (Andersen 1991; Myszkowska 2014; Pauling et al. 2014; Rodríguez-Rajo et al. 1983) or previous pollen concentrations (Castellano-Méndez et al. 2005). The aim of this research is to provide an assumption-free predictive model using a computational intelligence technique known as random forests (RF) (Breiman 2001). The study lets the RF select the most influential features from a purely data point of view according to their predictive significance and provides this information allowing for interpretability of the results. Earlier applications of computational intelligence methods can be found, for example, in Aznarte et al. (2007).

Very few of the previous predictive studies for pollen were able to provide this type of information about the relevance of the variables. And most of them dealt with forecasts horizons ranging from 1 to 10 days (Andersen 1991; Castellano-Méndez et al. 2005; Myszkowska 2014). The procedure presented in this work provides long-term predictions, up to 180 days, expanding their usefulness to prevent allergy symptoms.

The aim of this study is to provide a framework to forecast and identify the main factors which influence high pollen concentrations, and do this from a purely data-driven point of view. These long-term predictions could help research centers and clinical institutions to plan in advance the implications of high airborne concentrations and their duration, as well as allergy patients to be able to limit their exposure to risky pollen levels. Furthermore, this study is also aimed to provide support to phenological studies by identifying the relevant pollination factors from the information obtained from the data.

Materials and methods

Data description

Weather data Meteorological data are provided by Ayuntamiento de Madrid for the weather stations located in Casa de Campo, Plaza de España and Cuatro Caminos. Weather observations consist of average daily temperature in Celsius degrees, hours of sunlight per day, wind speed measured in m/s, daily rainfall in mm/h, pressure in mbar, degree of humidity in percentage, and ultraviolet radiation in mW/m^2. Very few missing observations appear in the meteorological series, and these were linearly interpolated.

Pollen data Pollen observations correspond to daily Poaceae concentrations registered at the Faculty of Pharmacy of Complutense University of Madrid, Spain (located at 40°26′52.1″ N, 3°43′41.1″ W) from 2000 to 2013. These data have been kindly provided by Red Palinológica de la Comunidad de Madrid and were obtained following the standard methodology of the Spanish Aerobiological Network. They are measured in grains per cubic meter of air.

Missing values in the pollen time series may lead to an artificial delay of the season start, especially when those appear in the critical months of February, March, and April, as it is when the daily concentrations are expected to increase. Table 1 shows, for instance, high presence of consecutive missing values on March 2001 and August 2009 compared to other months. These are a priori critical months as the season might start and end on those periods. Using the standard 'last observation carried forward' (LOCF) method

Table 1 Maximum number of consecutive days of missing data per month and year

	Month											
Year	1	2	3	4	5	6	7	8	9	10	11	12
2000	–	1	2	–	–	–	–	–	–	–	4	-
2001	–	–	8	–	2	–	–	–	–	–	–	-
2002	–	–	–	–	–	–	–	3	–	–	–	-
2003	–	–	–	–	–	–	–	–	–	–	–	-
2004	–	–	–	–	–	–	–	–	–	–	–	-
2005	2	1	3	1	–	–	–	–	–	3	–	-
2006	–	–	–	–	–	–	–	–	–	–	–	-
2007	–	–	–	–	–	–	–	–	–	–	–	-
2008	–	–	–	–	–	–	–	–	–	–	–	-
2009	–	–	–	–	1	–	–	18	2	1	–	11
2010	7	–	–	–	–	–	1	–	–	–	–	-
2011	–	–	–	–	–	–	–	–	–	–	–	-
2012	4	–	–	–	–	2	–	–	–	–	–	-

to estimate the missing observations does not fully solve this problem. Thus, we applied a redistribution of the data into a matrix of dimensions $N \times 365$, being N the number of available years. (No data were missing for the 29th of February in any year, so that day was removed from all the years to make the matrix dimensions match).

Out of this set up, two new matrices are generated to regress the missing data points by rows (within each year) and by columns (by years). Data suggests that concentration levels for the same day in different years do not imply similar levels in another,and hence, the resulting matrices are weighted to give more relevance to most recent data (within each year), as in:

$$p_t = \beta \cdot \mathrm{r_{row}} + (1 - \beta) \cdot \mathrm{r_{col}}, \quad (1)$$

where $\mathrm{r_{row}}$ is a linear regression within the year, $\mathrm{r_{col}}$ is a linear regression across years, and $\beta = 0.6833$ is estimated from the data.

Season definition

In literature, the main pollination season is defined according to two different approaches (Jato et al. 2006). The first one is based on daily cumulative airborne concentrations and the second considers the season started when a pollen concentration threshold is consistently surpassed.

There is no general consensus about the definition of the pollination season, and hence, season dates might differ according to their definition. Table 2 shows the differences between approaches on selected years and authors with their corresponding definition of the main pollination season. Threshold-based approaches such as Feher and Jarai-Komlodi (1997) and Sánchez-Mesa et al. (2003) tend to limit the season where peak concentrations appear, and this implies a high sensitivity to isolated peak concentrations. In contrast, cumulative approaches widen the pollination period being sensitive to early moderate concentrations, as is the case for 2002 in the table. Figure 1 shows how restrictive the proposal of Sánchez-Mesa et al. (2003) is compared to Nilsson and Persson (1981) and how the season period varies by reducing the threshold to 15 grains/m^3.

However, in order to forecast the season start as defined by the cumulative approaches, it would first be necessary to forecast the expected total yearly accumulation, which determines the percentages to define the pollination season. Of course, this is unfeasible as it implies forecasting one quantity (the yearly sum) in order to forecast the other (a quantile). Hence, this study will be restricted to threshold -based season definitions. In what follows, if u is a fixed

Table 2 Sample start and end of the pollination season according to different definitions

Approach	Definition	Year	Start	End
Nilsson and Persson 1981	The day in which the sum of daily	2002	09 Feb	03 Nov
	pollen concentration reaches a	2004	26 Feb	11 Jul
	value over 5 % (start) and 95 %	2009	04 Apr	09 Sep
	(end) of the total yearly sum.	2012	17 May	31 Aug
(Galán et al. 1995)	The day in which the sum of daily	2002	20 Jan	27 Dec
	pollen concentration reaches a	2004	11 Jan	15 Sep
	value over 1 % (start) and 99 %	2009	08 Mar	30 Oct
	(end) of the total yearly sum.	2012	7 Feb	30 Nov
(Andersen 1991)	The day in which the sum of daily	2002	26 Jan	01 Dec
	pollen concentration reaches a	2004	21 Jan	03 Aug
	value over 2.5 % (start) and 97.5 %	2009	14 Mar	29 Sep
	(end) of the total yearly sum.	2012	03 Mar	21 Sep
(Sánchez-Mesa et al. 2003)	The first day in which the daily	2002	17 May	03 Jun
	pollen concentration reaches	2004	11 Jan	15 Sep
	values over (start) and below	2009	07 May	30 Oct
	(end) 30 grains/m^3	2012	25 May	18 Jun
(Feher and Jarai-Komlodi 1997)	The first day in which the (start)	2002	05 Feb	19 Jun
	threshold reaches values over	2004	11 Apr	03 Jul
	and below (end) 3	2009	26 Sep	30 Oct
	grains/m^3 for 4 consecutive days	2012	22 Aug	12 Oct

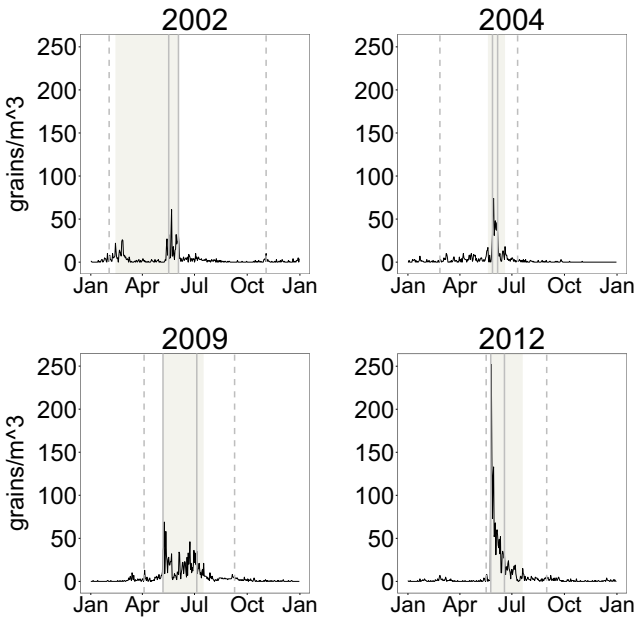


Fig. 1 Pollen concentrations for years 2002, 2004, 2009 and 2012 and definition of the season according to Nilsson and Persson (1981) (*vertical dashed line*) and Sánchez-Mesa et al. (2003) (*vertical solid line*). The shaded rectangle represents the latter approach relaxing the threshold to 15 grains/m^3

daily pollen concentration threshold, then the pollen season starts (ends) at the first (last) day that surpasses u.

In literature, pollen concentration levels show regional variations on pollen reactivity. For instance, according to Peternel et al. (2005) and Rantio-Lehtimäki et al. (1991), symptoms appear over 30 grains/m^3 in Finland and Croatia, while in Spain the first symptoms are observed between 25 grains/m^3 (Rodríguez-Rajo et al. 1983) and studies such as Sánchez-Mesa et al. (2003) use 30 grains/m^3. By far, the most common threshold level found in the literature is 30 grains/m^3 (Castellano-Méndez et al. 2005; Green et al. 2004; Sánchez-Mesa et al. 2003) which corresponds to the concentration at which the first allergy symptoms appear. Therefore, this level is selected as a representative in this study.

Features

In the pollen forecasting framework, the set of independent variables should contain relevant meteorological data as well as past pollen levels, as all of them are known to play a crucial role in predicting pollen concentrations. At the same time, due to the "curse of dimensionality" and to ease the computational burden, it is important to avoid including features which might not influence the pollen production at a certain time frame as it is. An example would be the rainfall registered 3 years before the forecast date: it will hardly be of interest to forecast the pollen season for that date. In our approach, feature relevance will be considered under different forecast horizons, thus enabling the proposed model to tell which set of independent variables are more influential for each horizon.

Cumulative pollen observations prior to the forecast date have been proved to serve as an indicator of the development stage of a plant (Ribeiro et al. 2007; Smith and Emberlin 2006). Correspondingly, 10- and 30-day cumulative sums of daily atmospheric concentrations prior to the forecast date are included as independent variables, along with the prior week daily concentrations for each date. Additionally, pollen accumulation within the year is also used as a proxy of the state of the plant.

The growth state of the buds is assumed to be linearly related to the amount of energy a plant has received (Cannell and Smith 1983). Sum of temperatures up to some point are usually considered as a good representation of this absorbed energy (Cannell and Smith 1983; Andersen 1991; Rodríguez-Rajo et al. 1983). Other authors (Pauling et al. 2014) however, use the concept of *chilling temperatures* and *forcing temperatures*, which are defined as the weighted sum of temperatures below or above certain levels for a fixed period. To allow for more flexibility, our study does not predefine the chilling and forcing periods, but chilling and forcing temperatures are calculated by accumulation of 30 and 60 days prior the forecast date:

$$F_{\text{sum}}(d) = \sum_{i=d-n}^{d} R_{\text{forc}}(i), \tag{2}$$

where

$$R_{\text{forc}}(i) = \begin{cases} 0 & \text{if } T(i) < T_{\text{forc}} \\ T(i) - T_{\text{forc}} & \text{if } T(i) \geq T_{\text{forc}} \end{cases}, \tag{3}$$

being d is the forecast date, n is the number of days which define the calculation period for the sum of forcing temperatures, $T(i)$ is the temperature for day i, and T_{forc} is the base temperature for forcing (all temperatures are in degrees Celsius). The same applies for chilling. Base forcing and chilling temperatures for a determined threshold are derived using geometrical relations from the reference of {1°C, 16°C} for the forcing period and {−6°C, 8°C} for the chilling period at thresholds of 10 grains/m^3 and 50 grains/m^3 respectively, as in Pauling et al. (2014).

The cumulative approach introduced for temperatures is also used to capture rainy and humid periods. Humidity and rain prevent pollen spread during pollination, and

humid and rainy weather causes grass species to become more abundant during the growing period of the plant urging to include short and long term periods prior the forecast date.

Pollen dispersion being a fundamental aspect of the problem, wind speed is recognized as an important influential factor (Palacios et al. 2000). Hence a 30-days cumulative sum of wind speed features is generated. For all climate data, similar as for the pollen concentrations, the prior 7 daily raw data observations are also included.

This leads to the availability of 70 features as detailed in Table 3, to which we added a dummy variable which represents the day of the year. This makes 71 features which are distributed in a matrix corresponding to the desired forecast horizon and the discretized class:

$$\left[\begin{array}{ccc|c} x_{1,1} & \dots & x_{1,71} & p_1 \\ \vdots & \ddots & \vdots & \vdots \\ x_{n,1} & \dots & x_{n,71} & p_n \end{array}\right] \rightarrow \left[\begin{array}{ccc|c} x_{1,1} & \dots & x_{1,71} & c_{1+t} \\ \vdots & \ddots & \vdots & \vdots \\ x_{n-t,1} & \dots & x_{n-t,71} & c_n \end{array}\right] \quad (4)$$

$$c_i = \begin{cases} 0 \ \textit{if} \ p_i < u \\ 1 \ \textit{if} \ p_i \geq u \end{cases}, \quad (5)$$

where p_i is the daily pollen observation at time i, t is the forecast horizon in number of days, and u is the threshold.

Table 3 Number of features generated by variable

	i	10	30	y	m	q	std
Pollen	7	1	1	1	–	–	–
Temperature	7	–	–	–	–	–	–
T_{forc}	–	–	–	–	1	1	–
T_{chill}	–	–	–	–	1	1	–
Humidity	7	–	–	–	1	–	–
Wind	7	–	–	–	1	–	–
Rain	7	–	–	–	1	–	1
Pressure	7	–	–	–	1	–	–
UV	7	–	–	–	1	–	–
Sun	7	–	–	–	1	–	–

i: previous $i \in [1, 7]$ day observation
10: previous 10-day cummulative sum
30: previous 30-day cummulative sum
y: year to date cummulative sum
m: previous month cummulative sum
q: previous 90-day cummulative sum
std: previous 15 days standard deviation

Random forest

Proposed for the first time in Breiman (2001), a random forest is an ensemble approach which leverages the performance of many simple decision trees that can be used to produce predictive models. It is a supervised learning procedure which combines several randomized decision trees and aggregates their predictions by averaging. The procedure operates over sample fractions of the data, grows a randomized tree predictor on each one and aggregate these predictors together.

The motivation to favor RF against other methods, like logistic regression (LR) is to avoid a correlation-based feature selection. It is known that LR is highly sensitive to variable collinearity and, as some features were generated from others , the parameterization of LR could be expensive in order to avoid overfitting. In this point we believe RF is a more robust approach. Given the relatively high number of instances and the presence of sudden high peaks in pollen concentrations as seen in Fig. 1, RF provides stability and accuracy in presence of outliers due to the *bagging* (Breiman 2001) technique.

Several decisions need to be made in order to build a RF model and to test its predictability. In order to optimize the execution, an analysis of the parameter search space needs to be done to precisely choose the parameter set up for each predictor.

To compare the performance of the different models resulting from the parameter set up, the area under the ROC curve generated by each model (AUC) is used. An ROC curve is a two-dimensional depiction of classifier performance (Fawcett 2003). The AUC of a classifier express the probability that the classifier will rank a randomly chosen instance which is correctly classified.

To test the optimal parameter set up the system performs a grid search to identify the best set of hyperparameters for the model based on the selected metric.

One of the strengths of random forests is that they are able to provide a measure of variable importance as a by-product of the model training. Breiman (2001) and Breiman (2002) proposed the evaluation of the importance of a variable x_i by adding up the weighted Gini impurity decreases for all nodes where x_i appears, and averaging over all the trees in the forest. Every node in a decision tree is designed to split the data set into two as a condition on a single variable. The measure on which the optimal split condition is chosen is called the Gini impurity. Thus when training a tree, it can be computed how much each feature decreases the weighted impurity in a tree being the average of these decreases the rank of the feature in the forest. This gives

a view on how important each variable is, and allows for further interpretability of the results.

Experimental design

Algorithm 1 System design

Require: $X_i = \{x_{i1}, x_{i2}...x_{i71}|c_i\} \quad i \in \{1...n\}$
Require: $u = threshold$
for all $[t, y]$ in $\{horizons, years\}$ **do**
  $AUC = 0$
  $X^S = \text{Preprocess}(X_i, u, t)$ ▷ Apply (4) (5)
  $X^S_{\text{test}} = X^S_{\text{y}}$
  $X^S_{\text{train}} = X^S - X^S_{\text{test}}$
  for $k \in [1, 15]$ **do**
    $parameter_k = \text{Grid.Search}(search\_space)$
    $model_k = \text{Random.Forest}(parameter_k, X^S_{\text{train}})$
    if $AUC \leq \text{AUC}(model_k)$ **then**
      $AUC = \text{AUC}(model_k)$
      $best = parameter_k$
    end if
  end for
  $prediction = \text{Random.Forest}(best, X^S_{\text{test}})$
  $E = \text{Error}(prediction)$
end for

The aim of this work is to help allergy patients and researchers in knowing in advance the period in which pollen concentrations will reach risk levels, and to identify the most influential factors for its prediction.

Given the very different shape of pollen concentrations and of the main pollination season across the observed years, as shown in Fig. 1, the experiments were tailored to find the best model available. From sudden high peak concentration levels in short periods to prolonged moderate atmospheric concentrations, the setup of the model has to be able to capture the inner available information to successfully predict the season.

Our approach is based on the idea that the pollen concentrations can be transformed into a binary classification problem where the featured instances represent influential factors. Daily pollen concentrations are mapped to {0, 1} depending on whether they are above the threshold (1) or not (0).

In order to avoid overfitting and to provide a more generalized overview of the performance of the model, a *leave-one-out* (LOO) cross validation approach was taken to split the data into train and test set. For each year, the observations of that year were taken out as a test set, leaving the remaining years to train the model so the final metrics consist of the average error for each iteration. By averaging the metrics from the LOO technique, the results provided are more representative than selecting, for instance, the last two years of the period as test set which would produce results very dependent on the characteristics of the selected years for testing.

As well, to provide a wider spectrum in order to give further information both for patients and researchers, the system provides forecasts for a wide set of time horizons, ranging from 1 day to 6 months. A forecast horizon of 15 days means that with the information available up to time t, the pollen concentration at day $t + 15$ is forecast.

Given the forecast horizons, vectors are build as in Eq. 4. The LOO approach is then applied by years. At each iteration, a random search is performed on the parameters taking into account the search boundaries and comparing the results for each set up. Finally, the best candidate is validated and its forecast metrics are provided. This process is summarized in Algorithm 1.

Results

In our setup, a set of forecast horizons were tested along with a threshold of 30 grains/m^3. An optimal parameterization of the RF model was done using the LOO technique for the years between 2000 and 2013, leaving the remaining year of each iteration as test set. At each iteration, several metrics are generated as an estimator of system performance for each horizon. Given the different characteristics of each year studied, this method provides generalization letting the model learn the particular characteristics of each pollination season.

The second aim of this study is to identify the best predictors of the main pollination season. It is intended to provide a robust and flexible framework to obtain a good estimation of the predictors according to different forecast horizons.

The performance of the model is tested by checking the error rate of the class when it is classified as positive, this is the daily pollen concentrations which surpass the threshold. This measure is known as sensitivity or recall, and it measures the proportion of atmospheric concentrations over the defined threshold of 30 grains/m^3 that were correctly classified as such. This measure is completed by the specificity which, on the contrary, measures the proportion of pollen concentrations below the threshold correctly classified. The global precision for both classes, above and below the threshold, is measured by the accuracy.

Forecast horizon

Table 4 shows the predictive metrics for each forecast horizon. Specificities and accuracies of over 90 % are achieved across the different horizons. The high values for specificity (true negative rate) indicate that the proposed model succeeds in identifying the periods of the main pollination

Table 4 Predictive Metrics. Totals based on LOO method for the study period between 2000 and 2013

Horizon	TP	FP	TN	FN	Sensitivity	Specificity	Accuracy	AUC
1	295	159	4198	36	0.891	0.964	0.958	0.972
5	282	244	4109	49	0.852	0.944	0.938	0.956
7	281	270	4081	50	0.849	0.939	0.932	0.939
15	302	333	4010	29	0.912	0.923	0.922	0.935
30	304	344	3984	27	0.918	0.921	0.918	0.935
60	308	326	3972	23	0.931	0.924	0.923	0.923
90	269	310	3972	48	0.849	0.928	0.924	0.922
120	264	318	3954	33	0.889	0.926	0.924	0.930
150	274	401	3686	22	0.926	0.902	0.904	0.924
180	262	320	3767	34	0.885	0.922	0.919	0.928

season with an acceptable rate of false negatives (predicting concentrations below the threshold inside the observed season). Figure 2 shows the prediction for 2001 with a forecast horizon of 1, 7, 15, and 90 days. Given the 30 grains/m^3 threshold-based definition of the pollination season, the model manages to identify season start and end dates having a maximal error of 17 days for season start with the 90 days horizon. On the other hand, sensitivities are somehow lower, but attaining percentages over 84 % in all cases. This means that the model struggles to predict concentrations below the threshold when they appear during the main pollination season, showing a high number of false positives

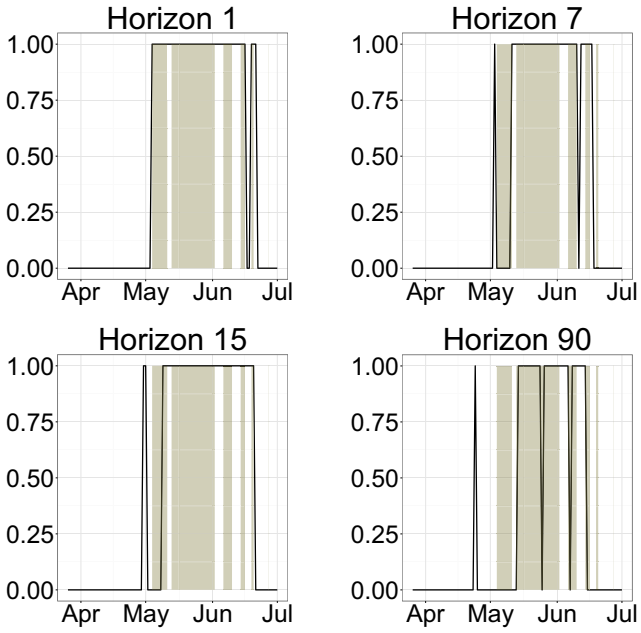


Fig. 2 Pollen observed over the threshold 30 grains/m^3 (*shaded*) for Apr-Jul 2001 with forecast with horizon 1, 7, 15 and 90 days (*solid lines*)

(FP). Weather conditions during the pollination season such as heavy sudden rainfall might directly affect airborne concentrations resulting in rapid drops of pollen concentrations below the threshold. As this information is not available in the predictors, the proposal does not identify this specific conditions. We also believe this is due to the fact that the classes are unbalanced, as the pollen concentrations over the selected threshold represent only around 7 % of the total observations. Even though at each iteration of the RF double trees were built, which means bootstrap sampling from the minority class and drawing the same number of cases from the majority class to finally aggregate the predictions, there might be an improvement in this metric by penalizing misclassification of the minority class or limiting the period studied to the potential dates where high concentrations appear. This however would imply making some assumptions over the period studied which could increase the presence of missing data. For instance, missing early season start dates, i.e., end of February, if the assumption limits the study period from March to August.

It is interesting to see how the model performs for the longer forecast horizons, which in general show lower specificity and higher sensibility and, consequently, lower accuracy. This means a higher number of false positives, as illustrated in Fig 2 for the 90-days threshold. In this case, the model incorrectly predicts an early start of the season. In general, for longer horizons, there is a clear tendency of expanding the main pollination season showing a more loose decision when defining the boundary dates, and consequently increasing the number of false positives as the horizon increases.

The model, on the other hand, manages to maintain a low and stable number of false negatives (FN) through the different horizons, which means that it succeeds in capturing the main periods where high concentrations appear.

It is noticeable that the decreasing accuracy pattern as the horizon increases is broken for the horizons of 60, 90, and 120 days, showing a small increase. This leads to think that the influential factors related to the previous winter period do play a key role in forecasting the start of the season.

Forecast horizon vs feature importance

In Fig. 3, the relative importance of the variables for a selected group of horizons is depicted. Each climate and pollen feature are labeled according to the method used to obtain it, as explained in Section 2. Hence, 'm' in the name of a variable denotes the accumulation of the daily featured data 30 days prior to the forecast date, 'q' represents the accumulation of daily data 90 days prior to the forecast date and 'y' the cumulative daily data from 1st January of the year in which the forecast lies. The data point of a variable x corresponding to the date d $- i$, being d the forecast date,

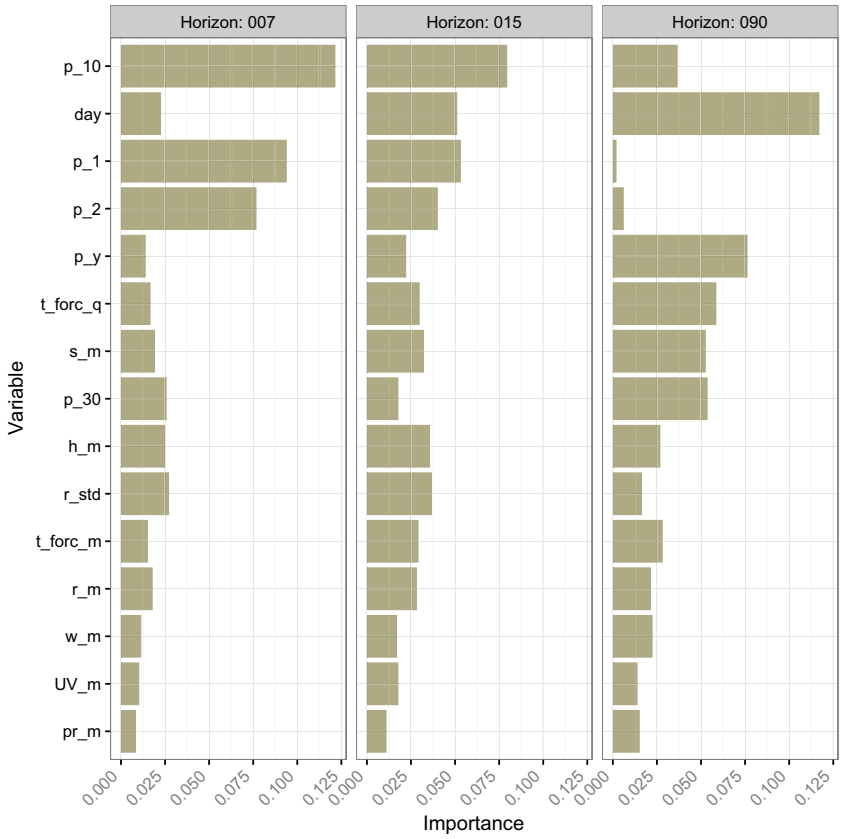


Fig. 3 Selection of the 15 most important variables by forecast horizon

is represented by x_i. Table 5 shows a detailed description of the most relevant features.

Clearly, for the 90 days horizon (rightmost graph), the influence of the forcing temperature is important for the prediction accounting a 6 % of the total importance compared to the 2 % and the 2.6 % for the 7 and 15 days horizons, respectively. On the other hand, the results for short-term horizons (leftmost graph) show that the most recent pollen concentrations are the most influential factors. Previous day (p_1) and the day before the previous (p_2) pollen

Table 5 Variable Description

Variable	Description
w_m	wind speed accumulation one month prior the forecast day
UV_m	ultraviolet radiation accumulation one month prior the forecast day
t_forc_q	accummulated forcing temperature 90 days prior the forecast day
t_forc_m	accummulated forcing temperature 30 days prior the forecast day
s_m	sun hours accumulation one month prior the forecast day
r_std	standard deviation of rainfall one month prior the forecast day
r_m	rainfall accumulation one month prior the forecast day
p_y	accummulated pollen daily concentration from the first of January until the forecast date
pr_m	pressure accumulation one month prior the forecast day
p_30	daily pollen accumulation one month prior the forecast day
p_2	pollen daily concentration 2 days prior the forecast date
p_10	pollen daily concentration 10 days prior the forecast date
p_1	pollen daily concentration 1 days prior the forecast date
h_m	humidity accumulation one month prior the forecast day
day	day of the year

observations add up around 17 % of importance for the 5 days horizon while the contribution for the same features decreases to a 8.5 % and barely 1 % as the horizon increases to 15 and 90 days, respectively.

For short and medium horizons, the most influential features among the meteorological variables are the monthly cumulative humidity (h_m) and rainfall (r_m) and the 15 days standard deviation of rainfall (r_std). It is known that rainy and humid conditions wash away airborne concentrations during the pollination of the flower. Cumulative temperature features (t_forc_q) and the monthly accumulation of sun hours (s_m) are believed to boost the plant formation during the pre-flowering, thus the model weights a total of around 11 % of importance for the 90 days horizon in contrast to the 4 % achieved for 7 days. It can be clearly seen in Fig. 3 how these two variables gain importance as the horizon increases.

Discussion

As seen in Table 4, our proposal achieves accuracies which compare favorably to other studies, for example, (Brighetti et al. 2013), who obtained a value of 89.1 % for sensitivity and a value of 30.4 % for specificity for the 5 days horizon and the same threshold, whereas our model achieves a 94.4 % for sensitivity and a 85.2 % for specificity. This boost in specificity of course means that our model success in capturing the precise period when the main concentrations appear, achieving a much lower error rate outside the main pollination period which leads to an accuracy of 93.8 %. For a 1-day horizon, our model achieves an accuracy of 95.8 % compared to an average of 94.5 % of the two validation sets in Castellano-Méndez et al. (2005), being able to provide a more general approach when forecasting regardless the nature of the pollen series. Compared to the findings from Nowosad (2016), which also uses RF, our model achieves 96.4 % specificity compared to an average of 97 % for the 1 day horizon which implies a slightly lower performance when identifying low pollen concentration levels. On the other hand, our proposal achieves a 89.1 % sensitivity compared to 61, 70, and 88 % in Nowosad (2016), providing a higher hit rate when identifying high levels. This is the cause for the higher global accuracy compared to the reference techniques.

Regarding variable importance, our proposal suggests that, for horizons over 90 days, the importance of the forcing temperatures is higher compared to its role in shorter horizons, supporting the proposal of the optimal parameters in Pauling et al. (2014). Additionally, chilling temperatures are not ranked within the most influential features, confirming conclusions from Pauling et al. (2014) which hinted that chilling temperatures might lead to smaller error reductions when forecasting. In addition, long-term horizons tend to weight more sunlight hours and rain features, which promote the formation of flowers during the pre-flowering months. Rainfall and humidity accumulations are positively related and influence the pollen release during the flowering period, in accordance to the findings of Aguilera et al. (2014). Hence, the model ranks these two features importances in accordance for short-term horizons.

Once the model is trained, producing forecasts takes less than a second on a 64-bit desktop Ubuntu machine with 6 cores and 32 GB of RAM. This of course allows the operational use of the approach.

Conclusions

The present paper introduces a new approach to forecast Poaceae pollen concentrations over different horizons making no assumptions on the phenology of the plant. It achieves consistent results in selecting the most influential factors given the forecast horizons. The selection of features from a purely data point of view is also consistent with different phenological studies while letting the model automatically select their relevance depending on the phases of the flower formation.

This study is tailored to help not only allergy patients but also research centers to prevent exposures to risk concentration levels for long-term horizons providing consistency up to 120 days prior the forecast data point. The model was tested on data from years 2000 to 2013, showing its adaptation and generalization regardless the specific characteristics of each pollen season.

The model proposed extends and supports the knowledge about the influence of meteorological factors on Poaceae pollen seasons. Although the results are promising, further efforts are required concerning the selection and generation of different features. Also, a wider experiment, using data from different sites, could shed more light into this interesting subject.

Acknowledgments This work has been partially funded by Ministerio de Economía y Competitividad, Gobierno de España, through a *Ramón y Cajal* grant (RYC-2012-11984).

The authors would like to thank Patricia Cervigón (Comunidad de Madrid) and Montserrat Gutiérrez Bustillo (Universidad Complutense de Madrid) for his assistance in obtaining the data for this study.

The authors would also like to thank the anonymous reviewers for their useful, constructive, and valuable comments, which greatly improved the original version of the manuscript.

References

Aguilera F, Fornaciari M, Ruíz-Valenzuela L, Galán C, Msallem M, Dhiab A, la Guardia CD, del Mar Trigo M, nd F Orlandi TB (2014)

Phenological models to predict the main flowering phases of olive (Olea europaea l.) along a latitudinal and longitudinal gradient across the Mediterranean region. Int J Bioeteorology 59:629–641
Andersen TB (1991) A model to predict the beginning of the pollen season. Grana 30:269–275
Aznarte JL, Benítez Sánchez JM, Lugilde DN, de Linares Fernández C, de la Guardia CD, Sánchez FA (2007) Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models. Expert Syst Appl 32(4):1218–1225
Breiman L (2001) Random forest. Mach Learn 45:5–32
Breiman L (2002) Manual on seeting up, using and understanding random forest. Stat Dept University of California Berkley v3.1
Brighetti MA, Costa C, Menesatti P, Antonucci F, Tripodi S, Travaglini A (2013) Multivariate statistical forecasting modeling to predict Poaceae pollen critical concentrations by meteoclimatic data. Aerobiologia 30:25–33
Cannell M, Smith R (1983) Thermal time, chill days and prediction of budburst in Picea sitchensis. J Appl Ecol 20:269–275
Castellano-Méndez M, Aira MJ, Iglesias I, Jato V, González-Manteiga W (2005) Artificial neural networks as a useful tool to predict the risk level of Betula pollen in the air. Int J Biometeorology 49:310–316
Fawcett M (2003) Roc graphs: Notes and practical considerations for data mining researchers. Tech rep, HP Laboratories
Feher Z, Jarai-Komlodi M (1997) An examination of the main characteristics of the pollen seasons in Budapest, Hungary (1991-1996). Grana 36:169–174
Galán C, Emberlin J, Domínguez E, Bryant RH, Villamandos F (1995) A comparative analysis of daily variations in the Gramineae pollen counts at Cordoba, Spain and London, UK. Grana 34:189–198
Green BJ, Dettman M, Yli-Panula E, Rutherford S, Simpson R (2004) Atmospheric Poaceae pollen frequencies and associations with meteorological parameters in Brisbane, Australia: a 5 year record, 1994–1999. Int J Biometeorology 40:172–178
Jato V, Rodríguez-Rajo FJ, Alcázar P, Nuntiis PD, Galán C, Mandrioli P (2006) May the definition of pollen season influence aerobiological results? Aerobiologia 22:13–25
Myszkowska D (2014) Predicting tree pollen season start dates using thermal conditions. Aerobiologia 30:307–321
Nilsson S, Persson S (1981) Tree pollen spectra in the Stockholm region (Sweden), 1973–1980. Grana 20:179–182
Nowosad J (2016) Spatiotemporal models for predicting high pollen concentration level of Corylus, Alnus and Betula. Int J Biometeorology 60:843–855
Palacios IS, Molina RT, Rodríguez AFM (2000) Influence of wind direction on pollen concentration in the atmosphere. Int J Biometeorology 44:128–133
Pauling A, Gehrig R, Clot B (2014) Toward optimized temperature sum parametrizations for forecasting the start of the pollen season. Aerobiologia 30:45–57
Peternel R, Srnec L, Culig J, Hrga I, Hercog P (2005) Poaceae pollen in the atmosphere of Zagreb (Croatia), 2002–2005. Grana 45:130–136
Rantio-Lehtimäki A, Koivikko A, Kupias R, Mäkinen Y, Pohjola A (1991) Significance of sampling height of airborne particles for aerobiological information. Allergy 46:68–76
Ribeiro H, Cunha M, Abreu I (2007) Definition of main pollen season using logistic model. Ann Agric Environ Med 14:259–264
Rodríguez-Rajo F, Frenguelli G, Jato M (1983) Effect of air temperature on forecasting the start of the Betula pollen season at two contrasting sites in the south of Europe (1995-2001). Int J of Biometeorology 47:117–125
Sánchez-Mesa J, Smith M, Emberlin J, Allitt U, Caulton E, Galán C (2003) Characteristics of grass pollen seasons in areas of southern Spain and the United Kingdom. Aerobiologia 19:243–250
Smith M, Emberlin J (2006) A 30-day-ahead forecast model for grass pollen in north London, UK. Int J Biometeorology 50:233–242
de Weger LA, Bergmann KC, Rantio-Lehtimaki A, Dahl A, Buters J, Déchamp C, Belmonte J, Thibaudon M, Cecchi L, Besancenot JP, Galán C, Waisel Y (2013) Impact of Pollen. In: Sofiev M, Bergmann KC (eds) Allergenic Pollen, Springer Netherlands, pp 161–215. doi:10.1007/978-94-007-4881-1_6

Chapter 5

What are the most important variables for Poaceae airborne pollen forecasting?

Type:	Published Article
Title:	*What are the most important variables for Poaceae airborne pollen forecasting?*
Journal:	Science of The Total Environment
Authors:	Ricardo Navares & José Luis Aznarte
Published:	December 2016
Impact Factor:	5.589
Quartile:	Q1
DOI	10.1016/j.scitotenv.2016.11.096

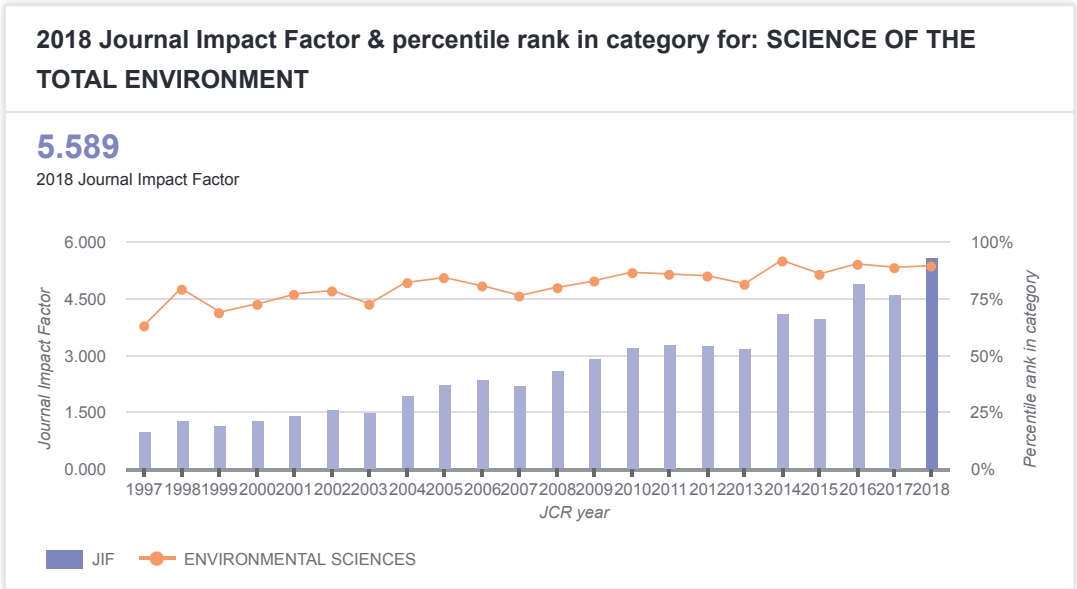


FIGURE 5.1: Impact factor Science of the Total Environment

ARTICLE IN PRESS

STOTEN-21398; No of Pages 9

Science of the Total Environment xxx (2016) xxx–xxx

Contents lists available at ScienceDirect

Science of the Total Environment

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

What are the most important variables for Poaceae airborne pollen forecasting?

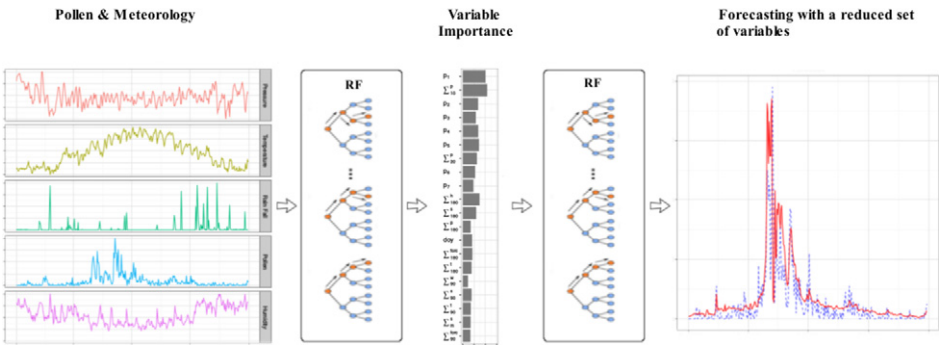
Ricardo Navares[a], José Luis Aznarte[b,*]

[a] Superior Technical School of Computer Engineering, UNED, Juan del Rosal, 16, Madrid 28040, Spain
[b] Department of Artificial Intelligence, UNED, Juan del Rosal, 16, Madrid 28040, Spain

HIGHLIGHTS

- A new data-driven approach to forecast airborne pollen concentrations is presented.
- The relative importance of variables is estimated and used to build the models.
- The assumption-free findings are coherent with previous phenological-based results.
- Non-parametric hypothesis testing confirms the statistical validity of the results.
- A reduced set of important variables renders more simple and accurate models.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 3 October 2016
Received in revised form 9 November 2016
Accepted 15 November 2016
Available online xxxx

Keywords:
Aerobiology
Prediction
Random forests
Nonparametric tests
Feature selection
Time series
Madrid

ABSTRACT

In this paper, the problem of predicting future concentrations of airborne pollen is solved through a computational intelligence data-driven approach. The proposed method is able to identify the most important variables among those considered by other authors (mainly recent pollen concentrations and weather parameters), without any prior assumptions about the phenological relevance of the variables. Furthermore, an inferential procedure based on non-parametric hypothesis testing is presented to provide statistical evidence of the results, which are coherent to the literature and outperform previous proposals in terms of accuracy. The study is built upon Poaceae airborne pollen concentrations recorded in seven different locations across the Spanish province of Madrid.

1. Introduction

The ability to anticipate future values of pollen concentrations in the air is crucial both for the allergic population, which can use predictions to foresee and adapt their needs concerning their outdoor presence, and for clinical institutions and public health organisms, which can prearrange resources before a predicted future outburst of pollen-related affections occurs. Many authors have faced the pollen forecasting problem in the last decades, with approaches that range from numerical models as SILAM (Sofiev et al., 2013), classic statistical time series analysis such as multivariate regression and nonlinear models (Cotos-Yáñez et al., 2004; Rodríguez-Rajo et al., 2004; Tassan-Mazzocco et al., 2015), to machine learning and computational intelligence, for example artificial neural networks

* Corresponding author.
E-mail address: jlaznarte@dia.uned.es (J. Aznarte).

http://dx.doi.org/10.1016/j.scitotenv.2016.11.096
0048-9697/

(Astray et al., 2016; Aznarte et al., 2007; Iglesias-Otero et al., 2015). Other approaches use numerical method to forecast pollen dispersions based on operational weather forecasts such as COSMO-ART (Vogel et al., 2008) or HYSPLIT (de Water et al., 2003).

However, there is a common underlying question which is independent of the chosen approach: which past information should be used when forecasting future values of pollen concentrations? For example, in univariate time series analysis, the models try to extract information from the past behaviour of just the pollen concentrations data (Aznarte et al., 2007). Of course, botany tells us that meteorology plays a crucial role in the development of the plants and hence in the pollen emission, and thus many authors have included meteorological variables in their models. In fact, there are studies about the influential factors in the growth state of plant buds (and, consequently, airborne pollen atmospheric concentrations) based on a phenological point of view (Cannell and Smith, 1983; Kmenta et al., 2016; Ribeiro et al., 2007; Smith and Emberlin, 2006), or based on the relation with climate conditions (Andersen, 1991; Pauling et al., 2014; Rodríguez-Rajo et al., 1983), or both. However, there is no consensus over which meteorological variables are more relevant.

For example, some studies employ meteorological daily data in order to forecast pollen concentrations, such as previous daily precipitation (Castellano-Méndez et al., 2005; Iglesias-Otero et al., 2015) or the relative humidity, wind speed and radiation upon the surface (Jones and Harrison, 2004). Others prefer the use of autoregressive indices, as, for example, thermal indices during plant formation season, in order to capture climatological information prior to pollen emission (Andersen, 1991; Myszkowska, 2014; Otero et al., 2013; Pauling et al., 2014). Some studies combine both approaches by using daily data and cumulative meteorological indices (Matyasovszky et al., 2015) researching the relationships between past and current weather conditions and current pollen levels (Deák et al., 2013), assuming that the timing of flowering is mostly driven by the accumulated temperature during a certain time period as the pollen release model in SILAM (Sofiev et al., 2013).

On the other hand, automatic feature selection is an important research field in computational intelligence. Feature selection techniques are used, among other reasons, to simplify the models in order to make them more interpretable and to shorten the training times. The idea behind automatic feature selection is that usually the data contains many features that are either redundant or irrelevant, and can thus be removed without incurring much loss of information.

An alternative to feature selection for pollen forecasting is factor analysis (Deák et al., 2013; Matyasovszky et al., 2015), which tries to find a latent representation of the observed variables that is good at explaining it. This technique can be used, for example, to avoid collinearity among the variables which might influence the predictive capability of the model. Given RF is robust enough against multicollinearity, its application would be sufficient to approach the problem.

The objective of this paper is to apply an automatic feature selection procedure to pollen forecasting, and validate it through statistical inference. By avoiding any *a priori* assumptions about the importance of the variables, neither based on the phenology of the plant nor on meteorological considerations nor on derived indices, we expect to question other author's assumptions and to provide new insight on the predictive power of the different available variables. Statistical inference through a nonparametric ranking-based statistical test (Friedman, 1937), along with a pairwise variable comparison in a post-hoc procedure, will allow to soundly establish the validity of the results.

As a case study, we chose to work on Poaceae airborne pollen concentrations in the Madrid region. Poaceae is the largest family of monocotyledonous flowering plants known as grasses and is considered to be one of the most important aeroallergens in Europe (Sánchez-Mesa et al., 2003). Poaceae pollen not only is one of the most prolific in Madrid but also is known to be one of the most aggressive, accounting 94% of positive reaction in patients (Subiza et al., 1995). The increase of allergy cases and the severity of the reactions (de Weger et al., 2013) motivates the need for prediction of Poaceae concentrations.

2. Materials and methods

2.1. Data description

Daily Poaceae airborne concentrations were provided by Red Palinológica de la Comunidad de Madrid and were obtained following the standard methodology of the Spanish Aerobiological Network (Galán Soldevilla et al., 2007). They are measured in grains per cubic meter of air. The observations come from 7 locations around the region of Madrid (Alcalá de Henares, Alcobendas, Aranjuez, Faculty of Pharmacy of Complutense University of Madrid, Getafe, Leganés and Villalba) and span periods that go from 2000 to 2013. Fig. 1 shows the location of the measuring stations.

On the other hand, we used weather observations consisting of average daily temperature in Celsius degrees, solar radiation in W/m^2, wind speed measured in m/s, daily rainfall in mm/h, pressure in mbar and degree of humidity in percentage. Data sets for locations Alcalá de Henares, Alcobendas, Aranjuez, Getafe and Leganés consist of 5 years of observations from 2005 to 2009. At the Faculty of Pharmacy data is available from 2001 to 2013 while in Villalba only three years are available starting 2007 to 2009 as shown in Fig. 2.

The geography of the Autonomous Community of Madrid provides several peculiarities given the locations studied. Villalba, which is located at 903 m above the sea level, has a mountain climate with a yearly average temperature of 10–11°C and a yearly average rainfall of 1250–1500 mm. As opposed to Villalba, Aranjuez has an elevation of 495 m above the sea level with an average yearly temperature above 13°C and a yearly average rainfall below than 400 mm. The remaining locations consist of metropolitan areas located between 594 and 668 m above the sea level.

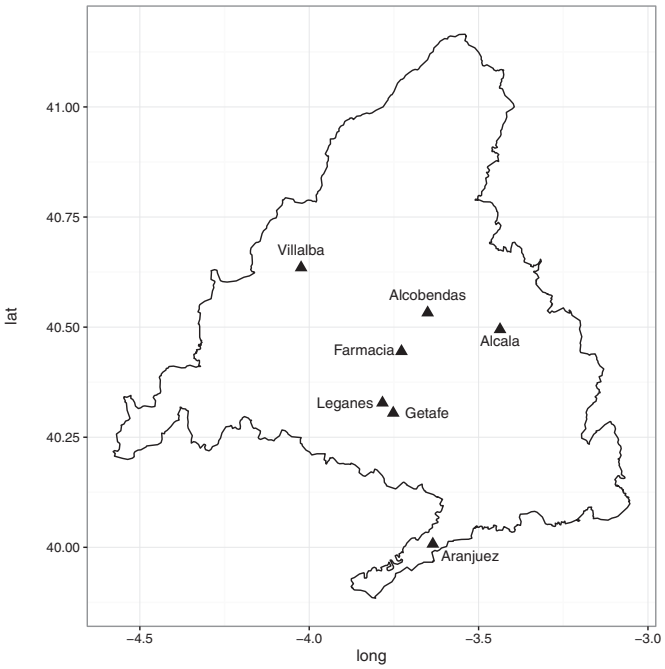


Fig. 1. Location of weather and pollen stations.

Please cite this article as: R. Navares, J. Aznarte, What are the most important variables for Poaceae airborne pollen forecasting?, Science of the Total Environment (2016), http://dx.doi.org/10.1016/j.scitotenv.2016.11.096

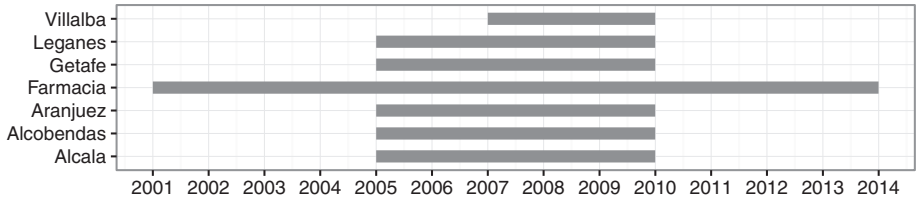


Fig. 2. Years available per observation station.

Very few missing data points were observed in the meteorological series so these were directly linearly interpolated to fill the gaps. On the other hand, pollen series contain missing observations in *a priori* critical months as February, March and April. In general, during these months pollen concentrations are meant to increase, thus the missing data is regressed within each year and across the years.

2.2. Features

In the pollen forecasting literature, the common approach is to include a set of independent variables which contain as much relevant information as possible. However, the core idea presented in this paper is to consider all the variables used by previous studies and compute a measure of variable importance to establish which set of independent variables are the most influential.

It is assumed that previous cumulative pollen observations serve as a proxy of the state of the plant (Ribeiro et al., 2007; Smith and Emberlin, 2006) and are considered as an indicator of plant development. Accordingly, cumulative sum of pollen concentrations within the year is included as an independent variable, along with the 10 and 30-day cumulative sums. As well, the presence of an autocorrelative pattern in the pollen time series suggests that previous daily concentrations have an influence in present values (Aznarte et al., 2007). In order to capture this inner relation between lagged airborne daily concentrations, previous single daily observations up to a week are also considered.

Weather conditions play a major role in determining the severity and length of the pollination season, as they are directly responsible for increases and decreases of the airborne pollen concentrations through its effect on the plants. The growth state of the buds is assumed to be linearly related to the amount of energy a plant has received (Cannell and Smith, 1983). This energy is usually represented by the sum of temperatures up to some point (Andersen, 1991; Cannell and Smith, 1983; Rodríguez-Rajo et al., 1983). Other recent studies (Pauling et al., 2014) introduce the concept of *chilling temperatures* and *forcing temperatures* instead, as they constitute a more optimal parameterization to represent the energy absorbed by the plants. Forcing temperatures are defined as the weighted sum of temperatures above a certain level for a fixed period:

$$F_{\text{sum}}(d) = \sum_{i=d-n}^{d} R_{\text{forc}}(i), \tag{1}$$

where

$$R_{\text{forc}}(i) = \begin{cases} 0 & if T(i) < T_{\text{forc}} \\ T(i) - T_{\text{forc}} & if T(i) \geq T_{\text{forc}} \end{cases}, \tag{2}$$

being d the forecast date, n the number of days which define the calculation period for the sum of forcing temperatures, $T(i)$ the temperature for day i, and T_{forc} the base temperature for forcing fixed at 8 °C. The same applies for chilling temperatures, but computing the temperatures below the threshold of 6 °C. To allow for more flexibility, our study does not predefine the chilling and forcing periods, but chilling and forcing temperatures are calculated by accumulation of 30, 90 and 180 days prior to the forecast date.

Sudden rainfall during the pollination cleans the air, washing pollen concentrations away, whereas wind speed contributes to pollen dispersion (Palacios et al., 2000), motivating the selection of the prior 7 daily raw data observations for these two variables. Conversely, rainy and humid periods prior to bloom is known to boost plant development when combined with sunny days. Thus, as was the case for chilling/forcing temperatures, the cumulative sums of 30, 90 and 180 days prior to the forecasting date are considered for rainfall and wind.

This leads to the availability of a total of 79 features as detailed in Table 1, to which we added a dummy variable which represents the day of the year. This makes 144 features which are distributed in a matrix corresponding to the desired forecast horizon. In order to train the model for a forecast horizon of $t = 1$ (one day-ahead forecast), vectors of the form $(x_{1,t}, \ldots, x_{144,t}|y_{t+1})$ where t is the time in which the forecast is done.

2.3. Random forests for regression

In the last years, there has been a growing interest in ensemble learning which aggregates the results of several independent models selected to boost their predictive performance. A well-known method is called *bagging* or bootstrap aggregating, proposed by Breiman (1996). Subsequently, Breiman (2001) presented a model called random forests (RF) which adds an additional layer of randomness to *bagging* providing robustness against overfitting with a limited number of parameters. These two characteristics favor RF against other computational intelligence methods such as neural networks.

The procedure combines several randomized regression trees generated over sample fractions of the data, and aggregates their prediction by averaging. This averaging process mitigates the influence of outlier data points giving RF advantage over other common

Table 1
Summary of features generated by variable.

	i	Σ_{10}	MA_i	Σ_{360}	Σ_{30}	Σ_{90}	Σ_{180}	Std
Pollen	7	1	4	1	1	1	1	–
Temperature	7	1	4	1	1	1	1	1
T_{forc}	–	–	–	1	1	1	1	–
T_{chill}	–	–	–	1	1	1	1	–
Humidity	7	1	4	1	1	1	1	1
Wind	7	1	4	1	1	1	1	1
Rain	7	1	4	1	1	1	1	1
Pressure	7	1	4	1	1	1	1	1
UV	7	1	4	1	1	1	1	1
Sun	7	1	4	1	1	1	1	1

i: previous $i \in [1, 7]$ day observation.
Σ_{10}: previous 10-day cumulative sum.
MA_i: max and min i-days moving average $i \in \{5, 15\}$.
Σ_{360}: year to date cumulative sum.
Σ_{30}: previous month cumulative sum.
Σ_{90}: previous 90-day cumulative sum.
Σ_{180}: previous 180-day cumulative sum.
Std: previous 15days standard deviation.

Please cite this article as: R. Navares, J. Aznarte, What are the most important variables for Poaceae airborne pollen forecasting?, Science of the Total Environment (2016), http://dx.doi.org/10.1016/j.scitotenv.2016.11.096

methods as support vector regression, which are highly sensitive in presence of outliers. As opposed to classification trees, the optimal split condition is the variance, which at the same time, is used to compute a measure of the importance of the independent variables.

The importance of a variable is estimated by measuring the increases in prediction error or variance when data from that variable is randomly permuted while the rest are left unchanged. The underlying idea is that if the variable is not important, then rearranging the values of that variable will not degrade prediction accuracy. For each tree, the prediction error is recorded before and after the permutation, the difference between both errors is averaged over all trees and normalized, thus providing the relative importances. The bigger the difference, the higher the importance of the permuted variable.

In order to check the performance of the forecasts, a general purpose error metric for numerical predictions as root mean squared error (RMSE), defined by Eq. (3), was used along with the coefficient of determination R^2, which indicates the proportion of variance of the observed data was predicted.

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{3}$$

where y_i is the observed ith data point, $\hat{y}_i$ the predicted and n the total number of data points in the test set.

2.4. Non-parametric hypothesis testing

Deciding if a variable is a better predictor than other is not a trivial task. Even if a measure of variable importance can be computed, it is not sufficient to evaluate it in a single case. In this work, we use statistical inference to investigate, in proper significance terms, which variables can be considered better than others.

When comparing a set of variables, the common statistical method for testing the differences between more than two related sample means is the repeated-measures ANOVA. Unfortunately, parametric hypothesis tests are based on assumptions such as normality or symmetry of the data distribution (Demšar, 2006). These assumptions are likely to be violated unless the data are well conditioned. On the other hand, nonparametric tests, whose main characteristic is that they are applied over nominal or ordinal data, are usually less restrictive (albeit also less robust). However, through a ranking-based transformation, they can be applied to continuous data.

The Friedman test (Friedman, 1937) is a multiple comparisons test to detect significant differences between a set of at least two samples. In our approach, the idea is to prove the existence of features which are more important than others across the years and the locations. The first step of the procedure is converting the original computed variable importance for each year and location to its correspondent rank within the set to obtain the average rank $R_j = \frac{1}{n}\sum_i r_i^j$ (where j denotes the feature, i refers to each year and location and n is the total number of pairs {location, year}). Then the null hypothesis of equality of medians is tested through the statistic

$$F = \frac{12n}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right], \tag{4}$$

where k is the number of variables and $F \sim \chi^2_{k-1}$. This test only allows to detect significant differences in the whole variable space without comparing each one against each other. A conversion of the rankings can be computed to obtain the p-value of each pair (Conover, 1999). The main drawback is that these p-values are not suitable for multiple comparison as the probability error of a certain comparison, does not take into account the remaining comparisons belonging to the family.

To overcome this limitation, it is needed to take into account that multiple tests are conducted via adjusted p-values which can be directly compared with a significance level α. A post-hoc test adjusts the value of α when dealing with multiple comparisons. One of the most commonly used of these adjustments is the Holm procedure (Holm, 1979) which adjusts the value of α by ordering, from smallest to largest, the p-values of each test. Then starting with the most significant p_i tests the hypothesis of $H_i : p_i > \alpha/(k-i)$, being k he total number of variables in our proposal. If H_i is rejected then allows to test H_{i+1} and so on.

An extension of Holm's step-down method was proposed by Shaffer (1986), which uses a logical relation between the combination of the hypotheses of all pairwise comparisons. For instance, if a variable v_1 is more/less important than v_2, it is not possible that v_1 is as important as v_3 and v_2 has the same importance as v_3. Based on this argument and following Holm's method, instead of rejecting $H_i : p_i \leq \alpha/(k-i)$, rejects $H_i \leq \alpha/t_i$, being t_i the maximum number of hypotheses which can be true given the number of false hypotheses in $j \in \{1, \ldots, i\}$

In order to contrast the difference between the importance of two variables we can use as an estimator the medians of the differences of each computed variable importance across locations and years (García et al., 2010). Being i the number of sets composed by each pair {location, year}, the median of the difference of each pair of variables Z_{v_i,v_j} is computed, then for each variable the average of the medians where the variable is involved is calculated as follows:

$$m_{v_i} = \frac{\sum_{j=1}^{k} Z_{v_i,v_j}}{k}, \tag{5}$$

where k is the total number of variables. The estimator of each pair of variables is defined as $m_{v_i} - m_{v_j}$, which provides how far a pair variables are in terms of importance.

2.5. Experimental design

Algorithm 1. Prediction.

Require: series; iterations = 50
for l in locations **do**
  for y in years(location) **do**
    $X^l_{\text{test}} = X^l_y$
    $X^l_{\text{train}} = X^l - X^l_{\text{test}}$
    for i in iterations **do**
      $\text{RF}_i = \text{RF}(X^l_{\text{train}})$
      $\text{importance}[y][i] = \text{RF.importance}(X^l_{\text{train}})$
      $\text{prediction}[y][i] = \text{RF}_i(X^l_{\text{test}})$
    end for
  end for
  $\text{total.importance}[l] = \text{average}(\text{importance}[y][i])$
  $\text{total.prediction}[l] = \text{average}(\text{prediction}[y][i])$
end for

Algorithm 2. Nonparametric test.

Require: series; iterations = 50; $i = 0$
Require: $M \leftarrow NULL$
for all $[l, y]$ in {location, years} **do**
  $\text{var.imp} \leftarrow NULL$
  for all iterations **do**
    $\text{var.imp} = \text{var.imp} + \text{RF.imp}(\text{series}_{l,y})$
  end for
  $\text{var.imp} = \text{var.imp}/\text{iterations}$
  $M[i,] = \text{var.imp}$
  $i++$
end for
$F = \text{friedman.test}(M)$
if $F \leq \alpha$ **then**
  $\text{post.hoc}(M)$
end if

Please cite this article as: R. Navares, J. Aznarte, What are the most important variables for Poaceae airborne pollen forecasting?, Science of the Total Environment (2016), http://dx.doi.org/10.1016/j.scitotenv.2016.11.096

In order to study the importance of the different available variables for pollen forecasting, the experiment is divided in three parts. Firstly, we use the full set of variables to build models designed to predict one day-ahead pollen concentrations for each location. From these models, we obtain a first insight of the most important variables in the different locations. Subsequently, in order to generalize and verify if the average rankings of variables per location are statistically significant, we apply statistical inference through non-parametric tests applied to the importance obtained per location and per year. Finally, we use the results of the hypothesis testing process to compare the models, in terms of forecasting precision and computational cost, with different subsets of the most important variables.

The first part of the experiment faces the one day-ahead pollen concentration forecasting problem in a standard fashion. By using a *leave-one-out* setup, we split the series into training and testing set at each location, saving one of the available years in each iteration and training the models with the rest. Given that one year can be seen in this case as one observation, and given the limited amount of observations, the differences from other cross validation techniques, such as 10-fold cross validation, are not expected to be substantial. The process is summarized in Algorithm 1. Through this cross validation approach, we obtain results which are more independent of the particular characteristics of each year, in the form of point-forecasts and estimates of the relative variable importances for each location. From these estimates we expect to see already some common patterns in the set of variables which are important across the different locations.

However, in order to produce a more rigorous and general result, independent from the location and the particular shape of the pollen curve of each year, the second part of the experiment employs a different partition of the data. The idea is distributing the importances resulting from the RF in a matrix of dimension $N \times M$ where N represents each pair {location, year} present in the dataset and M the available variables. Given the stochastic nature of RF, an iterative approach was designed to avoid the likelihood of overfitting, the final result consisting of the average of 50 instances of each model. Over the resulting matrix a Friedman test is conducted to give evidence on whether to proceed with the post-hoc analysis. This procedure is outlined in Algorithm 2.

Finally, with the results of the statistical inference process, we want to compare the forecasts obtained using the full set of variables in the first part of the experiment with those obtained by using reduced sets of only the most important ones. Precisely we chose to select the 5 and the 15 most important variables.

3. Results

3.1. Predicting with all the variables

Firstly, we applied RF using a leave-one-out setup, building the models with the full set of variables at our disposal. The models

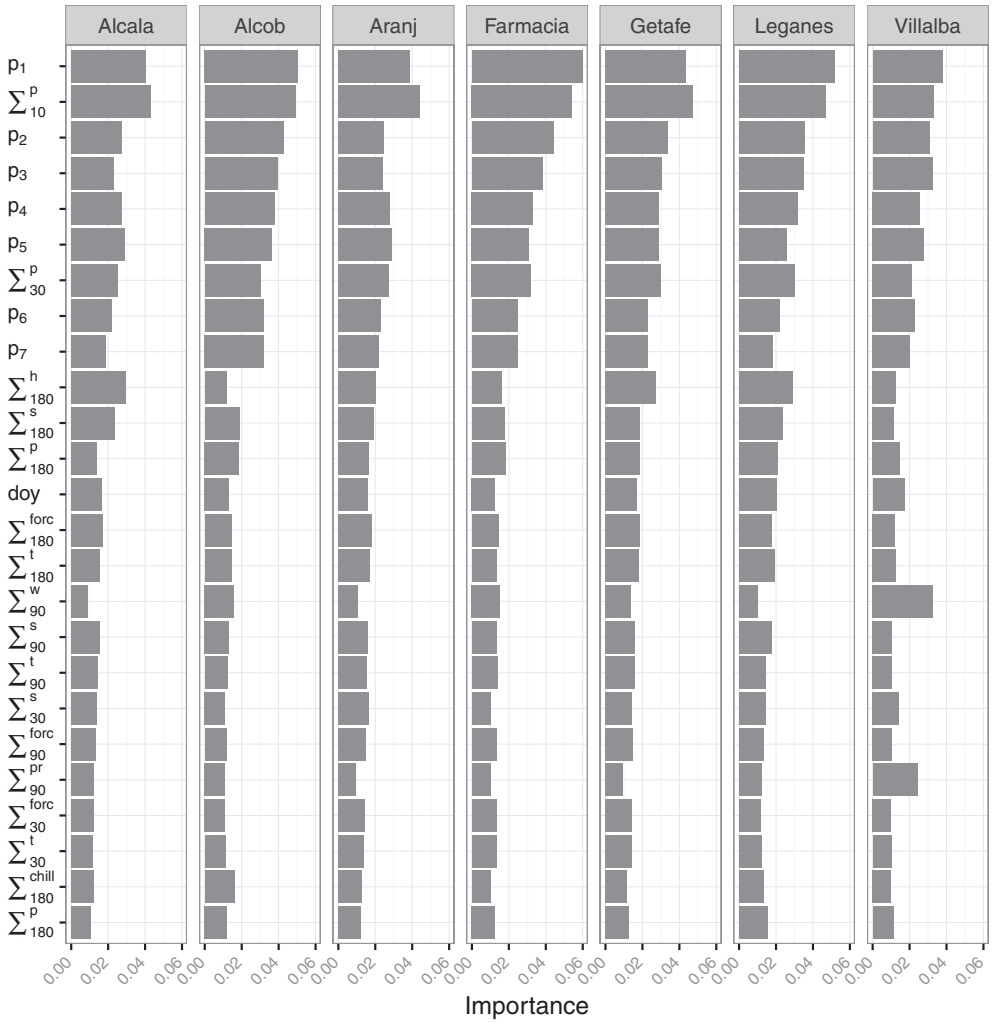


Fig. 3. Relative importance of the most important variables by location (variables are ordered by the sum of their importance across locations).

Please cite this article as: R. Navares, J. Aznarte, What are the most important variables for Poaceae airborne pollen forecasting?, Science of the Total Environment (2016), http://dx.doi.org/10.1016/j.scitotenv.2016.11.096

Table 2
Average ranking of the 16 most important features.

i	Variable	Ranking	i	Variable	Ranking
1	Σ^p_{10}	4.121	9	p_5	8.682
2	p_1	4.463	10	doy	8.682
3	p_2	6.902	11	Σ^w_{90}	8.853
4	p_3	6.829	12	p_7	10.365
5	Σ^h_{180}	7.756	13	Σ^t_{90}	10.878
6	Σ^p_{30}	8.121	14	s_3	11.048
7	p_4	8.512	15	Σ^s_{30}	11.097
8	p_6	8.658	16	MA^t_5	12.195

were trained independently for all the years available at each location, leaving one different year on each iteration as a test set. At each location, the relative importance of the variables, averaged by iteration and test sets, is obtained and shown in Fig. 3.

From this figure, it seems clear that the relation among the four most important variables p_1, Σ^p_{10}, p_2 and p_3 is maintained across locations . In fact, the ordering of relative importances remains very similar for all of them except for Villalba, where wind plays a much more crucial role. At this location, the importance of daily wind speed accumulated during 90 days prior the forecast date (Σ^w_{90}) accounts for more than 3% of importance, which is much higher than in the rest of locations. However, as we can see in Fig. 1, Villalba is the most northern and western site. Furthermore, it is located in the Guadarrama mountains at 903 m above the sea level while for example Aranjuez and Getafe have an elevation of 495 m and 622 m respectively. The particular meteorological conditions of Villalba, related to mountain climate, produce a higher correlation of wind with pollen concentrations during the period of study. For instance, during those years Σ^w_{90} is correlated at 32.69% with the daily pollen concentration in Villalba, while in the same period at Alcalá is 12.74% and a 8.64% when the full study period (2005–2009) is considered. This explains the increase in its importance at this location with respect to the rest. Similar applies to Σ^{pr}_{180}, which also gains importance due to the elevation, dropping average daily pressure and influencing flower formation and consequently pollen release.

Notwithstanding, the rest of the patterns of variable importance are quite stable across locations. Variable Σ^p_{10} keeps the best rank among them except for Farmacia, Villalba and Leganés, where its position is exchanged with the second most important variable from the test (p_1). On the other hand, ranks for p_2 and p_3 are perfectly maintained accross all locations as well as their relation to the top 2 ranked features.

3.2. Nonparametric tests to compare variable importance

As stated above, the relative importance of the variables computed by location seems quite stable. However, in order to give statistical evidence of the existence of features which are more influential than others, a non-parametric test was used. To do so, a RF model was trained, in this case for each location and for each available year, and the relative importance for each variable was thus computed. Given the large number of variables, the pairwise computation is expensive, hence in this case the hypothesis test is applied on a reduced set of variables including those which represent more than 1% of the total variance, which leaves us with 16 variables.

Out of this setup the Friedman statistic obtained is $F = 148.09$ which is distributed according to chi-square with 15 degrees of freedom with a critical values of $\chi^2_{15} = 24.99$ at $\alpha = 0.05$, leading to a computed p-value of 1.11e − 10, which strongly suggests the existence of significant differences among the variables. Table 2 shows the average Friedman ranking of the variables.

Due to the fact that the null hypothesis for Friedman's test is rejected, a post-hoc test can be applied to detect the pairs which produce the differences. Table 3 shows the contrast estimation of medians of the importance in percentage, and it is noticeable how p_1 and Σ^p_{10} obtain, respectively, an average of around 6% and 5.5% more importance than other variables, supporting the conclusions from Friedman's test. These two variables are followed by p_2 and p_3 which outperform around 3% and 2.2%, respectively, the remaining variables. On the other hand, MA^t_5 achieves the lowest relative importance across all stations and years.

Carrying out a post-hoc pairwise test will tell the evidence in the differences among pairs of variables. Table 4 and Fig. 4 show the rejected hypothesis (p-value $\leq\alpha$) with a significance level of $\alpha = 0.05$ for each compared pair. It can be seen how there is evidence that Σ^p_{10} and p_1 significantly differ from most of the other variables, although there is no evidence that they differ from each other. As shown above, the contrast estimation and the ranks from Friedman indicate that these two variables retain the highest importance among the whole set. Hence, there is statistical evidence that Σ^p_{10} and p_1 can be considered, in general, the most influential variables regardless the year and location.

There also exists evidence of differences between p_2 and p_3 and the group composed by {MA^t_5, Σ^s_{30}, s_3, Σ^t_{90}, p_7} whose higher rank is 10.36, separating the importance of these two variables from the lowest ranked group in this study.

Table 4 does not show clear distinction between {Σ^p_{10}, p_1} and {p_2, p_3} but, referring to the logical relation between the combination of the pairwise hypotheses proposed by Shaffer (1986), there is

Table 3
Contrast estimation in %.

	Σ^t_{90}	Σ^p_{10}	p_6	doy	Σ^s_{30}	p_1	Σ^w_{90}	s_3	p_2	Σ^h_{180}	p_4	p_3	p_5	MA^t_5	p_7	Σ^p_{30}
Σ^t_{90}	0.00	−5.45	−0.95	−0.64	−0.03	−5.98	−0.72	−0.09	−3.06	−1.51	−1.26	−2.21	−1.30	0.12	−0.71	−1.19
Σ^p_{10}	5.45	0.00	4.50	4.81	5.43	−0.52	4.73	5.36	2.39	3.94	4.20	3.24	4.15	5.57	4.74	4.26
p_6	0.95	−4.50	0.00	0.31	0.92	−5.03	0.23	0.86	−2.11	−0.56	−0.31	−1.26	−0.35	1.07	0.24	−0.24
doy	0.64	−4.81	−0.31	0.00	0.62	−5.33	−0.07	0.55	−2.42	−0.87	−0.61	−1.57	−0.66	0.77	−0.07	−0.54
Σ^s_{30}	0.03	−5.43	−0.92	−0.62	0.00	−5.95	−0.69	−0.07	−3.04	−1.48	−1.23	−2.18	−1.28	0.15	−0.68	−1.16
p_1	5.98	0.52	5.03	5.33	5.95	0.00	5.26	5.88	2.91	4.47	4.72	3.77	4.67	6.10	5.27	4.79
Σ^w_{90}	0.72	−4.73	−0.23	0.07	0.69	−5.26	0.00	0.62	−2.34	−0.79	−0.54	−1.49	−0.58	0.84	0.01	−0.47
s_3	0.09	−5.36	−0.86	−0.55	0.07	−5.88	−0.62	0.00	−2.97	−1.42	−1.16	−2.12	−1.21	0.22	−0.62	−1.09
p_2	3.06	−2.39	2.11	2.42	3.04	−2.91	2.34	2.97	0.00	1.55	1.81	0.85	1.76	3.19	2.35	1.87
Σ^h_{180}	1.51	−3.94	0.56	0.87	1.48	−4.47	0.79	1.42	−1.55	0.00	0.25	−0.70	0.21	1.63	0.80	0.32
p_4	1.26	−4.20	0.31	0.61	1.23	−4.72	0.54	1.16	−1.81	−0.25	0.00	−0.95	−0.05	1.38	0.55	0.07
p_3	2.21	−3.24	1.26	1.57	2.18	−3.77	1.49	2.12	−0.85	0.70	0.95	0.00	0.91	2.33	1.50	1.02
p_5	1.30	−4.15	0.35	0.66	1.28	−4.67	0.58	1.21	−1.76	−0.21	0.05	−0.91	0.00	1.42	0.59	0.11
MA^t_5	−0.12	−5.57	−1.07	−0.77	−0.15	−6.10	−0.84	−0.22	−3.19	−1.63	−1.38	−2.33	−1.42	0.00	−0.83	−1.31
p_7	0.71	−4.74	−0.24	0.07	0.68	−5.27	−0.01	0.62	−2.35	−0.80	−0.55	−1.50	−0.59	0.83	0.00	−0.48
Σ^p_{30}	1.19	−4.26	0.24	0.54	1.16	−4.79	0.47	1.09	−1.87	−0.32	−0.07	−1.02	−0.11	1.31	0.48	0.00

Please cite this article as: R. Navares, J. Aznarte, What are the most important variables for Poaceae airborne pollen forecasting?, Science of the Total Environment (2016), http://dx.doi.org/10.1016/j.scitotenv.2016.11.096

evidence of difference between Σ^p_{10} and p_1 and p_4, something which does not exist for p_2 and p_3. This leads to conclude that Σ^p_{10} and p_1 are more important than p_4, but there is no evidence that p_2 and p_3 are more important than p_4 so the conclusion is that Σ^p_{10} and p_1 are more influential than p_2 and p_3.

In summary, the non-parametric tests prove the existence of features which are more relevant than others. Among them, there are groups of features which significantly differ from other groups, while there is no statistical evidence of differences between group members. This means that, within a group, it is expected that their members maintain or alternate their ranks within the bounds of the rank of the group. For instance, we have seen that Σ^p_{10} and p_1 differ from the rest of the variables but do not differ from one another, constituting the top ranked group. As shown in Table 2, these variables have a Friedman's rank of 4.121 and 4.463 respectively, which is translated to position 1 and 2 in importance. It is expected that Σ^p_{10} and p_1 maintain their correspondent position or that p_1 takes position 1, dragging Σ^p_{10} to position 2 as they do not differ from each other.

3.3. Predicting with reduced sets of variables

Once we investigated the relative importance of the variables through statistical inference, we moved on to empirically verify how does selecting a subset of the most important variables affect the precision of the models and their computational efficiency.

In order to do so, we repeated the experiment in Section 3.1 for two reduced set of variables. First we selected the 15 variables which were proven to be the most important according to the tests of Section 3.2, and secondly we repeated the experiment reducing the set to just the 5 most important variables. Then, for both reduced sets and for each location, 50 instances of a RF model were trained to produce one day-ahead forecasts. Again, the models were trained through the LOO cross-validation approach by entire years.

Table 4
Pairwise rejected hypothesis at $\alpha = 0.05$ with unadjusted p-value and adjusted Holm and Shaffer p-values.

i	Hypothesis	p	p_{holm}	p_{shaff}
1	Σ^p_{10} vs MA^t_5	1.62e−14	1.94e−12	1.94e−12
2	p_1 vs MA^t_5	1.94e−13	2.31e−11	2.04e−11
3	Σ^p_{10} vs Σ^s_{30}	3.27e−11	3.86e−09	3.43e−09
4	Σ^p_{10} vs s_3	4.47e−11	5.23e−09	4.70e−09
5	Σ^t_{90} vs Σ^p_{10}	1.32e−10	1.53e−08	1.38e−08
6	Σ^s_{30} vs p_1	2.81e−10	3.23e−08	2.95e−08
7	p_1 vs s_3	3.78e−10	4.31e−08	3.97e−08
8	p_2 vs MA^t_5	5.09e−10	5.75e−08	5.34e−08
9	Σ^t_{90} vs p_1	1.06e−09	1.19e−07	1.11e−07
10	Σ^p_{10} vs p_7	2.89e−09	3.20e−07	3.03e−07
11	p_1 vs p_7	1.99e−08	2.18e−06	2.08e−06
12	Σ^s_{30} vs p_2	2.31e−07	2.52e−05	2.42e−05
13	s_3 vs p_2	2.96e−07	3.19e−05	3.11e−05
14	p_3 vs MA^t_5	4.82e−07	5.16e−05	5.06e−05
15	Σ^t_{90} vs p_2	6.91e−07	7.33e−05	7.26e−05
16	Σ^p_{10} vs Σ^w_{90}	6.80e−06	0.00	0.00
17	p_2 vs p_7	7.58e−06	0.00	0.00
18	Σ^p_{10} vs doy	1.44e−05	0.00	0.00
19	Σ^p_{10} vs p_5	1.44e−05	0.00	0.00
20	Σ^p_{10} vs p_6	1.60e−05	0.00	0.00
21	Σ^h_{180} vs MA^t_5	2.43e−05	0.00	0.00
22	Σ^p_{10} vs p_4	2.98e−05	0.00	0.00
23	p_1 vs Σ^w_{90}	2.98e−05	0.00	0.00
24	doy vs p_1	6.00e−05	0.01	0.01
25	p_1 vs p_5	6.00e−05	0.01	0.01
26	p_6 vs p_1	6.62e−05	0.01	0.01
27	Σ^s_{30} vs p_3	6.62e−05	0.01	0.01
28	s_3 vs p_3	8.04e−05	0.01	0.01
29	MA^t_5 vs Σ^p_{30}	0.00	0.01	0.01
30	p_1 vs p_4	0.00	0.01	0.01
31	Σ^p_{10} vs Σ^p_{30}	0.00	0.01	0.01
32	Σ^t_{90} vs p_3	0.00	0.01	0.01
33	p_4 vs MA^t_5	0.00	0.04	0.04
34	p_1 vs Σ^p_{30}	0.00	0.04	0.04
35	Σ^p_{10} vs Σ^h_{180}	0.00	0.05	0.04
36	p_6 vs MA^t_5	0.00	0.07	0.06

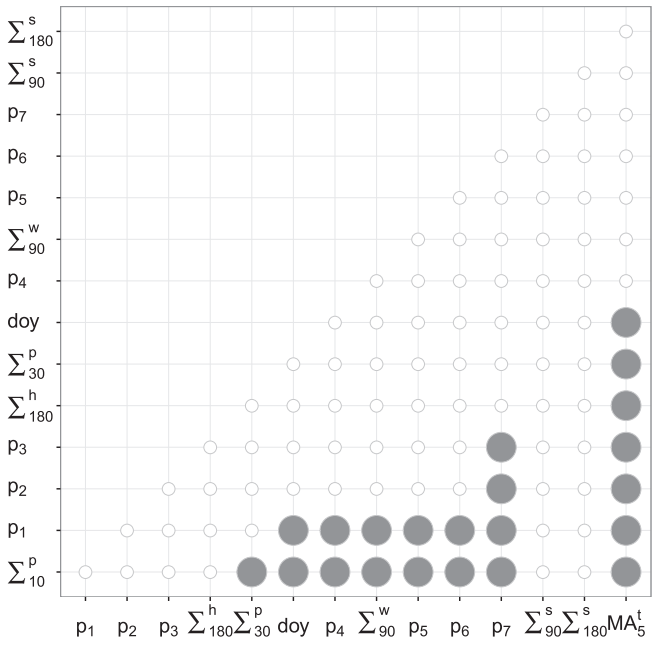


Fig. 4. Pairwise rejected hypothesis at $\alpha = 0.05$ with adjusted Shaffer p-values. The dots indicate pairs of variables which are not equally important.

The predictive performance of the models in terms of RMSE and R^2, for the setup using all variables (setup A), the reduced sets of 15 (setup B) and 5 (setup C) most important variables is shown in Table 5. Their respective execution times (for a single instance) are shown in Table 6.

From the study of both tables, we see that the reduction from the setup A using 144 variables to just the 15 most important of them (setup B) implies a reduction of almost a third of the execution time. Most importantly, models built with the reduced set of 15 variables yield better RMSE results in average (only in Alcobendas there is a slightly worse error result). In terms of R^2, both setups are approximately equal. Therefore, a first conclusion is that removing redundant non-important variables helps the models to converge to better results.

Regarding setup C, which uses just the 5 most important variables, the execution time is cut to one half with respect to setup A. However, the RMSE and R^2 values are worst in average, which indicates that such a reduced set is not enough to capture the inner characteristics of the pollen time series. In other words, the reduction leaves some important variables out, resulting in worse models.

By closely studying the results, we have identified a pattern which could explain some of the difficulties of the model in capturing the

Table 5
Average RMSE and R^2 of the test years studied at each location.

Station	$RMSE_A$	$RMSE_B$	$RMSE_C$	R^2_A	R^2_B	R^2_C
Alcalá	19.05	18.06	18.64	0.62	0.57	0.50
Alcobendas	15.42	15.57	14.99	0.70	0.68	0.59
Aranjuez	16.64	16.41	18.05	0.64	0.58	0.48
Farmacia	18.66	17.64	18.16	0.63	0.59	0.50
Getafe	16.34	16.11	21.87	0.69	0.70	0.49
Leganés	18.43	16.41	20.22	0.71	0.71	0.54
Villalba	17.46	16.43	20.61	0.58	0.69	0.55
Average	17.43	16.66	18.90	0.65	0.64	0.52

Please cite this article as: R. Navares, J. Aznarte, What are the most important variables for Poaceae airborne pollen forecasting?, Science of the Total Environment (2016), http://dx.doi.org/10.1016/j.scitotenv.2016.11.096

Table 6
Average execution time (in seconds) for the models built using different subsets of variables.

Setup	# variables	Time
A	144	29.00
B	15	20.29
C	5	14.75

inner behaviour of the data series. Precisely, the best performing test year across locations (2007) obtain an average RMSE equal to 3.84 grains/m^3, while the average RMSE in the worst performing year (2010) goes up to 28.16 grains/m^3. Concretely, this situation is related to the usual appearance of sudden extreme pollen concentration peaks during the pollination season.

These findings indicate that the models consistently fail to forecast the whole height of the peaks, producing an increase of the error: the higher the peak, the bigger the error. Another example can be seen in Fig. 5, where it is shown how, for the Alcobendas site in 2008, there are a number of sudden peaks, which result in a high average RMSE of 31.02 grains/m^3, while, in 2007 for the same location, the RMSE obtained was of 6.89 grains/m^3. This is most probably related to the lack of concentration peaks over 100 grains/m^3 for that year.

Finally, further evidence of this can be obtained from Table 5, were we can see that the model obtains low precision when forecasting concentrations exceeding 150 grains/m^3, as is the case for the Faculty of Pharmacy (Farmacia) and Alcalá where peak concentrations up to 400 grains/m^3 were observed.

4. Discussion

As we have seen, there is statistical evidence of the existence of relevant groups of variables when forecasting one day-ahead airborne pollen concentrations. The proposed approach succeeds in ranking and identifying these influential features, finding that previous pollen daily observations and 10-day cumulative airborne concentrations (which have both been proved to serve as indicators of the state of development of the plant (Ribeiro et al., 2007; Smith and Emberlin, 2006)) are the two top ranked features. These results, entirely based on the data and free from *a priori* assumptions are supported by nonparametric hypothesis tests and *post-hoc* procedures, and are coherent with previous phenological studies.

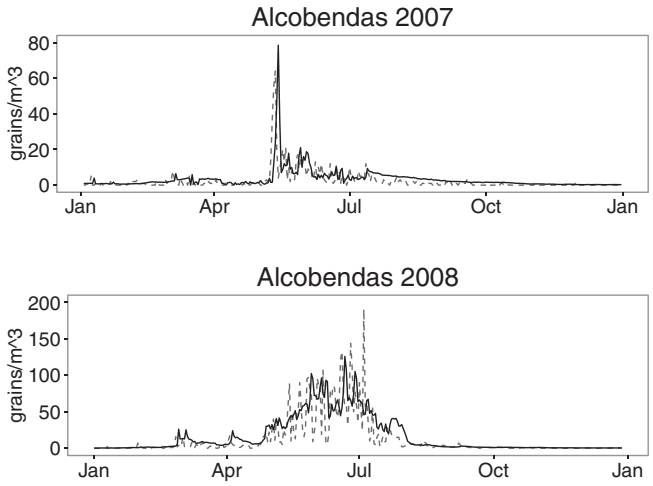


Fig. 5. Forecast (solid line) and observed (dashed) pollen concentration values for two years in the site of Alcobendas.

They have immediate practical application, as proven by the fact that subsets of top-ranked variables produce better results while reducing computational complexity.

However, the appearance of sudden high airborne concentrations increases the error on the performance of the model, as it is not able to capture extreme pollen levels given the limited amount of observations of this class. This phenomenon is common to other approaches, and could be investigated in the framework of the extreme value theory. However, it does not seem to be a particularly worrying issue, as the metric could be improved by limiting the observed concentrations to avoid the high peaks without hampering the usefulness of the proposal. For example, it could make sense to impose an upper limit to the values of the series as in Navares and Aznarte (2016). This limit should be related to the concentration levels which are considered to be risky for human health, for instance, symptoms are reported to appear over 30 grains/m^3 in Finland and Croatia (Peternel et al., 2005; Rantio-Lehtimäki et al., 1991), while in Spain the first symptoms are reported at 25 grains/m^3 (Rodríguez-Rajo et al., 1983). Another option is to transform the data into a logarithmic scale (or similar) as in Aznarte et al. (2007).

However, the results presented in this work are comparable in terms of accuracy to other studies such as Iglesias-Otero et al. (2015), which reports an average R^2 of 0.66 compared to an average 0.65 in our setup B with 15 variables. Furthermore, the approach proposed by Iglesias-Otero et al. (2015) uses artificial neural networks (ANN) which, as shown in the research, is strongly dependent on the configuration of the network. RF is considered as a more robust approach to mitigate overfitting, limiting the assumptions to be taken by the practitioner, avoiding decisions such as network structures and requiring minimal preprocessing of the data. The same applies to Astray et al. (2016) who achieves a RMSE of 22.56 on the best topology of the ANN. In average, our models achieve an RMSE of 17.43 across all configurations, being the setup B with 15 variables the best configuration: it obtains an RMSE of 16.66. These values clearly outperform the best regression models proposed by Csépe et al. (2014) which achieves a RMSE of 28.26 on its best performing algorithm (M5P), which is an implementation for regression trees. The *bagging* technique used in RF gives an edge over M5P algorithm reducing estimation variance and consequently prediction error. RF are relatively complex in computational terms, hence the focus on variable selection, which has shown that setup B (15 variables) as the most optimal among the configurations tested.

Additionally, the use of the LOO approach across multiple locations produces more generalized results as opposed to a predefined test set, on which the results heavily rely on the specific characteristics of the years selected to define the test set.

5. Conclusions

This paper presents a new approach to forecast airborne Poaceae pollen concentrations in the region of Madrid by identifying the most influential among the set of available variables. It provides statistical evidence, through non-parametric hypothesis tests, of the benefits of selecting the most influential variables for a one day-ahead forecast horizon.

Concretely, from all the considered pollen and meteorological variables, the data indicate that previous days pollen concentrations and cumulative sums of recent pollen concentrations are among the most important. They are followed by cumulative sums of humidity and solar radiation. These findings support, from a pure data-based point of view, the conclusions of previous phenology-based studies.

However, the statistical significance of the results from a single experiment are always to be questioned, and for this reason, in this paper, a sound statistical procedure based on Friedman tests and *post-hoc* analysis is used to support and validate the conclusions.

Please cite this article as: R. Navares, J. Aznarte, What are the most important variables for Poaceae airborne pollen forecasting?, Science of the Total Environment (2016), http://dx.doi.org/10.1016/j.scitotenv.2016.11.096

Finally, the usefulness of the proposal is shown by training the model with a reduced set of the most important variables, eliminating redundancies. This reduced model has been proven to increase the accuracy while limiting the computational burden.

The results clearly outperform those obtained by other authors, although there is still room for improvement and further research. For example, the models have difficulties in predicting extremely high concentrations, something that could be addressed by imposing upper limits according to risk levels for allergy patients. Also, a systematic exploration of the different combinations that can be drawn from the set of important variables should be studied, together with the optimization of the trade-off between the number of variables and the accuracy.

Acknowledgments

This work has been partially funded by the Ministerio de Economía y Competitividad, Gobierno de España, through a *Ramón y Cajal* grant (RYC-2012-11984).

The authors would like to thank Patricia Cervigón (Comunidad de Madrid) and Montserrat Gutiérrez Bustillo (Universidad Complutense de Madrid) for their assistance in obtaining the data for this study.

References

Andersen, T.B., 1991. A model to predict the beginning of the pollen season. Grana 30, 269–275.

Astray, G., Fernández-González, M., Rodríguez-Rajo, F., López, D., Mejuto, J., 2016. Airborne castanea pollen forecasting model for ecological and allergological implementation. Sci. Total Environ. 548–549, 110–121.

Aznarte, J.L., Benítez Sánchez, J.M., Lugilde, D.N., de Linares Fernández, C., de la Guardia, C.D., Sánchez, F.A., 2007. Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models. Expert Syst. Appl. 32 (4), 1218–1225.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 25, 123–140.

Breiman, L., 2001. Random forest. Mach. Learn. 45, 5–32.

Cannell, M., Smith, R., 1983. Thermal time, chill days and prediction of budburst in *Picea sitchensis*. J. Appl. Ecol. 20, 269–275.

Castellano-Méndez, M., Aira, M.J., Iglesias, I., Jato, V., González-Manteiga, W., 2005. Artificial neural networks as a useful tool to predict the risk level of Betula pollen in the air. Int. J. Biometeorology 49, 310–316.

Conover, W.J., 1999. Nonparametric methods. In: Wiley, B., O'Sullivan, M. (Eds.), Practical Nonparametric Statistics. John Wiley and Sons., pp. 233–305. http://dx.doi.org/10.1002/bimj.19730150311.

Cotos-Yáñez, T., Rodríguez-Rajo, F., Jato, M., 2004. Short-term prediction of Betula airborne pollen concentration in Vigo (NW Spain) using logistic additive models and partially linear models. Int. J. Biometeorol 48, 179–185.

Csépe, Z., Makra, L., Voukantsis, D., Matyasovszky, I., Tusnády, G., Karatzas, K., Thibaudon, M., 2014. Predicting daily ragweed pollen concentrations using computational intelligence techniques over two heavily polluted areas in Europe. Sci. Total Environ. 542–552, 476–477.

de Water, P.K.V., Keever, T., Main, C.E., Levetin, E., 2003. An assessment of predictive forecasting of *Juniperus ashei* pollen movement in the Southern Great Plains, USA. Int. J. Biometeorol 48, 74–82.

de Weger, L.A., Bergmann, K.C., Rantio-Lehtimaki, A., Dahl, A., Buters, J., Déchamp, C., Belmonte, J., Thibaudon, M., Cecchi, L., Besancenot, J.-P., Galán, C., Waisel, Y., 2013. Impact of pollen. In: Sofiev, M., Bergmann, K.-C. (Eds.), Allergenic Pollen. Springer Netherlands., pp. 161–215. http://dx.doi.org/10.1007/978-94-007-4881-1_6.

Deák, A., Makra, L., Matyasovszky, I., Csépe, Z., Muladi, B., 2013. Climate sensitivity of allergenic taxa in Central Europe associated with new climate change related forces. Sci. Total Environ. 442, 36–47.

Demšar, J., 2006, Dec. Statistical comparisons of classifiers over multiple data sets. J. Mach. Learn. Res. 7, 1–30.

Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc. 32, 674–701.

Galán Soldevilla, C., Cariñanos González, P., Alcázar Teno, P., Domínguez Vílches, E., 2007. Manual de Calidad y Gestión de la Red Espa nola de Aerobiología. Universidad de Córdoba .

García, S., Fernández, A., Luengo, J., Herrera, F., 2010. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: experimental analysis of power. Inf. Sci. 180, 2044–2064.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.

Iglesias-Otero, M.A., Fernández-González, M., Rodríguez-Caride, D., Astray, G., Mejuto, J.C., Rodríguez-Rajo, F.J., 2015. A model to forecast the risk periods of Plantago pollen allergy by using ANN methodology. Aerobiologia 31, 201–211.

Jones, A., Harrison, R., 2004. The effects of meteorological factors on atmospheric bioaerosol concentrations—a review. Sci. Total Environ. 326, 151–181.

Kmenta, M., Bastl, K., Kramer, M., Hewings, S., Mwange, J., Zetter, R., Berger, U., 2016. The grass pollen season 2014 in Vienna: a pilot study combining phenology, aerobiology and symptom data. Sci. Total Environ. 566–567, 1614–1620.

Matyasovszky, I., Makra, L., Csépe, Z., Sümeghy, Z., Deák, A., Pál-Molnár, E., Tusnády, G., 2015. Plants remember past weather: a study for atmospheric pollen concentrations of Ambrosia, Poaceae and Populus. Theor. Appl. Climatol. 122, 181–193.

Myszkowska, D., 2014. Predicting tree pollen season start dates using thermal conditions. Aerobiologia 30, 307–321.

Navares, R., Aznarte, J., 2016. Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features. Int. J. Biometeorol http://dx.doi.org/10.1007/s00484-016-1242-8.

Otero, J., García-Mozo, H., Hervás, C., Galán, C., 2013. Biometeorological and autoregressive indices for predicting olive pollen intensity. Int. J. Biometeorol 57, 307–316.

Palacios, I.S., Molina, R.T., Rodríguez, A.F.M., 2000. Influence of wind direction on pollen concentration in the atmosphere. Int. J. Biometeorology 44, 128–133.

Pauling, A., Gehrig, R., Clot, B., 2014. Toward optimized temperature sum parametrizations for forecasting the start of the pollen season. Aerobiologia 30, 45–57.

Peternel, R., Srnec, L., Culig, J., Hrga, I., Hercog, P., 2005. Poaceae pollen in the atmosphere of Zagreb (Croatia), 2002–2005. Grana 45, 130–136.

Rantio-Lehtimäki, A., Koivikko, A., Kupias, R., Mäkinen, Y., Pohjola, A., 1991. Significance of sampling height of airborne particles for aerobiological information. Allergy 46, 68–76.

Ribeiro, H., Cunha, M., Abreu, I., 2007. Definition of main pollen season using logistic model. Ann. Agric. Environ. Med. 14, 259–264.

Rodríguez-Rajo, F., Dopazo, A., Jato, V., 2004. Environmental factors affecting the start of pollen season and concentrations of airborne Alnus pollen in two localities of Galicia (NW Spain). Ann. Agric. Environ. Med. 11, 35–44.

Rodríguez-Rajo, F., Frenguelli, G., Jato, M., 1983. Effect of air temperature on forecasting the start of the Betula pollen season at two contrasting sites in the south of Europe (1995–2001). Int J. of Biometeorology 47, 117–125.

Sánchez-Mesa, J., Smith, M., Emberlin, J., Allitt, U., Caulton, E., Galán, C., 2003. Characteristics of grass pollen seasons in areas of southern Spain and the United Kingdom. Aerobiologia 19, 243–250.

Shaffer, J., 1986. Modified sequentially rejective multiple test procedures. J. Am. Stat. Assoc. 81, 826–831.

Smith, M., Emberlin, J., 2006. A 30-day-ahead forecast model for grass pollen in north London, UK. Int. J. Biometeorology 50, 233–242.

Sofiev, M., Siljamo, P., Ranta, H., Linkosalo, T., Jaeger, S., Rasmussen, A., Rantio-Lehtimaki, A., Severova, E., Kukkonen, J., 2013. A numerical model of birch pollen emission and dispersion in the atmosphere. Description of the emission module. Int. J. Biometeorol 57, 45–58.

Subiza, J., Jerez, M., Jiménez, J., Narganes, M., Cabrera, M., Varela, S., Subiza, E., 1995. Allergenic pollen pollinosis in Madrid. J. Allergy Clin. Immunol. 96, 15–23.

Tassan-Mazzocco, F., Felluga, A., Verardo, P., 2015. Prediction of wind-carried Gramineae and Urticaceae pollen occurrence in the Friuli Venezia Giulia region (Italy). Aerobiologia 31, 559–574.

Vogel, H., Pauling, A., Vogel, B., 2008. Numerical simulation of birch pollen dispersion with an operational weather forecast system. Int. J. Biometeorol 52, 805–814.

Please cite this article as: R. Navares, J. Aznarte, What are the most important variables for Poaceae airborne pollen forecasting?, Science of the Total Environment (2016), http://dx.doi.org/10.1016/j.scitotenv.2016.11.096

Chapter 6

Forecasting Plantago pollen: improving feature selection through random forests, clustering and Friedman tests

Type:	Published Article
Title:	*Forecasting Plantago pollen: improving feature selection through random forests, clustering and Friedman tests*
Journal:	Theoretical and Applied Climatology
Authors:	Ricardo Navares & José Luis Aznarte
Published:	August 2019
Impact Factor:	2.720
Quartile:	Q2
DOI	10.1007/s00704-019-02954-1

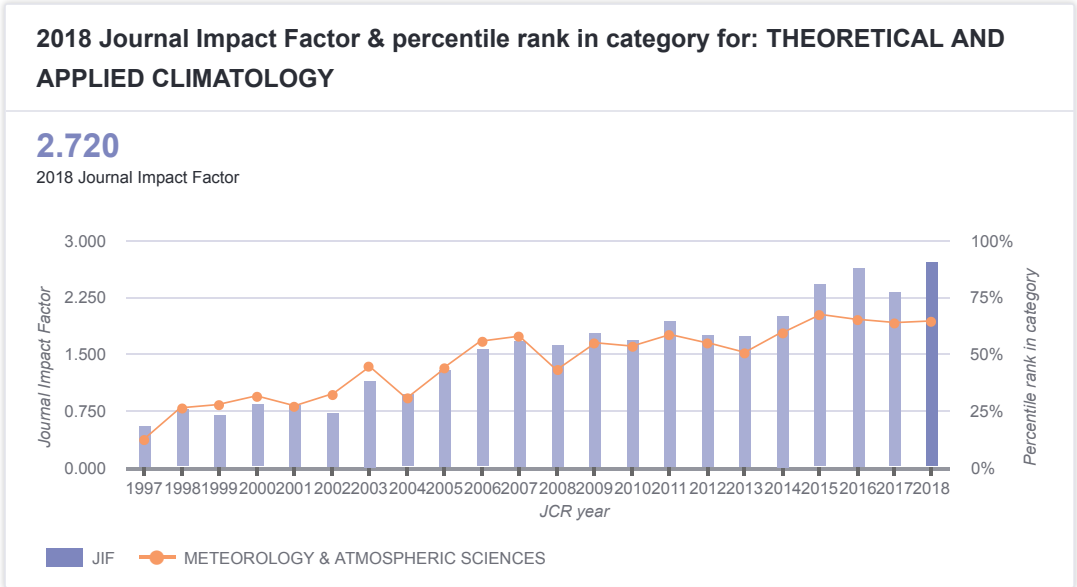


FIGURE 6.1: Impact factor Theoretical and Applied Climatology

Theoretical and Applied Climatology
https://doi.org/10.1007/s00704-019-02954-1

ORIGINAL PAPER

Forecasting *Plantago* pollen: improving feature selection through random forests, clustering, and Friedman tests

Ricardo Navares[1] · **José Luis Aznarte**[2]

Received: 12 August 2018 / Accepted: 14 July 2019

Abstract

Predicting concentrations of pollen is of great importance both for patients and for public health institutions. In this paper, we present a forecasting approach which relies on data and makes no assumptions on the underlying phenomena affecting the plants and the pollination process. Machine learning is used to build a model and to select the most important variables for prediction. Through nonparametric hypothesis testing, we show how some variables are indeed more important than others and how the careful combination of these variables can lead to more accurate and parsimonious models which avoid the huge computational times of more complex models while outperforming them in terms of the precision of the forecasts. By increasing the richness of the selected variables based on the clustered Friedman importance ranks, prediction error is reduced from 4.57 to 4.40 grains/m^3 as an average, which accounts for a 3.5% average improvement across locations studied with a 50% reduction of execution times.

1 Introduction

Allergy symptoms and their severity have been increasing in Western Europe during the last decades (de Weger et al. 2013), a fact which clearly implies the usefulness of the prediction of pollen concentrations, which are used not only to limit the exposure of patients to allergens but also to prearrange resources for clinical institutions. *Plantago* is one of the most common species among the herbaceous plants. Even though its airborne atmospheric concentrations are low, positive reactions appear nearly in 50% of the sensitive patients (Subiza et al. 1995).

Up to date, several studies have been proposed to approach the pollen forecasting problem based on techniques that range from classic multivariate regression to nonlinear models (Cotos-Yáñez et al. 2004; Rodríguez-Rajo et al. 2004; Tassan-Mazzocco et al. 2015; Tseng et al. 2018), to artificial neural networks (Astray et al. 2016; Aznarte et al. 2007; Iglesias-Otero et al. 2015). Due to the *curse of dimensionality*, in every approach, variable selection plays a crucial role. For example, complex multivariate models require a precise parameterization in order to avoid overfitting or, in the case of neural networks, the number of instances needed to train them increase with the topology of the network (which in turn makes overfitting likely).

Regardless of the chosen approach, the first aim of the practitioner is to select a set of variables which are relevant when forecasting airborne pollen concentrations. In general, it is assumed that weather conditions directly influence these concentrations. For example, heavy sudden rains during pollination wash away pollen grains from the atmosphere. Consequently, studies include meteorological variables such as previous daily precipitation (Castellano-Méndez et al. 2005; Iglesias-Otero et al. 2015) or wind speed, relative humidity, and solar radiation as a proxy of dryness, which in turn creates optimal conditions for pollen proliferation (Jones and Harrison 2004; Myszkowska 2014; Rodríguez-Rajo et al. 1983). Alternatively, other studies prefer the use of climatological indices to capture meteorological information prior the pollen release. Some such indices are cumulative weather variables (Andersen 1991; Matyasovszky et al. 2015; Myszkowska 2014; Otero et al. 2013; Pauling et al. 2014), which capture the influence of past and current weather conditions in current airborne

✉ Ricardo Navares
rnavares2@alumno.uned.es

José Luis Aznarte
jlaznarte@dia.uned.es

[1] Superior Technical School of Computer Engineering, UNED, Juan del Rosal, 16, 28040, Madrid, Spain

[2] Department of Artificial Intelligence, UNED, Juan del Rosal, 16, 28040, Madrid, Spain

Published online: 07 August 2019

concentrations (Deák et al. 2013). On the other hand, in the literature, there are also models which take a phenological point of view (Cannell and Smith 1983; Kmenta et al. 2016; Ribeiro et al. 2007; Smith and Emberlin 2006; García-Mozo et al. 2008), assuming that future airborne pollen release depends on the current and past growth state of plant buds. Finally others considered the problem as a univariate time series problem, developing predictors based exclusively on past pollen concentrations (Aznarte et al. 2007).

In any case, there is a shared necessity to reduce model complexity in order to ease its interpretability and shorten calculation times. Feature selection constitutes an important research field in Computational Intelligence and Statistics. Automatic feature selection is aimed at obtaining a subset of variables which are relevant for model construction, removing those which are redundant or irrelevant thus minimizing the loss of information.

As an alternative of automatic feature selection, some authors used statistical and probabilistic methods to reduce the dimensionality of the variables in time series forecasting. Among these techniques, factor analysis (Deák et al. 2013; Matyasovszky et al. 2015) and partial mutual information (Li et al. 2015; Tran et al. 2015) are widely used. The underlying idea is to find a new representation of the observed variables that is good at explaining the phenomenon under study at the same time avoiding collinearity among them.

The objective of this paper is to provide a framework to select the optimal combination of features to achieve a more precise forecast of airborne *Plantago* pollen concentrations while avoiding a priori assumptions about the influence of the variables, either meteorological or phenological, on future pollen releases. Through the use of random forests (RF), we perform an automatic feature selection which lets the model capture the inner information from the observed data and decide the importance of each one of the available features. Once a parsimonious model is built with the more relevant variables, it is used to forecast the series under study with good results compared with the literature. These results are in turn validated through a nonparametric ranking-based statistical test (Friedman 1937), showing the applicability of the proposal to the *Plantago* pollen concentration forecast problem in the region of Madrid.

2 Material and methods

2.1 Data description

Pollen observations correspond to daily *Plantago* concentrations registered at 7 locations distributed around the region of Madrid: Alcalá de Henares, Alcobendas, Aranjuez, Faculty of Pharmacy of Complutense University of Madrid, Getafe, Leganés and Villalba ,as shown in Fig. 1 . Pollen counts were conducted following the standard methodology of the Spanish Aerobiological Network (Galán Soldevilla et al. 2007) and were provided by Red Palinológica de la Comunidad de Madrid.

Meteorological observations were obtained from sensors placed in the close surroundings of the pollen stations (each pair of stations in the same municipality separated less than 2 km), consisting of daily average temperature in Celsius degrees, solar radiation in W/m^2, wind speed measured in m/s, daily rainfall in mm/h, pressure in hPa, and degree of humidity in percentage. The data was provided by the Autonomous Region of Madrid[1] and was measured following the criteria from the Spanish government meteorological agency (AEMET).

Both pollen and meteorological datasets span from year 2005 to 2009 at Alcalá de Henares, Alcobendas, Aranjuez, Getafe, and Leganés. Observations at the Faculty of Pharmacy are available from 2001 to 2013 while in Villalba, only 3 years are available starting 2007 to 2009.

Even though the climate in the region of Madrid is continental (with a strong Mediterranean influence), its geographical characteristics provide several peculiarities at each location. Situated at Guadarrama mountains, Villalba is 903 m above the sea level with a yearly average temperature of 10–11 °C and a yearly average precipitation of 1250–1500 mm. These characteristics are related to mountain climate. Conversely, Aranjuez, which is located 495 m above the sea level, has milder yearly average temperatures (above 14 °C) with a yearly average rainfall below 400 mm. The remaining locations are metropolitan areas between 594 and 668 m above the sea level with yearly average temperatures above 15.2 °C and precipitations around 440 mm.

2.2 Features

Previous studies (Navares and Aznarte 2016b) propose a set of features based on the influence that meteorological conditions have in the development of the plant, together with several variables inspired by phenological studies along with purely analytical approaches, which extract information directly from the time series. In order to perform a deeper analysis of the optimal selection of important features for forecasting airborne concentration, this study increases the number of features from which the models are constructed.

A total of 143 features were generated as shown in Table 1. In addition, an extra variable is included which

[1] https://gestiona.madrid.org/azul_internet/html/web/AvisosAccion.icm?ESTADO_MENU=1

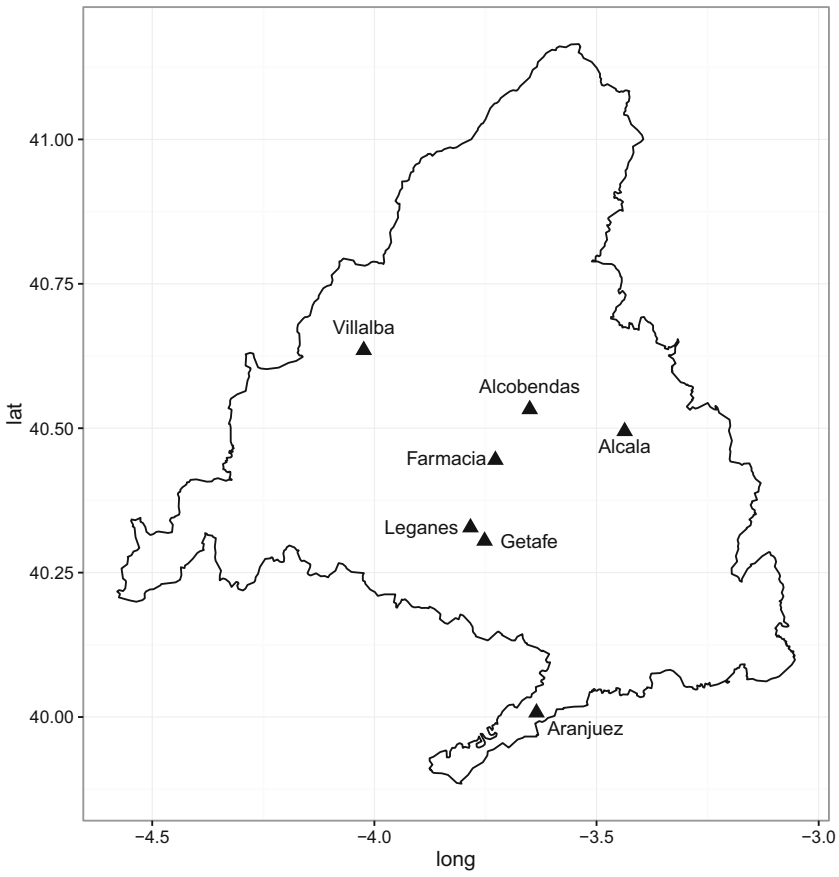


Fig. 1 Location of weather and pollen stations in the region of Madrid

Table 1 Summary of number of features generated by variable

	i	Σ_{10}	MA_i	Σ_{ytd}	Σ_{30}	Σ_{90}	Σ_{180}	std
Pollen	7	1	4	1	1	1	1	–
Temperature	7	1	4	1	1	1	1	1
T_{forc}	–	–	–	1	1	1	1	–
T_{chill}	–	- -	–	1	1	1	1	–
Humidity	7	1	4	1	1	1	1	1
Wind	7	1	4	1	1	1	1	1
Rain	7	1	4	1	1	1	1	1
Pressure	7	1	4	1	1	1	1	1
UV	7	1	4	1	1	1	1	1
Sun	7	1	4	1	1	1	1	1

Features include i-lagged variables, previous t days cumulative sums represented by Σ_t, moving averages (MA), and previous 15 days standard deviation (std)

i: previous $i \in [1, 7]$ day observation

Σ_{10}: previous 10-day cumulative sum

MA_i: max and min i-days moving average $i \in \{5, 15\}$

Σ_{ytd}: year to date cumulative sum

Σ_{30}: previous month cumulative sum

Σ_{90}: previous 90-day cumulative sum

Σ_{180}: previous 180-days cumulative sum

std: previous 15 days standard deviation

represents the Julian day of the correspondent year. Thus, the model is trained using a matrix of vectors of the form $(x_{1,t}, \ldots, x_{144,t}|y_{t+1})$, where t is the time in which the forecast is done, y_{t+1} represents next day pollen concentration (day-ahead forecast), and $x_{i,t}$ represents the value at time t of the variable i.

2.3 Random forests for regression

Proposed for the first time in Breiman (2001), a random forest (RF) is an ensemble approach which leverages the performance of many weak learners (trees) by combining them to form a strong learner. RF is a supervised learning procedure which generates several randomized regression trees over sample fractions of data, and combines them by averaging. Different random selections are computed by each tree improving stability and accuracy; this technique is known as bootstrap aggregating or *bagging* (Breiman 1996).

Bagging intervenes at two levels, first in data selection and second in variable selection, ensuring that each tree uses a different set of data and a different set of variables. Then, by averaging, the likelihood of *overfitting* diminishes providing a clear advantage over other computational intelligence methods such as neural networks. Also, this procedure gives robustness against the presence of outliers

which favors the selection of RF when compared with support vector regressions.

Several measures have been proposed to estimate variable importance in RF. The most advanced is the permutation accuracy importance measure. Its rationale is based on randomly permuting each predictor variable so the original association with the response is broken. Thus, when a variable is permuted, the accuracy is recorded and averaged over all trees providing the relative variable importance, which increases as the difference in accuracy becomes bigger before and after the permutation. As opposed to classification trees, the accuracy (or prediction error) in the regression problem is measured by the variance.

In our case, model performance is checked using the root mean squared error (RMSE) defined by Eq. 1, along with the coefficient of determination R^2:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{1}$$

where y_i is the observed i^{th} data point, $\hat{y}_i$ the predicted, and n the total number of data points in the test set.

2.4 Nonparametric hypothesis testing

In order to evaluate the variable relevance obtained by the RF algorithm, we use the nonparametric Friedman test (Friedman 1937) over variable importances with a post hoc procedure as described in Navares and Aznarte (2016b). The aim is to investigate, in significance terms, which variables are considered better predictors.

2.5 Experimental design

A first experiment (which will be considered as a benchmark for the subsequent) consists in building, for each location in which we have data, a model designed to predict one day-ahead pollen concentrations using the full set of 144 variables. As a result, we obtain the model's performance along with the identification of the most important variables. Secondly, a verification step is done through the application of statistical inference. The purpose is to generalize and corroborate if the average rankings of variables per location are statistically significant. This is done through the non-parametric tests applied to the importance of variables obtained per location and per year. Lastly, we compare the results obtained from the benchmark with different models created using subsets of the most important variables ranked in the hypothesis tests.

The benchmark model approaches the one day-ahead airborne concentration forecast problem by building a RF model per location. Given the number of years available, as a cross-validation technique, we use a *leave-one-out* (LOO) setup by years. Hence, the series are split into training and test set at each location, leaving one of the available years in each iteration as a test set and training the model with the remaining years. Through this validation technique, we aim to increase the generalization power of the test set, providing it with data points across all seasonal characteristics. Also, by averaging the outcome of the LOO, the results obtained are more independent of peculiar characteristics of each year.

The second experiment uses a different subdivision of the data to add an extra layer of generalization, which is location independent and it is not subordinated to the shape of the pollen curve of each year. For each combination of year and location, the variable importances resulting from building a RF model are distributed in a $N \times M$ matrix where N represents each pair {location, year} present in the dataset and M the available variables. The random nature of RF justifies the averaging of 50 instances of each model in order to obtain the expected variable importances as they might slightly change from one RF execution to other. A Friedman hypothesis test is then conducted to prove the existence of significant differences within the full set of variables and, if applicable, a post hoc analysis is performed.

Once the test is applied, we compare the models in terms of accuracy and time complexity with the benchmark from the first experiment, which uses the full set of variables. Thus, the results of this full-sized model are compared with those obtained by using a reduced set of the 5 and 15 most important variables out of the statistical inference process.

Considering the trade-off between accuracy and computational complexity, the last experiment is tailored to find the combinations of variables that improve current accuracies subject to minimizing the complexity of the model. As the number of combinations is large, and consequently the number of models, a restricted set of combinations is defined based on clustering the ranked importances obtained from the nonparametric Friedman test.

3 Results

3.1 Predicting with all the variables

In our first setup, RF is applied using the LOO technique across all locations with the full set of variables. The relative importance of the variables is calculated and then averaged, resulting in Fig. 2.

It can be seen that the variables Σ^{p}_{10}, p_1, and Σ^{h}_{180} stay ranked as the top 3 most important variables across the locations except for Villalba. At this location, 90 days cumulative pressure (Σ^{pr}_{90}) and wind speed (Σ^{w}_{90}) increase up to 6% and 4% of the total importance, respectively.

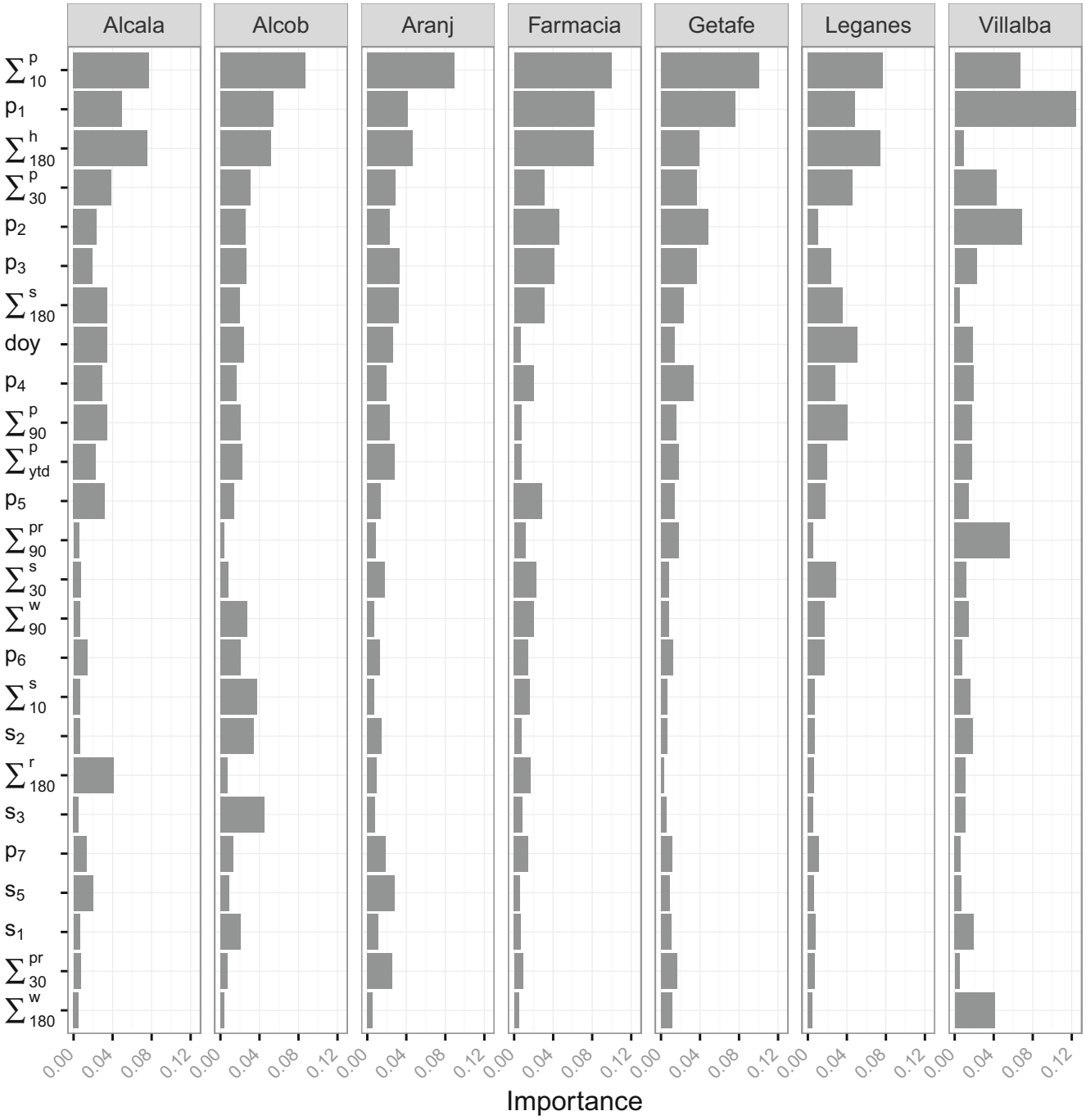


Fig. 2 Relative importance of the most important variables by location using the permutation accuracy importance measure (variables are ordered by the sum of their importance across locations)

Due to the fact that this station is located at 903 m above the sea level, its particular climate conditions, related to mountain climate, produce higher correlation between the wind variable and airborne concentrations. Also, lower atmospheric pressure due to the elevation eases plant formation and consequently pollen releases, which explains higher importance for Σ^{pr}_{180}, compared for example with Aranjuez which is located 495 m above the sea level. Furthermore, the different urban configuration, for example, with no high buildings, as compared with urban areas such as Getafe, Leganés, Farmacia, Alcobendas, and Alcalá, causes pollen concentrations to differ considerably.

However, excluding the particularity of Villalba, Σ^{p}_{10} stays as the most important variable across locations, followed by p_1 and Σ^{h}_{180} which alternate positions depending on the location.

3.2 Nonparametric tests to compare variable importance

As we have seen in Section 3.1, there is some stability in the top-ranked variables. However, there is no clear pattern on the grouping of the variables and their interrelations. In order to investigate this, and to provide statistical evidence of the existence of variables (or groups of variables), which are more important than others, a nonparametric test was performed.

For each location and year available, a RF model was trained using each combination {*location*, *year*} in isolation and the relative importance of each variable is computed. Out of this setup, the Friedman rank test was performed over 16 variables, which represent more than 1% of the total variance. This reduced set is considered as the pairwise computation with the full set is expensive. A Friedman statistic of $F = 165.84$, which is distributed according to chi-square with 15 degrees of freedom, obtains a p value of 1.04e−10 with $\alpha = 0.05$, which provides strong evidence of the existence of significant differences between the variables, which are ranked as shown in Table 2.

Figure 3a shows the dendrogram based on the rankings from Friedman. It can be clearly seen that there are two main groups which consist of the top four ranked variables by importance and the remaining ones. As Friedman's null hypothesis was rejected, a post hoc procedure can be carried out in order to check the differences between pairs of variables.

The rejected hypotheses with $\alpha = 0.05$ from the pairwise comparison is shown in Table 3. As stated above, there is no significant difference between the four most important variables Σ^p_{10}, p_1, Σ^h_{180}, and Σ^p_{30}, suggesting the first cluster of variables shown in the dendrogram from Fig. 3a. This means it is likely these variables alternate ranking positions depending on the location as seen in Fig. 2. It is noticeable that there is no evidence of the difference between the group variables $\{p_1, \Sigma^h_{180}, \Sigma^p_{30}\}$ and $\{p_2, p_3\}$ although there is significant evidence that Σ^p_{10} is different from p_2 and p_3 implying that $\{p_1, \Sigma^h_{180}, \Sigma^p_{30}\}$ has to be different from $\{p_2, p_3\}$, according to the logical relation between the combination of the hypotheses proposed by Shaffer (1986).

Regarding other groups of variables, there is a statistical evidence that p_2 is different from Σ^s_{30} and Σ^w_{90} which constitute the third cluster (Fig. 3a, top), and between p_3 and Σ^s_{180}, which also belongs to the third group. In summary, the tests prove the existence of features which are more relevant than others which are grouped according the three top levels of the dendrogram in Fig. 3a. There is no evidence that members of the same group differ from each other, which means that variables from the same cluster might alternate rank positions within the array of its constituent variables.

Table 2 Rank of the 16 most important features computed by averaging the rank of each variable importance for each year and location

i	Variable	Ranking	i	Variable	Ranking
1	Σ^p_{10}	3.219	9	p_5	9.073
2	p_1	5.073	10	Σ^p_{90}	10.024
3	Σ^p_{30}	5.591	11	p_6	10.097
4	Σ^h_{180}	5.975	12	doy	10.170
5	p_2	6.975	13	Σ^p_{ytd}	10.268
6	p_3	7.878	14	Σ^w_{90}	10.804
7	Σ^s_{180}	8.487	15	Σ^s_{30}	11.487
8	p_4	8.682	16	Σ^r_{180}	11.829

3.3 Predicting with a reduced set of variables

After verifying the statistical evidence of the relative importance of the variables, we repeated the experiment in Section 3.1 for two reduced set of variables selected from the findings in Section 3.2. The two sets consist of the 15 (setup B) and the 5 (setup C) most important variables from Friedman test, which are tested against the full set of 144 variables (setup A) from Section 3.1, in order to check how the setup influences model performance in terms of predictability and computational efficiency.

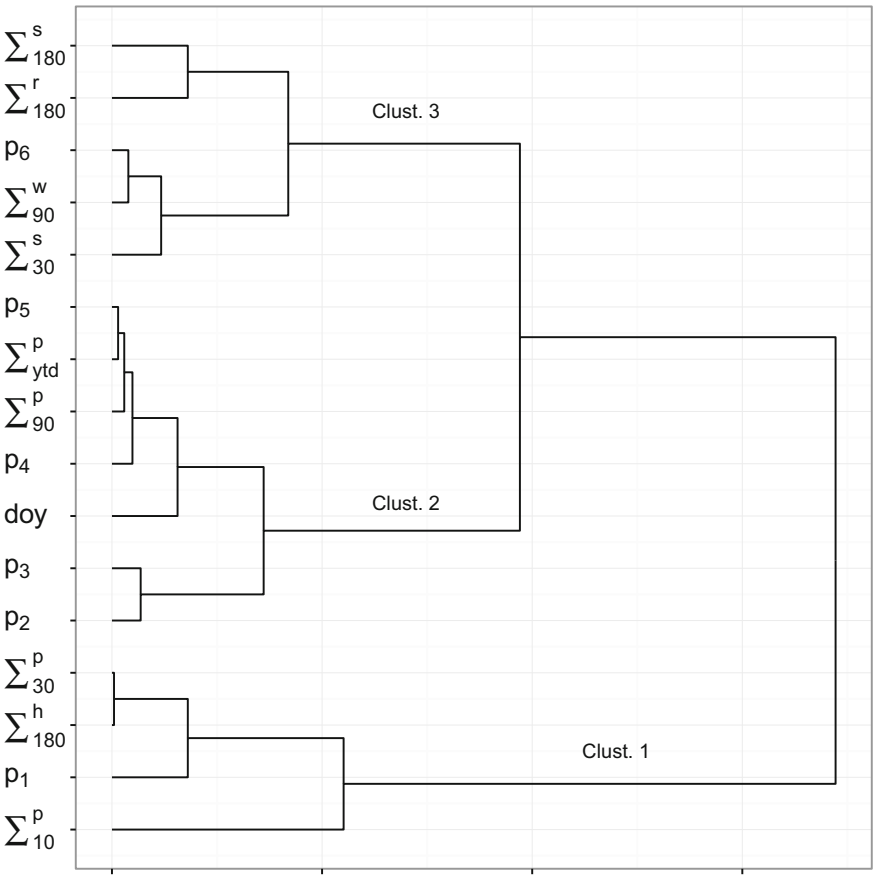
For each reduced set, 50 instances of a RF model were trained to forecast one day-ahead *Plantago* concentrations via the LOO cross-validation by years and observation station . Execution times for a single instance of each setup are 22.66, 17.08, and 11.86 s for setup A (144 variables), setup B (15 variables), and setup C (5 variables), respectively.

By using the setup B of 15 variables, execution time reduces by 25% per iteration while setup C (5 variables) produces a substantial reduction of almost 50%. However, the aim is to find a balanced setup which provides the shortest execution times without an excessive impact in the predictive performance of the models in terms of RMSE and R^2. Table 4 shows the performance metrics per location and setup; it can be seen that setup A produces in average the least accurate results with an RMSE equal to 4.80 compared with 4.64 for both setups B and C, which clearly indicates the benefits of reducing the number of variables both in terms of accuracy and computational performance. Having a close look at the locations, reducing the number of variables to 5 (C) produces mixed results when compared with setup B being the performance dependent on the location and, in any case, the error increases or decreases in a relatively small amount when both setups are compared.

It is clear that there is a consistent improvement on the errors across locations when a reduced set of variables

Fig. 3 **a** Dendrogram based on Friedman's ranks. **b** Combinations of 5 variables tested by clusters and groups used in Section 3.4. ${}^{n}C_{k}$ represents the combination of a subset of length k selected from a set of n variables being $k \leq n$

a



b

	Cluster 1	Cluster 2	Cluster 3
Group 1	${}^{4}C_{4}$	${}^{7}C_{1}$	
Group 2	${}^{4}C_{3}$	${}^{7}C_{2}$	
Group 3	${}^{4}C_{3}$	${}^{7}C_{1}$	${}^{4}C_{1}$
Group 4	${}^{4}C_{2}$	${}^{7}C_{2}$	${}^{4}C_{1}$

Table 3 Pairwise rejected hypothesis at $\alpha = 0.05$ with unadjusted p value and adjusted Holm and Shaffer p values

i	Hypothesis	p	p_{holm}	p_{shaff}
1	Σ^p_{10} vs Σ^r_{180}	2.66e-16	3.19e-14	3.19e-14
2	Σ^p_{10} vs Σ^s_{30}	3.74e-15	4.46e-13	3.93e-13
3	Σ^w_{90} vs Σ^p_{10}	5.44e-13	6.42e-11	5.72e-11
4	Σ^p_{10} vs Σ^p_{ytd}	2.04e-11	2.38e-09	2.14e-09
5	doy vs Σ^p_{10}	3.83e-11	4.44e-09	4.02e-09
6	Σ^p_{10} vs p_6	6.11e-11	7.02e-09	6.41e-09
7	Σ^p_{90} vs Σ^p_{10}	9.70e-11	1.11e-08	1.02e-08
8	p_1 vs Σ^r_{180}	1.32e-10	1.49e-08	1.38e-08
9	Σ^s_{30} vs p_1	1.06e-09	1.19e-07	1.11e-07
10	Σ^p_{30} vs Σ^r_{180}	2.27e-08	2.52e-06	2.38e-06
11	Σ^p_{10} vs p_5	2.59e-08	2.85e-06	2.72e-06
12	Σ^h_{180} vs Σ^r_{180}	2.59e-08	2.85e-06	2.72e-06
13	Σ^w_{90} vs p_1	5.01e-08	5.41e-06	5.26e-06
14	Σ^p_{30} vs Σ^s_{30}	1.40e-07	1.50e-05	1.47e-05
15	Σ^s_{30} vs Σ^h_{180}	1.59e-07	1.68e-05	1.67e-05
16	Σ^p_{10} vs p_4	2.04e-07	2.14e-05	2.14e-05
17	Σ^p_{10} vs Σ^s_{180}	5.44e-07	5.66e-05	5.00e-05
18	p_1 vs Σ^p_{ytd}	7.79e-07	8.02e-05	7.17e-05
19	doy vs p_1	1.25e-06	1.27e-04	1.15e-04
20	p_6 vs p_1	1.77e-06	1.79e-04	1.63e-04
21	Σ^p_{90} vs p_1	2.49e-06	2.49e-04	2.29e-04
22	p_2 vs Σ^r_{180}	3.91e-06	3.88e-04	3.60e-04
23	Σ^w_{90} vs Σ^p_{30}	3.91e-06	3.88e-04	3.60e-04
24	Σ^w_{90} vs Σ^h_{180}	4.38e-06	4.24e-04	4.03e-04
25	Σ^p_{10} vs p_3	9.41e-06	9.03e-04	8.66e-04
26	p_2 vs Σ^s_{30}	1.78e-05	0.00	0.00
27	Σ^p_{30} vs Σ^p_{ytd}	4.03e-05	0.00	0.00
28	Σ^h_{180} vs Σ^p_{ytd}	4.46e-05	0.00	0.00
29	doy vs Σ^p_{30}	6.00e-05	0.01	0.01
30	doy vs Σ^h_{180}	6.62e-05	0.01	0.01
31	Σ^p_{30} vs p_6	8.04e-05	0.01	0.01
32	p_6 vs Σ^h_{180}	8.85e-05	0.01	0.01
33	Σ^p_{90} vs Σ^p_{30}	1.07e-04	0.01	0.01
34	Σ^p_{90} vs Σ^h_{180}	1.18e-04	0.01	0.01
35	p_1 vs p_5	1.42e-04	0.01	0.01
36	Σ^r_{180} vs p_3	1.72e-04	0.01	0.01
37	p_2 vs Σ^w_{90}	2.71e-04	0.02	0.02
38	p_2 vs Σ^p_{10}	3.54e-04	0.03	0.03
39	p_4 vs p_1	5.97e-04	0.05	0.05

(setups B and C) are used, compared with the full set. This provides evidence about the convenience of ignoring redundant or noisy variables, which eases the predictive capability of the model. However, this gain is not shown in the R^2 measure, which obtains as an average of 60% in setup A compared with 60% in setup B and 55% in setup C. Given the similarity of results between choosing 5 or 15 variables,

Table 4 Average RMSE in grains/m^3 and R^2 of the forecasted test years studied at each location

Station	$RMSE_A$	$RMSE_B$	$RMSE_C$	R^2_A	R^2_B	R^2_C
Alcalá	5.75	5.69	5.80	0.58	0.54	0.51
Alcobendas	6.64	6.39	6.10	0.66	0.62	0.58
Aranjuez	3.83	3.76	3.81	0.54	0.51	0.49
Farmacia	3.71	3.63	3.63	0.67	0.65	0.59
Getafe	5.58	5.50	5.68	0.65	0.62	0.59
Leganés	5.48	5.15	5.11	0.59	0.64	0.52
Villalba	2.79	2.36	2.20	0.52	0.59	0.55
Average	4.82	4.64	4.64	0.60	0.60	0.55

there is still an open question on how to improve setup C, which is objectively more convenient in terms of execution, on those locations where it is less accurate than setup B.

On the other hand, through a close examination of the results, we have identified some difficulties of the model in capturing sudden peak pollen concentrations over 100 grains/m^3 as can be seen in Fig. 4. Concretely, this sudden extreme peaks increase RMSE to 10.06 grains/m^3 at the Faculty of Pharmacy (Farmacia) in 2013, while year 2009 produces a RMSE = 2.27 grains/m^3. The reduced amount of observed data with these characteristics explains the difficulties of the model to identify the inner behavior of the data when forecasting. However, this situation can be mitigated by limiting airborne concentration levels, imposing a threshold which substitutes the data points above it (Navares and Aznarte 2016a). This limit should be related to the concentration levels which are considered to be risky for human health; for instance, symptoms appear over 30 grains/m^3 in Finland and Croatia (Peternel et al. 2005; Rantio-Lehtimäki et al. 1991), while in Spain, the first symptoms are observed at 25 grains/m^3 (Rodríguez-Rajo et al. 1983).

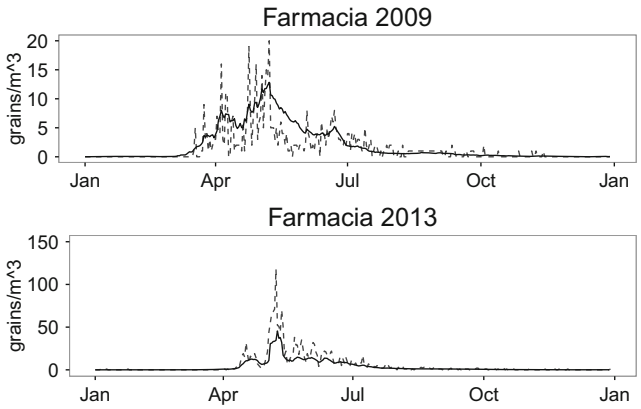


Fig. 4 Forecast (solid line) and observed (dashed) pollen concentration values for two years in the Faculty of Pharmacy (Farmacia)

3.4 Optimizing predictions with reduced sets of variables

In Section 3.3 we left an open question about how the setup C, consisting of 5 variables, can be improved, knowing it is the optimal configuration in terms of execution time and that it does not significantly differ in accuracy from setup B (15 variables). Concretely, the following question is relevant: is there a subset of 5 variables out of the 15 most important variables which outperform setup B accuracies?

Although random forests are known to be in general robust against overfitting, in the presence of correlated features, strong features can obtain low scores and the method can be indeed biased towards certain variables. For this reason, the experiment in Section 3.1 was repeated with a reduced set of 5 variables selected from the results obtained in the nonparametric test from Section 3.2. However, Fig. 3a showed that the 15 most important variables showed a cluster structure based on the results of Friedman test, which leads to the aforementioned question. The idea here is to use the cluster structure to find combinations of the 15 variables of setup B taken in groups of 5 which minimize the model's RMSE.

Of course, there are 3003 possible combinations of 15 variables taken in groups of 5, and this represents a high burden of executions of the experiment if one wants to explore them all. Thus, we propose to limit the number of combinations in such a way that the cluster structure, and the possible interactions between variables from different clusters, is taken into account. Taking as a reference setup C which includes the 5 most important variables, the idea is to step by step increase the presence of variables with lower importance by replacement in the original setup C. Hence, a first group includes the 4 most important variables from setup C while the fifth variable is replaced by one from cluster #2. This is done a number of times equal to the number of variables in cluster #2, resulting in the execution of 7 models as the second cluster contains 7 variables. The second group reduces the representation of the first cluster to 3 variables and increments the number of variables selected from cluster #2 to 2. The number of models executed is equal to the number of combinations resulting from taking 3 variables from cluster #1 and 2 variables from cluster #2 (always keeping 5 as the total number of variables). This logic is followed for subsequent groups selecting the correspondent number of combined variables per cluster as summarized in Fig. 3b.

As shown above, based on the pairwise comparison test, 3 clusters were obtained as in Fig. 3a. Cluster #1 contains the 4 most important variables, cluster #2 the next 7 variables in importance, and cluster #3 the last 4. Notice we excluded the least important variable from cluster #3 (Σ^s_{180}) for direct comparison with setup B of 15 variables. Given the restrictions mentioned, the groups of combinations explored are those shown in Fig. 3b.

In that figure, ${}^nC_k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$, being n the total number of variables belonging to a cluster and k the number of selected variables of the same cluster. The best results obtained are shown in Table 5 along with the reference benchmark, that is, represented by the best performing setup from Section 3.3 in terms of accuracy (RMSE).

The results obtained support the initial statement of this section about producing biased models when variables are selected regarding only importance, as this rules out the consideration that they might be related to each other. Increasing the richness of the variable selection space, in terms of including candidates from less important clusters according to Friedman test, and consequently with statistical evidence of differences, produces an increase of accuracy across all locations. Table 5 shows that including at least 2 variables from cluster #2 or cluster #3 or both increases accuracy by an average of 4%, with the largest increase of 7% in Getafe. Most of the improved results were obtained from groups 3 and 4, which include combinations of variables extracted from all the clusters considered, thus diversifying the distribution of importances and avoiding redundancies among the top-ranked variables.

4 Discussion

Automatic feature selection has been gaining relevance in computer sciences and statistics since the end of last century, when specific issues were published in specialized

Table 5 Comparison of the benchmark RMSE in grains/m^3 obtained from the best setup in Section 3.3 with the best performing combination along with its constituent variables and the test group they belong

Station	Benchmark	$RMSE_{best}$	Group #	Cluster #1	Cluster #2	Cluster #3
Alcalá	5.69 (B)	5.67	4	$\{p_1, \Sigma^h_{180}\}$	$\{p_2, p_5\}$	Σ^r_{180}
Alcobendas	6.10 (C)	5.73	4	$\{\Sigma^p_{10}, p_1\}$	$\{p_3, p_4\}$	p_6
Aranjuez	3.76 (B)	3.65	4	$\{\Sigma^p_{10}, p_1\}$	$\{p_2, p_4\}$	Σ^s_{30}
Farmacia	3.63 (C)	3.49	3	$\{\Sigma^p_{10}, p_1, \Sigma^h_{180}\}$	$\{p_5\}$	Σ^r_{180}
Getafe	5.50 (B)	5.11	4	$\{\Sigma^p_{10}, p_1\}$	$\{p_4, p_5\}$	Σ^r_{180}
Leganés	5.11 (C)	4.95	3	$\{\Sigma^p_{10}, p_1, \Sigma^h_{180}\}$	$\{p_4\}$	Σ^s_{30}
Villalba	2.20 (C)	2.20	4	$\{\Sigma^p_{10}, p_1\}$	$\{p_2, p_5\}$	p_6

journals, including many related papers (Blum and Langley 1997; Kohavi and John 1997). Numerous approaches and extended literature exist on the topic (Bolón-Canedo et al. 2013), and there is no widespread agreement on a so-called "best method" as the solution is always domain and problem specific. In this paper, we used the Friedman nonparametric hypothesis test to support the conclusions about our proposal selected as well as to provide statistical evidence of the results.

As seen in Section 3, using the proposed approach makes it possible to identify the most important variables for one day-ahead airborne pollen concentrations forecasting. Variable ranks obtained from the model (shown in Fig. 2) are consistent with the rank-based nonparametric hypothesis test, which provides statistical evidence of the findings without any a priori assumption about the influence of each climate or phenological feature.

The different geographical locations considered in the study, which do not share the same climatic conditions, prove the generalization abilites of the proposed approach, which obtain stable variable ranks across the observation stations. The dendrogram in Fig. 3a shows the expected rank interchangeability among variables which belong to the same cluster, being these clusters corroborated by a pairwise comparison post hoc procedure, which provides statistical evidence of the difference between pairs of variables.

Our approach identifies the previous day pollen observation and 10- and 30-day cumulative daily concentrations as the most important features in one day-ahead pollen forecast. In phenological studies, these features have been previously proved to serve as an indicator of the state of the plant (Ribeiro et al. 2007; Smith and Emberlin 2006). From the point of view of time series analysis, previous day pollen observation is generally found to be a relevant variable (Astray et al. 2016; Iglesias-Otero et al. 2015; Levetin 2014). Notwithstanding, there is weak evidence that 10-day cumulative sum has a tendency to be the most influential as it is the only variable in the top-ranked cluster which is clearly set aside from the variables of next group (composed by two- and three-day past pollen observations).

Along with these three variables, the model identifies previous 180-day relative humidity accumulation as one of the most influential features which is also in accordance with the findings in Jones and Harrison (2004), Levetin (2014), and Rodríguez-Rajo et al. (2004), as it promotes plant growth during the development state and mitigates pollen spread during release phase of the plant. Furthermore, relative humidity has more effect on airborne pollen concentrations compared with 180-day rainfall accumulation which is shown in cluster # 3 (Fig. 3a). This result is inline with previous studies which show non significant correlation between pollen concentrations and rainfall (Bartková-Scevková 2003) while humidity shows a significant effect. Pollen grains release allergen due to hydration which it becomes more intense before rainfalls (Grote et al. 2001).

Even though Barnes et al. (2001) show a higher influence of heavy rains on pollen concentrations when intradaily measured only daily pollen counts were available for this study, consequently, it was not possible to detemine intraday strong weather conditions changes which have effects on pollen concentrations (Barnes et al. 2001). For instance, sporadic short heavy rains followed by long dry periods within the day does not considerably change daily pollen counts (Bartková-Scevková 2003).

The sensitivity analysis presented by Puc (2012) indicates that the most influential variables to predict airborne pollen concentrations are maximum daily temperature and humidity, being minimum daily temperature ranked the least influential. Since only daily average temperatures were used in this proposal, there is a mitigation effect in terms of maximun daily temperature influence. However, cumulative solar radiation appears among the top most important features computed (Table 2). A positive radiative forcing involves climate warming and, as a result, an increase of temperatures (Leanh and Rind 1998).

In Section 3.3, we applied three configurations based on three sets of variables selected by importance from the rankings. The results from setups B and C outperform the initial model with 144 variables. This implies that removing redundant or irrelevant features tend to improve model performance both in accuracy and execution time. Given the parity of results of these setups and to reduce model complexity, in Section 3.4, we tested different combinations of 5 variables selected among the 15 most important (setup B). This mitigates redundancies among variables which do not significantly differ according to Friedman test and, as a result, improves both model accuracy and performance (measured in execution time).

The accuracy of the results are comparable with other studies such as Iglesias-Otero et al. (2015), who reports a R^2 of 0.66 compared to 0.60 in setup B and an average of 0.55 from the experiments in Section 3.4. We believe that optimizing the number of variables to select, instead the arbitrary 5 or 15, would improve this metric. However, Iglesias-Otero et al. (2015) used artificial neural networks, which require substantial data preprocessing as the inputs shall be limited when the number of training instances is not large enough, increasing the number of assumptions to be taken *a priori*. Not only does RF mitigate these drawbacks, but also is considered more robust against overfitting.

When compared to similar algorithms, our proposal outperforms the most accurate regression model among those proposed by Csépe et al. (2014) which achieves a RMSE of 28.26 using an implementation of regression trees. With a RMSE of 4.17, RF reduces the estimation variance

(and consequently the prediction error) due to the *bagging* technique. Also, averaging the results through the LOO approach provides more sound and general results compared with predefined test sets, whose results rely on the specific characteristics of the years selected.

The proposal presented in this paper has immediate practical application as it is able to identify the top-ranked variables producing satisfactory results. However, there are some drawbacks which deserve further investigation: (1) model error increases with the appearance of sudden high concentrations levels, concretely those over 100 grains/m^3, given the reduced number of observations of this type. However, this situation is persistent in most approaches found in the literature. This circumstance can be addressed by limiting the observations without altering the practicality of the model by setting meaningful thresholds (Navares and Aznarte 2016a, 2017). For instance, Rodríguez-Rajo et al. (1983) conclude that first allergy symptoms in Spain are reported over the risk threshold of 25 grains/m^3, which could be used as threshold. (2) There is an open question about the number of variables which guarantee optimal results.

In summary, we have presented statistical evidence that the approach proposed in this paper identifies the most important variables without any a priori assumptions over their influence. Removing redundant and irrelevant features improve model accuracy, being setup C with 5 variables the best performer among the subset tested and producing satisfactory results. Including a diverse subset of variables avoid biased results due to the mitigation of redundancies, thus improving the accuracy of the model.

5 Conclusions

The present study introduces a feature selection approach for forecasting airborne pollen concentrations. Statistical evidence of the consistency in identifying the most important features is provided through nonparametric hypothesis testing, using as a case study a one day-ahead forecast of *Plantago* airborne concentrations in the region of Madrid.

The results indicate that the proposed approach is a valid and efficient way to rank independent variables in the task of airborne pollen time series prediction. Precisely, we have determinated that the cumulative sums of recent pollen concentrations along with previous days pollen observations are among the most important features. This represents a data-driven confirmation of main findings from phenological studies. Also, the influence of cumulative daily relative humidity was shown as an important factor during plant formation and pollen release state, which coincides with the conclusions of previous researches based on the influence of meteorological variables.

Furthermore, from this ranking of variables, it has been shown how it is possible to build new, more parsimonious models which produce better results than benchmark approaches: by selecting the best set of independent variables, we achieved a 3.5% average improvement across locations with an average 50% reduction in execution times. This effect, due to the elimination of redundancies and irrelevant features, is to be expected in any regression technique using the same set of variables, and, as shown, is the main strength of our proposal when compared with other techniques.

Acknowledgments The authors would like to thank Patricia Cervigón (Comunidad de Madrid) and Montserrat Gutiérrez Bustillo (Universidad Complutense de Madrid) for his assistance in obtaining the data for this study.

References

Andersen TB (1991) A model to predict the beginning of the pollen season. Grana 30:269–275

Astray G, Fernández-González M, Rodríguez-Rajo F, López D, Mejuto J (2016) Airborne castanea pollen forecasting model for ecological and allergological implementation. Sci Total Environ 548–549:110–121

Aznarte JL, Benítez Sánchez JM, Lugilde DN, de Linares Fernández C, de la Guardia CD, Sánchez FA (2007) Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models. Expert Syst Appl 32(4):1218–1225

Barnes C, Pacheco F, Landuyt J, Hu F, Portnoy J (2001) The effect of temperature, relative humidity and rainfall on airborne ragweed pollen concentrations. Aerobiologia 17(1):61–68

Bartková-Scevková J (2003) The influence of temperature, relative humidity and rainfall on the occurrence of pollen allergens (betula, poaceae, ambrosia artemisiifolia) in the atmosphere of Bratislava (Slovakia). Int J Biometeorol 48(1):1–5

Blum A, Langley P (1997) Selection of relevant features and examples in machine learning. Artif Intell 97:245–271

Bolón-Canedo V, no NSM, Alonso-Betanzos A (2013) A review of feature selection methods on synthetic data. Knowl Inform Syst 34:483–519

Breiman L (1996) Bagging predictiors. Mach Learn 25:123–140

Breiman L (2001) Random forest. Machn Learn 45:5–32

Cannell M, Smith R (1983) Thermal time, chill days and prediction of budburst in Picea sitchensis. J Appl Ecol 20:269–275

Castellano-Méndez M, Aira MJ, Iglesias I, Jato V, González-Manteiga W (2005) Artificial neural networks as a useful tool to predict the risk level of Betula pollen in the air. Int J Biometeorology 49:310–316

Cotos-Yáñez T, Rodríguez-Rajo F, Jato M (2004) Short-term prediction of Betula airborne pollen concentration in Vigo (NW Spain) using logistic additive models and partially linear models. Int J Biometeorol 48:179–185

Csépe Z, Makra L, Voukantsis D, Matyasovszky I, Tusnády G, Karatzas K, Thibaudon M (2014) Predicting daily ragweed pollen concentrations using computational intelligence techniques over

two heavily polluted areas in Europe. Sci Total Environ 542–552:476–477
de Weger LA, Bergmann KC, Rantio-Lehtimaki A, Dahl A, Buters J, Déchamp C, Belmonte J, Thibaudon M, Cecchi L, Besancenot JP, Galán C, Waisel Y (2013) Impact of pollen. In: Sofiev M, Bergmann KC (eds) Allergenic pollen. Springer, Netherlands, pp 161–215, https://doi.org/10.1007/978-94-007-4881-1_6
Deák A, Makra L, Matyasovszky I, Csépe Z, Muladi B (2013) Climate sensitivity of allergenic taxa in Central Europe associated with new climate change related forces. Sci Total Environ 442:36–47
Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Statist Assoc 32:674–701
Galán Soldevilla C, Cariñanos González P, Alcázar Teno P, Domínguez Vílches E (2007) Manual de Calidad y Gestión de la Red Española de Aerobiología. Universidad de Córdoba
García-Mozo H, Chuine I, Aira M, Belmonte J, Bermejo D, de la Guardia CD, Elvira B, Gutiérrez M, Rodríguez-Rajo J, Ruiz L, Trigo M, Tormo R, Valencia R, Galán C (2008) Regional phenological models for forecasting the start and peak of the quercus pollen season in Spain. Agr Forest Meteorol 148:372–380
Grote M, Vrtala S, Niederberger V, Wiermann R, Valenta R, Reichelt R (2001) Release of allergen-bearing cytoplasm from hydrated pollen: a mechanism common to a variety of grass (poaceae) species revealed by electron microscopy. J Allergy Clin Immunol 108(1):109–115
Iglesias-Otero MA, Fernández-González M, Rodríguez-Caride D, Astray G, Mejuto JC, Rodríguez-Rajo FJ (2015) A model to forecast the risk periods of Plantago pollen allergy by using ANN methodology. Aerobiologia 31:201–211
Jones A, Harrison R (2004) The effects of meteorological factors on atmospheric bioaerosol concentrations: a review. Sci Total Environ 326:151–181
Kmenta M, Bastl K, Kramer M, Hewings S, Mwange J, Zetter R, Berger U (2016) The grass pollen season 2014 in Vienna: a pilot study combining phenology, aerobiology and symptom data. Sci Total Environ 566–567:1614–1620
Kohavi R, John G (1997) Wrappers for feature subset selection. Artif Intell 97:273–324
Leanh J, Rind D (1998) Climate forcing by changing solar radiation. J Climate 11(12):3069–3094
Levetin E (2014) Daily ragweed pollen forecasting. J Allergy Clin Immunol 133:AB17
Li X, Maier H, AC Z (2015) Improved PMI-based input variable selection approach for artificial neural network and other data driven environmental and water resource models. Environ Model Softw 65:15–29
Matyasovszky I, Makra L, Csépe Z, Sümeghy Z, Deák A, Pál-Molnár E, Tusnády G (2015) Plants remember past weather: a study for atmospheric pollen concentrations of Ambrosia, Poaceae and Populus. Theor Appl Climatol 122:181–193
Myszkowska D (2014) Predicting tree pollen season start dates using thermal conditions. Aerobiologia 30:307–321
Navares R, Aznarte J (2016a) Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features. Int J Biometeorol. https://doi.org/10.1007/s00484-016-1242-8
Navares R, Aznarte J (2016b) What are the most important variables for poaceae airborne pollen forecasting? Sci Total Environ 579:1161–1169
Navares R, Aznarte J (2017) Forecasting the start and end of pollen season in madrid. In: Advances in time series analysis and forecasting. Springer International Publishing, pp 387–399. chap 26
Otero J, García-Mozo H, Hervás C, Galán C (2013) Biometeorological and autoregressive indices for predicting olive pollen intensity. Int J Biometeorol 57:307–316
Pauling A, Gehrig R, Clot B (2014) Toward optimized temperature sum parametrizations for forecasting the start of the pollen season. Aerobiologia 30:45–57
Peternel R, Srnec L, Culig J, Hrga I, Hercog P (2005) Poaceae pollen in the atmosphere of Zagreb (Croatia), 2002–2005. Grana 45:130–136
Puc M (2012) Artificial neural network model of the relationship between betula pollen and meteorological factors in Szczecin (Poland). Int J Biometeorol 56(2):395–401
Rantio-Lehtimäki A, Koivikko A, Kupias R, Mäkinen Y, Pohjola A (1991) Significance of sampling height of airborne particles for aerobiological information. Allergy 46:68–76
Ribeiro H, Cunha M, Abreu I (2007) Definition of main pollen season using logistic model. Ann Agric Environ Med 14:259–264
Rodríguez-Rajo F, Frenguelli G, Jato M (1983) Effect of air temperature on forecasting the start of the Betula pollen season at two contrasting sites in the south of Europe (1995-2001). Int J of Biometeorology 47:117–125
Rodríguez-Rajo F, Dopazo A, Jato V (2004) Environmental factors affecting the start of pollen season and concentrations of airborne Alnus pollen in two localities of Galicia (NW Spain). Ann Agric Environ Med 11:35–44
Shaffer J (1986) Modified sequentially rejective multiple test procedures. J Am Stat Assoc 81:826–831
Smith M, Emberlin J (2006) A 30-day-ahead forecast model for grass pollen in north London, UK. Int J Biometeorology 50:233–242
Subiza J, Jerez M, Jiménez J, Narganes M, Cabrera M, Varela S, Subiza E (1995) Allergenic pollen pollinosis in madrid. J Allergy Clin Immunol 96:15–23
Tassan-Mazzocco F, Felluga A, Verardo P (2015) Prediction of wind-carried Gramineae and Urticaceae pollen occurrence in the Friuli Venezia Giulia region (Italy). Aerobiologia 31:559–574
Tran H, Muttil N, Perera B (2015) Selection of significant input variables for time series forecasting. Environ Model Softw 64:156–163
Tseng Y, Kawashima S, Kobayashi S, Takeuchi S (2018) Algorithm for forecasting the total amount of airborne birch pollen from meteorological conditions of previous years. Agr Forest Meteorol 249:35–43

Chapter 7

Forecasting hourly NO2 concentrations by ensembling neural networks and mesoscale models

Type:	Published Article
Title:	*Forecasting hourly NO2 concentrations by ensembling neural networks and mesoscale models*
Journal:	Neural Computing and Applications
Authors:	Damir Valput & Ricardo Navares & José Luis Aznarte
Published:	August 2019
Impact Factor:	4.664
Quartile:	Q1
DOI	10.1007/s00521-019-04442-z

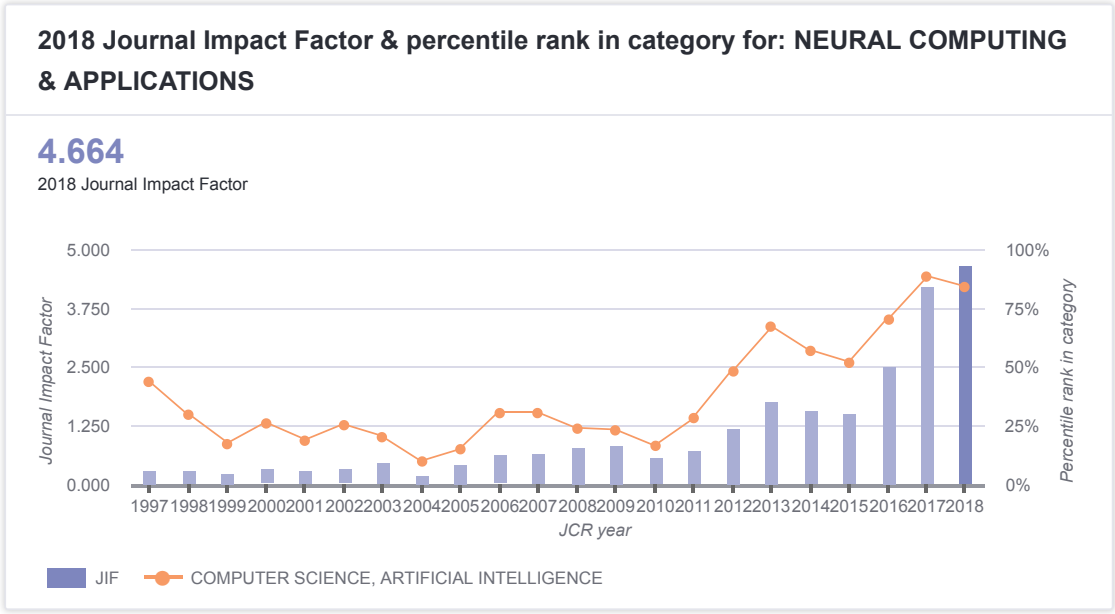


FIGURE 7.1: Impact factor Neural Computing and Applications

Neural Computing and Applications
https://doi.org/10.1007/s00521-019-04442-z

ORIGINAL ARTICLE

Forecasting hourly NO_2 concentrations by ensembling neural networks and mesoscale models

Damir Valput[1] · Ricardo Navares[1] · José L. Aznarte[1]

Received: 21 August 2018 / Accepted: 8 August 2019

Abstract
In the framework of extreme pollution concentrations being more and more frequent in many cities nowadays, air quality forecasting is crucial to protect public health through the anticipation of unpopular measures like traffic restrictions. In this work, we develop the core of a 48 h ahead forecasting system which is being deployed for the city of Madrid. To this end, we investigate the predictive power of a set of neural network models, including several families of deep networks, applied to the task of predicting nitrogen dioxide concentrations in an urban environment. Careful feature engineering on a set of related magnitudes as meteorology and traffic has proven useful, and we have coupled these neural models with mesoscale numerical pollution forecasts, which improve precision by up to 10%. The experiments show that some neural networks and ensembles consistently outperform the reference models, particularly improving the Naive model's results from around (20%) up to (57%) for longer forecasting horizons. However, results also reveal that deeper networks are not particularly better than shallow ones in this setting.

Keywords Neural networks · Deep learning · Air quality · Nitrogen dioxide · Forecasting · Madrid

1 Introduction

Air pollution is an increasingly worrying health problem in many urban regions of the world. Various studies have linked the exposure to high air pollution levels with a number of short-term and long-term dangerous effects [1]. One of the pollutants that is causing a greater deal of concern is nitrogen dioxide (NO_2). It is a pollutant linked mostly to traffic emissions and industrial activities [2–4]. In many cities, including Madrid, governments are implementing alert systems that enforce traffic restrictions when NO_2 levels are high [5]. In this context, NO_2 concentrations forecasting and early warnings about upcoming restrictions are of high importance, making it possible to timely alert the general public, which can in turn adapt its mobility plans accordingly.

Machine learning, and especially artificial neural networks (ANN), is increasingly popular alternatives to classic statistical methods for time-series forecasting. Additionally, with the availability of more powerful computer processing units, deep architectures of ANN are becoming more appealing [6]. The literature on this issue is vast, being the applications to the particular field of air quality forecasting less common, albeit the interest is growing at a fast pace. The previous approaches range from simple multi-layer perceptrons to stacking of complex networks, and most of them deal with the prediction of particulate matter or ozone, while NO_2 is less frequently considered [7–15]. Applications of deep neural networks to this particular problem are starting to bloom, but the literature is still scarce [16].

In order to predict air quality, meteorologists and physicists rely on numerical weather predictions coupled with a chemical layer which takes into account the emissions of pollutants and their interactions within the atmosphere. This approach uses complex models of the atmosphere, and it produces numerical pollution forecasts [17] in a regional or continental scale (thus the term "mesoscale" models, as opposed to global models). An

✉ José L. Aznarte
jlaznarte@dia.uned.es

[1] Artificial Intelligence Department, Universidad Nacional de Educación a Distancia — UNED, c/ Juan del Rosal, 16, Madrid, Spain

Published online: 20 August 2019

example of such models is the regional air quality predictions of the Copernicus Atmosphere Monitoring Service (CAMS) from the European Centre for Medium-Range Weather Forecasts (ECMWF) [18, 19]. These predictions are based on seven regional numerical air quality models, which consider meteorological parameters settings, boundary conditions for chemical species and emissions inventories. Given that they are designed to model relatively large parts of the atmosphere, the spatial resolution of these models is too low as to properly infer the local conditions in the streets of a given city, but it remains to be seen if this data source can be used as a predictor for local models (in the spirit of the downscaling models routinely applied in meteorology).

However, in order to downscale regional numerical pollution predictions into local forecasts, a clear picture of the local conditions is needed. In this case, it is known that NO_2 concentrations are closely related to several other time series which correlate with it. Locally recorded features including meteorology (temperature, humidity and similar), traffic (traffic intensity, speed, etc.) and the concentrations for other pollutants (CO, SO_2, NO) should then be incorporated to the forecasting system.

Our general objective is to design an operational forecasting system for the city of Madrid which predicts hourly values of NO_2 concentrations up to 48 h ahead. This system must provide efficient and sound forecasts able to raise early warnings regarding air quality. Through the experiments presented in this paper, we intend to investigate the predictive performance of several types of neural networks and ensembles, comparing them with the ECMWF regional model and other statistical learning models. We also intended to identify which predictors are the most relevant and, particularly, if the use of numerical pollution predictions might improve the performance of local models.

To the best of our knowledge, there are no previous thorough comparisons of neural network-based models to predict NO_2 concentrations. Furthermore, the usefulness of the numerical pollution predictions produced by the ECMWF as a predictor in this framework has not been studied. Finally, another minor contribution of our work is the modified version of a sequential backward feature selection algorithm which has proven useful to reduce the dimensionality while increasing the performance of the models.

The rest of the paper is organised as follows. After this introduction, we present the available data set in Sect. 2. In Sect. 3, we move on to describe how the features have been engineered, and how we ranked and selected the available features by their relative importance. After that, in Sect. 4, we introduce the chosen models and compare them with other forecasting approaches. The paper ends with conclusions on the achievements, the insights gained from the work done so far and the plans for the future development of the forecasting system.

2 Data description

The available data consist of the following sets:

2.1 Pollution data

The city of Madrid has an atmospheric pollution monitoring system which consists of 24 measuring stations around the city. The data collected through this monitoring system are made publicly available on the open data website of the Municipality of Madrid [20]. From there, we obtained the average hourly values of NO_2 concentrations (in $\mu g\,m^{-3}$) at the measuring station of Plaza de España. The selected data set consists of the data collected in the period from 1 January 2013 until 25 December 2015, both days included.

2.2 Meteorological data

The second relevant data set available consists of meteorological variables. Some of the 24 measuring stations in the aforementioned atmospheric pollution monitoring system record values of certain meteorological variables as well. For the selected station, we have the hourly values of average temperature, wind speed, accumulated precipitation and relative humidity. Related studies show that meteorological variables correlate with NO_2 concentrations and can therefore improve the accuracy of forecasting models [7, 8, 15, 21]. Thus, we decided to include them in our data set and study their relative importance.

2.3 Other pollutants

In addition to recording levels of NO_2 in the air, the measuring station at Plaza de España also monitors the levels of carbon monoxide (CO), sulphur dioxide (SO_2) and nitrogen monoxide (NO). Even though these pollutants are not the primary object of interest in this study, there is evidence that their concentrations correlate with the concentration of NO_2 [21, 22]. Hence, we decided that they could also be useful predictors for NO_2 concentration.

2.4 Traffic data

Traffic data are also available in the open data website of the Municipality of Madrid [20]. The city has a network of 3613 measuring points around the city which register a number of traffic variables: intensity, road occupation and average velocity of the vehicles, among other. The data are

recorded and integrated over periods of 15 min. We aggregated it into hourly time stamps to adjust it with the rest of the available series. Among the different recorded variables, we selected traffic intensity, which is expressed as the number of vehicles over the observed time period. It would be interesting to include the average velocity as well; however, this data are not available for the observed location. We deem this data to be potentially useful for the prediction of NO_2 concentrations because of the fact that traffic emissions highly influence the NO_2 levels [15, 21].

Since the series in this data set are quite scarce and present many missing data points, and since the traffic monitoring network is very dense (having ten times more measuring points than the pollution monitoring network), we took into account six monitoring stations in the proximity of the air quality measuring station (the busy streets surrounding it). The measurements from those six stations have been united into a new traffic indicator by averaging the data points from the observed stations. This way we obtained a new time series which serves as a representation of the traffic situation around the NO_2 measuring station.

2.5 Numerical pollution predictions

As mentioned above, one of our goals is to explore the usefulness of regional numerical pollution predictions in designing local forecasting systems. We considered forecasts from the ensemble model of the Copernicus Atmosphere Monitoring Service, which are produced up to 96 h ahead by the European Centre for Medium-Range Weather Forecasts (ECMWF) [18, 19].

We used these forecasts for the period 1 January 2013–25 November 2015. These forecasts are produced in a continental scale and are based on seven state-of-the-art numerical air quality models developed in seven European countries. The ensemble model takes forecasts of the seven air quality models and combines them by calculating their median value. This results, on average, in better air pollutant concentration forecasts than each individual model.

The time resolution of the ensemble model is 1 h, and the spatial resolution is 0.1°. New forecasts are published every 24 h, and the data are forecast for the surface level and up to 5000 m above it, although we used only the surface forecasts.

2.6 Missing data

The quality of the available data is deteriorated by missing data points and invalid measurements. For the latter, in the data set of the historical NO_2 values, the data provider specifically flagged some data points as invalid. In addition, certain data samples are missing from the hourly time series (which sometimes turn out to be even whole days). The reason for missing and invalid data samples is usually faults or malfunctions in the sensors or in the telecommunications system.

We decided to exclude from the final data set the days when two data points are not valid. If there were up to two missing/invalid data points in a day, we performed linear interpolation to recover them. The reason we chose this approach is the shape of NO_2 hourly concentrations signal: it changes fairly slowly throughout the day, and, observed locally, it can be approximated well with linear interpolation.

The final data set comprises of the data points from 902 days, out of a total of 1059 day in the observed period. That means that around 15% of data is missing due to invalid or unrecorded measurements.

2.7 Exploratory data analysis

As stated above, the series under study consist of hourly data samples. Let us first introduce the notation we are going to use throughout this text. We say that for some time series y_t of observed values at the time t, $\hat{y}_{(t+h)}$ is the forecast of the value of y at the time $t + h$. We call the natural number h the forecasting horizon, measured in hours in our case.

In Fig. 1, we can observe the characteristics of hourly, daily and monthly variations in concentrations of our NO_2 measures. These observations are helpful in creating the initial set of input features. Looking at hourly averages, we can observe two peaks in NO_2 concentrations that happen on an average day: one around 9 a.m. and other around 9 p.m. Intuitively, these peaks correspond to the traffic rush hours that happen around that time. The lowest NO_2 concentrations are recorded in the night time from 1 a.m. until 5 a.m. when the traffic intensity is at its lowest.

Looking at the distribution of hourly NO_2 concentrations shown by months, we note that winter months have, on average, slightly higher hourly concentrations of NO_2 than summer months. This is in accordance with some previous studies [10] that have shown that warm, dry weather decreases the concentration of NO_2. In the case of Madrid, a key issue is the stability of the atmosphere when winter anticyclones are over the Spanish peninsula.

Observing the hourly NO_2 levels by days of the week, we can see that they decrease during weekends, and that in general, Fridays have the highest average NO_2 concentrations. Again, it is possible to draw a parallel between these observations and the traffic activities around our air quality measuring station.

Upon this analysis, we decided to introduce additional dummy variables of hour of the day ("hod"), day of the week ("dow"), day of the year ("doy") and type of day

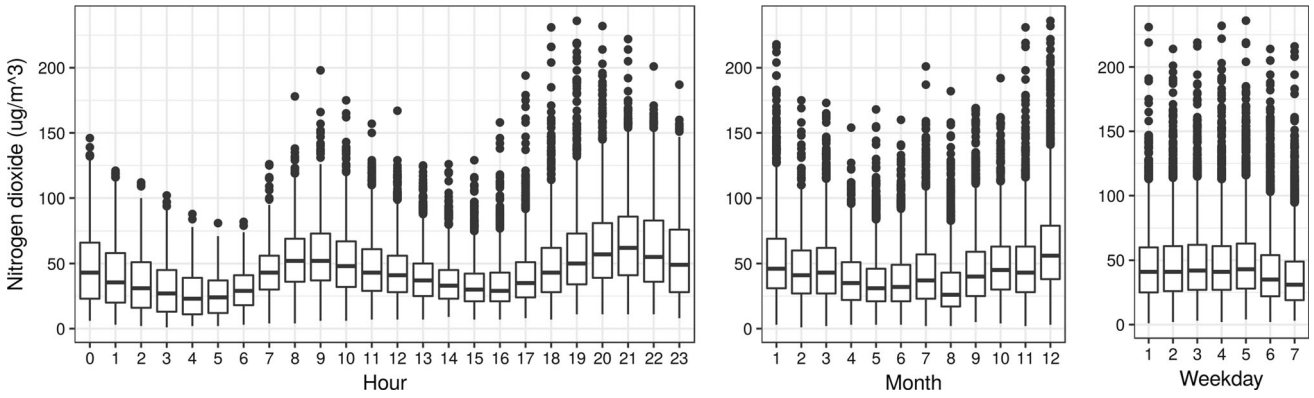


Fig. 1 Distribution of NO_2 concentrations considering the hour, the month and the day of the week (week starts on Monday)

("tod"), a categorical variable distinguishing between workdays, Saturdays and Sundays or public holidays. These variables encode time-related parameters of the observed series, and they are explicitly fed as additional features into the models in order to help them learn time dependencies.

Next, we studied the correlation dependency between the time series of NO_2 concentrations and meteorological variables, other observed pollutants, CAMS ensemble forecasts and traffic intensity. We calculated the Pearson correlation coefficient to examine the linear correlation between the hourly NO_2 levels and Spearman correlation coefficient to measure the monotonic relationship between two time series.

From Table 1, we can observe that NO_2 levels correlate the strongest with wind speed (negative correlation), other pollutants, CAMS forecasts and traffic intensity (positive correlation, as one might expect). It is weakly and negatively correlated with the hourly average temperature, with rainfall and relative humidity being the weakest correlated variables with NO_2 levels.

Below, we show how a feature ranking process is capable of giving us a more detailed insight on the importance of each feature.

2.8 Evaluation criteria

In order to asses the accuracy of our forecasting models, we used the root-mean-squared error (RMSE) and mean bias as defined in [15]. These measures have been estimated using a tenfold cross-validation method.

3 Feature engineering

This section describes how the features were extracted from the data set, and it presents the analysis of the feature importance and the algorithm for the elimination of features that reduces the forecasting error. All models are trained for six forecasting horizons: $h \in \{1, 2, 6, 12, 24, 48\}$.

3.1 Feature extraction

While constructing the initial set of input features, we relied on the observations from the exploratory analysis of the data, on findings in related works and on knowledge of urban emissions of NO_2 [1, 15, 21, 22].

For each available time series in our data set, the relevant past samples for each forecasting horizon h were selected. For instance, if we were to build a forecasting model that will be run at the time t with $h = 12$, one of the relevant NO_2 concentrations for this task is the one at $t - 12$. This follows from observing that the hourly NO_2 concentrations series can be seen as a quasi-periodical signal with a period of 24 h, as well as from looking into the correlation between the NO_2 levels at the times t and $t - 12$.

When it comes to forecast from the CAMS ensemble model, considering that their data points belong to a spatial grid with a resolution of 0.1°, we took nine points that form the neighbourhood of our observing site. We tried experimenting with bigger and smaller neighbourhoods,

Table 1 Correlation between NO_2 (with no lag) and other variables

Coefficient	Temperature	Wind	Rainfall	Humidity	CO	NO	SO_2	CAMS	Traffic intensity
Pearson	− 0.107	− 0.346	− 0.009	− 0.022	0.708	0.634	0.352	0.576	0.287
Spearman	− 0.115	− 0.421	− 0.010	0.010	0.787	0.753	0.567	0.575	0.316

including solely one point on the grid (the one closest to the coordinates of the air quality station), but the model with the neighbourhood produced the smallest errors. The ECMWF publishes new forecasts every 24 h ahead, and we included only the freshest forecasts as features.

All features have been re-scaled into the range $[0, 1]$ to allow for faster convergence.

3.2 Feature ranking and selection

In addition to forecasting the levels of NO_2 concentrations, we are interested in gaining insight into the importance of the selected features. In order to rank them according to their relative importance, we used a random forest of 100 decision trees. Decision trees are a natural model for this task since they inherently rank the features by their importance every time a split is performed according to the function that measures the quality of splits [23, 24]. As the splitting criterion, we took the mean squared error. The feature rankings can then be easily obtained by weighing the improvements in the criterion in all the nodes where the feature appears as a splitter. At the end of this process, each feature ends up with a score determining its relative rank.

Furthermore, we used the ranked features to determine whether there exists a subset of initially engineered features that yields better accuracy (lower generalisation error). For this purpose, we implemented a slightly modified version of a sequential backward feature selection algorithm [24]. The algorithm relies on a so-called wrapper method in which different subsets of features are tried out on a machine learning model (neural networks in our case), and the accuracy is measured, while the algorithm searches for the combination of features that gives the lowest generalisation error. The search is performed by finding the threshold score such that all the eliminated features (the ones with a lower score than the threshold) give a higher RMSE on the validation data set when being used to train the neural network. In our experiments, it turned out that eliminating all the rain variables yields better generalisation error.

Known benefits of feature selection are reducing computation time, improving prediction performance by removing irrelevant and redundant features, improving the generalisation capability of the model and facilitating understanding of the data [24]. In comparison with principal component analysis (PCA) [25], which reduces the initial dimensionality of the set of features by exploiting their linear dependencies and creating a new set of uncorrelated features, feature selection simply filters out redundant features from the original set. The benefits of a wrapper method of feature selection over PCA are: greater transparency by keeping original features intact, further insights about the importance of each of the engineered features, taking into account the target value when estimating predictive power of features and not relying only on a linear relationship between the features. Moreover, our hand-engineered set of features does not suffer from high dimensionality problem, and thus, in our application, we benefit more from feature selection and ranking than PCA method.

However, one of the disadvantages of ranking features with random forests is that, when it comes to correlated features, after choosing one of them, the importance of the other ones is significantly reduced. Furthermore, the algorithm of backward feature selection is computationally very expensive, and for that reason, we applied it only using one model of shallow FNN. The selected features were then used for all other models for a fair comparison between them.

In Table 2, we present the 12 most important features as ranked by this method, for three forecasting horizons: $h \in \{1, 12, 48\}$. The algorithm was run several times, and it proved to give consistent results throughout the runs. We note how the importance of predictors changes when we compare shorter-term (1 h) to longer-term (48 h) forecast. For $h = 1$, the current value of NO_2, as well as other pollutants (CO, NO), seems to play an important role, whereas when $h = 48$ the different neighbouring CAMS points and the features encoding the time parameters (such as hour of the day, type of the day and day of the year) seem to gain on importance. Intuitively, as we predict further into the future, the current conditions are less important and the general common characteristics of the

Table 2 Ranking of the most important features according to their relative importance, for horizons 1, 12 and 48

Rank	1	12	48
Forecasting horizon			
1	no2(0)	cams7	hod
2	no2(1)	cams4	no2(0)
3	co(0)	cams3	cams4
4	cams1	no2(12)	cams7
5	cams7	cams1	tod
6	no(0)	cams6	cams3
7	cams4	hod	cams6
8	cams3	no2(13)	no2(96)
9	hod	cams9	no2(1)
10	cams6	tod	cams1
11	cams9	traffic(13)	no2(24)
12	no2(2)	traffic(12)	doy

The number in parenthesis is the number of lag hours from the runtime of the forecast

forecast moment convey more information useful for the prediction.

4 Results and discussion

As discussed in [26], and mentioned above, the training examples were shuffled following a uniform distribution and divided into tenfold in order to apply cross-validation. Considering the stochastic nature of some models, we ran the training algorithm five times for each fold and averaged the results from the runs.

4.1 Reference models

In order to compare the results of the neural networks, we selected other benchmarking models:

- *Naive predictor* Also known as random walk is a family of predictors that simply take a past value and assign it to a forecast. They are a good ground estimate for benchmarking the performance of other algorithms. In our case, the Naive predictor forecasts the NO_2 level at the time $t + h$ by taking the value at the current time t: $\hat{y}_{t+h} = y_t$, for y being the NO_2 concentration.
- *CAMS* This model takes the predictions from the CAMS ensemble model, selects the eight-neighbourhood of the observed site and produces a forecast as the average of the eight CAMS neighbours. It is important to note that it is not possible to estimate the k-hours ahead forecast for every $k \in [1, 48]$ due to the fact that the new forecasts in CAMS are not published on hourly basis, but every 24 h. The consequence is that our estimate is equal for every $k \in \{1, 2, 6, 12, 24\}$, and it differs only for $k = 48$ where we used 2 day old forecasts from the CAMS ensemble model.
- *Linear regression (LinReg)* with l_2 penalty function has been used to compare linear with nonlinear forecasting methods [27].

4.2 Shallow feed-forward artificial neural networks

The shallow ANN we are using is a fully connected multilayer perceptron that has an input layer, an output layer and one hidden layer. In the process of tuning the ANN, the oldest sample has been chosen experimentally, from the set of $\{24, 48, 72, 96, 120\}$ h before the time of forecast. The data samples older than three days did not improve the forecasting accuracy of the ANN.

The optimisation algorithm we used for training the network is adaptive moment estimation (ADAM) [28], with a learning rate experimentally set to 10^{-4} and mean squared error as loss function. Over time, ADAM has become a widespread optimiser in ML application as it has been empirically shown to outperform other available optimisers when training neural nets, as it combines two great ideas of deep learning: RMSprop and momentum [29]. We have chosen it here due to its demonstrated efficiency. Furthermore, early stopping method and $l2$-regularisation, which adds a squared magnitude of coefficients as a penalty term to loss function, were also applied. Since the $l2$-regularisation is very sensitive to the used regularisation parameter, we tuned the model with respect to that hyper-parameter as well.

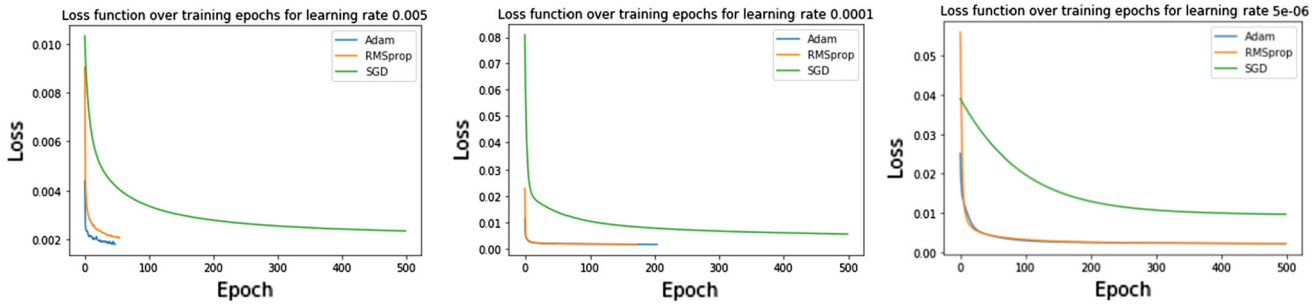
Early stopping was implemented by monitoring the value of the loss function on the validation data set and stopping the training if the monitored value did not change for at least 1^{-6} after 20 iterations in one run of the optimisation algorithm.

In Fig. 2, we show the performance of three common optimisers: ADAM, RMSprop [29] and stochastic gradient descent (SGD) [29], on the training data set, across a range of learning rates: $\{0.005, 0.0001, 0.000005\}$. As early stopping has been used, training has either been performed for 500 epochs (maximum number of epochs set in this example) or interrupted according to the stopping criterion. As can be seen, RMSprop and ADAM have similar performances, with ADAM performing visibly better at lower values of learning rate, while SGD exhibits an inferior performance. We have chosen the learning rate of 10^{-4} as it gave us the most satisfying ratio between accuracy and speed of convergence: higher values do not achieve the desired accuracy, while further lowering of the learning rate slows down the training without tangibly improving the performance.

In Fig. 3, we show the convergence of the loss function on both training and validation data sets, for onefold and the chosen learning rate 10^{-4}. SGD (Fig. 3c) has the smoothest cost function, but the longest computation time, and it achieves far inferior error rate. ADAM (Fig. 3a) takes a bit longer to train the network than RMSprop (Fig. 3b), but provides a smoother loss function on the test data. The difference in their error rates across all the experiments we performed is almost negligible. After performing this analysis, we decided to continue our experiments using ADAM optimiser.

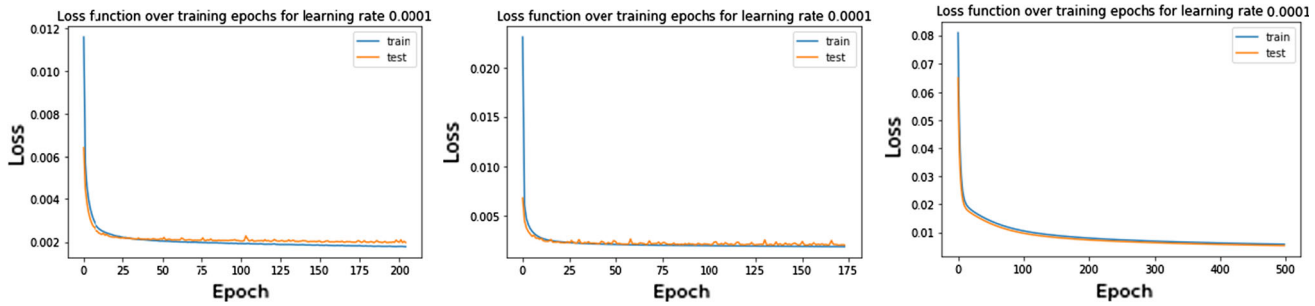
To speed up the convergence of the optimisation algorithm, mini-batch training was used with the size of the batch set to 100. The number of training epochs was limited to 500, but in most cases, the network training came to halt before that number was reached due to early stopping.

The activation function chosen in the hidden layer was the rectifier linear unit (ReLu), and in the output layer, the



(a) Learning rate = 0.005 **(b)** Learning rate = 0.0001 **(c)** Learning rate = 5e-6

Fig. 2 Comparison of optimisers: ADAM, RMSprop and SGD, across varying learning rates, training data



(a) ADAM **(b)** RMSprop **(c)** SGD

Fig. 3 Loss function on training and validation data set, for ADAM, RMSprop and SGD and learning rate = 0.0001

hyperbolic tangent function. Substituting ReLu with a hyperbolic tangent function in the last layer has given as a much smoother convergence of cost function and an overall better performance.

Finally, the number of neurons in the hidden layer has been chosen experimentally so that the neural network does not over-fit. In every case, we tried a range of values, capped at the number of input dimensions and settled at the one that gave the lowest error (RMSE), which was achieved with 85 neurons in the hidden layer. The ANN was implemented using the Keras library for Python [30].

4.3 Deep feed-forward neural networks

With the intention to check if they improved on the results obtained by the shallow feed-forward network, we decided to test out deeper configurations. We experimented with gradually increasing the depth of the network. The deep networks inherited the characteristics of the aforementioned shallow network: early stopping, l_2-regularisation, adaptive moment estimation with the same learning rate and mini-batch training. All hidden layers use rectifier linear units, and the output layer uses the hyperbolic tangent activation function.

Experiments with deeper networks have shown that the gradual increase in depth can provide us a better forecasting accuracy . We succeeded in developing deeper models that reduced the generalisation error while keeping the number of neurons in each hidden layer relatively small. (The results shown were achieved with 45–60 neurons per hidden layer.) The improvement over the shallow neural network's error is not overwhelming, but it is noticeable. Furthermore, the shallow neural network performs comparably to the deep ones for smaller values of forecasting horizons, but the benefit of the deeper ANN is more significant at higher forecasting horizons (concretely, from 1.7% smaller RMSE for $h = 12$ up to 2.1% for $h = 48$). The reason for this is that, given the persistence of the series, the 1-h ahead forecast of NO_2 concentrations is not a very difficult problem and all models, including the Naive model, perform comparably.

Deeper architectures, using more than three layers, did not improve the generalisation error. We experimented with architectures of up to 20 hidden layers. In the process, we experimented with various network topologies and observed that, in terms of the number of neurons in each hidden layer, the hourglass topology (decreasing and then again increasing number of neurons per layers) performs slightly better than other ones we tried (although all were giving comparable results).

This can be seen in Fig. 4 where we show the RMSE on the validation test for 1, 2, 3, 5, 10, 15 and 20 hidden layers used. The error (RMSE) on the validation data set even increases slightly when more than three hidden layers are

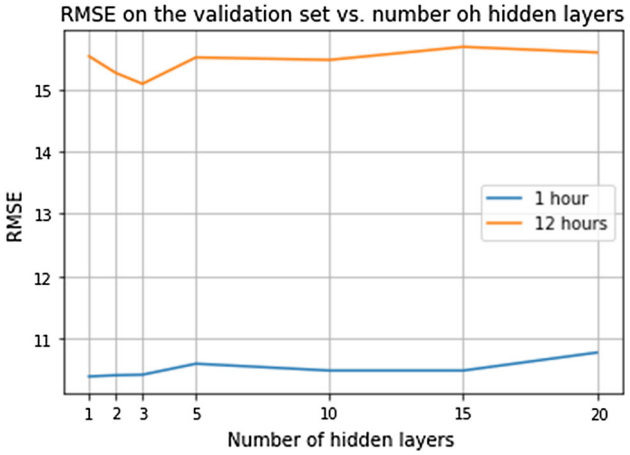


Fig. 4 RMSE on the validation data set versus number of hidden layers used in FNN

used. Figure 4 shows the error for forecasting horizons of 1 and 12 h only since the behaviour similar across all forecasting horizons.

4.4 Long short-term memory neural networks

We also tested a recurrent neural network-based model known as long short-term memory (LSTM) [31]. An LSTM unit possesses self-loops which enable the flow of the gradient for long durations, enabling it to deal with the vanishing gradient problem. Together with an input gate, an output gate and a forget gate, this architecture models the short-term memory that allows the network to learn over many time steps. Thanks to that, they are suitable for forecasting time series.

Since the sequences that we are using for training have different lengths (for example, the NO_2 sequence is longer than the other ones), we used zero-padding on shorter sequences. The time delay was determined by the length of the NO_2 sequence. During the building phase, one or more feed-forward layers were added after the LSTM layer. However, the best accuracy was obtained with only one layer after the LSTM. The results obtained compare favourably to the ones obtained with deep feed-forward networks.

Furthermore, we experimented with deeper architectures of LSTM and a bidirectional LSTM. The deeper architecture of LSTM consisted of three layers, with the number of neurons in each layer being 60–20–10. A bidirectional LSTM was applied so that the input can be run in two directions: one from past samples towards the future ones and the other one that goes in the opposite direction, from the future samples to the past ones. Running the input in two ways gives us the ability to have two hidden states, in no way communicating with each other, combining the information from the past and the future in any point in time.

4.5 Convolutional neural network

Convolutional neural networks (CNN) [32] differ from feed-forwards neural networks mainly by the existence of convolutional layers, which are hidden layers that utilise the power of mathematical convolution to transform inputs. Convolution allows for the encoding of the local properties of the input in such a way that propagates the information in a more efficient manner.

In the case of high-dimensional inputs, it can be impractical to fully connect all hidden layers and in such cases, CNN can be used to connect to reduce the size of the inputs. Reducing the size of the input is done by applying filters of reasonable size to perform convolution in convolutional layers.

In the selected one-dimensional CNN, a filter is composed of a subset of input features. The convolutional layers are then down-sampled by the pooling layers, which further alleviates the computational burden. Finally, the pooling layer is fully connected to ten ReLU neurons which are then passed to the output layer. The proposed one-dimensional CNN uses 32 filters of size 5 which are connected to a max pool layer of size 2.

4.6 Ensemble models

Lastly, we turned to the meta-learning approach of ensembling single learners' forecasts to improve the accuracy of our forecasting system [33]. We fed the forecasts obtained from all single learners but Naive and CAMS models into a meta-learner. From our experiments, the averaging model and weighted average model ('MetaAvg' and 'MetaWAvg', respectively) employed as meta-learners clearly improve the generalisation error. We tried also ensembling through decision trees, linear regression and FNN as meta-learners, but we did not find an architecture that would outperform the best single model out of the developed ones. The weights of the weighted average were calculated using RMSE of the single learners in such a way they are inversely proportional to the RMSE of a learner.

4.7 Discussion of the experimental results

Figure 5 and Table 3 show the root-mean-squared error and mean bias obtained with all the developed models for several forecasting horizons. Naive and CAMS models, shown in the table, were excluded from the figure so to place the focus on the advanced models.

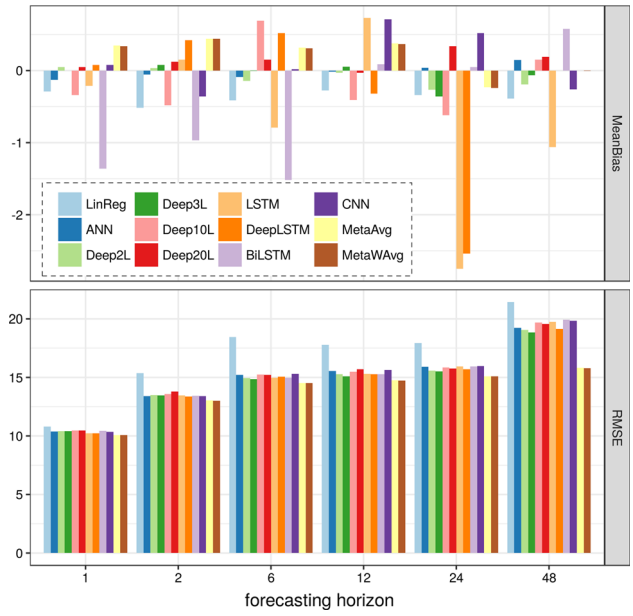


Fig. 5 Root-mean-squared error and mean bias versus forecasting horizon for the implemented models

At the lower forecasting horizons such as $h = 1$, the performance of the models is comparable. As mentioned above, when the forecasting horizon grows, deep neural networks slightly outperform shallow ones. Looking at RMSE, we can see that our deep learning system managed to improve the baseline prediction given by the Naive model for around 20% for the 1-h ahead prediction, 23% for the 48-h ahead prediction, while the biggest improvement of approximately 56% was achieved for the 6-h ahead prediction. The explanation lies in the fact that the forecasting horizon of 6 h (and similar ones) is specifically difficult for models such as the Naive predictor and linear regression because of the 24 h quasi-period of the NO_2 signal. Using the weighted average learner as a meta-learner, the performance was boosted up for additional 1.5–2.5%, depending on h.

In terms of bias, the less biased model is the Naive, as expected. This fact is related to the bias–variance trade-off [34], which explains that optimising variance-based error measures (as RMSE) is incompatible with optimising bias-based measures (as MB). However, it is remarkable that the neural network models, shallow and deep, obtain small absolute values for MB as well as for RMSE, whereas the LSTM gets bigger values for MB. It is also interesting to note how the ensembling methods get higher values of MB than some of its constituents.

Interestingly, bidirectionality in the LSTM model does not introduce any benefit in the forecasting performance. Due to the nature of this problem, LSTM in this context does not seem to learn any information by passing the input in two different directions. Moreover, other more complex models such as deeper LSTM and CNN did not manage to outperform simpler or shallower models. Likewise, deeper neural networks (with up to 20 hidden layers) do not introduce any benefit either. As a matter of fact, the error seems to slightly increase already after using more than three hidden layers in an ANN. It is an effect that has been noticed before in practical machine learning [6]. We suspect reasons for not being able to get better performance with more complex machine learning algorithms to lie in the fact that we have at our disposal a fairly small number of training examples. Indeed, we experimented with only 3 years of data, which in the era of deep learning does not represent a very large data set. To confirm our hypothesis, we would need to repeat the experiments on a larger data set, which we hope obtain.

Ensembles of models do outperform any single ML model, with no big differences between the simple average and weighted average model. However, even when taking out results of more complex models (such as BiLSTM or CNN), the performance of the ensemble model does not change much.

Comparing algorithm performances is not a trivial task even if error metrics can be computed. In order to provide statistical evidence to the evaluation of the results, a non-parametric Friedman rank test [35] is applied over the RMSE of each algorithm at each forecast horizon with a post hoc procedure as described in [36]. Since Friedman's null hypothesis of equality of medians is rejected (Table 4), the post hoc pairwise comparison was carried out to compare the algorithms. Table 5 shows there is strong evidence of differences in the performance of the group of algorithms composed by the ensemble (MetaWAvg and MetaAvg), the neural networks (Deep2L and Deep3L) and BiLSTM represented in the hypotheses 1–10 and the group composed by the Naive approach, Linear Regression and CAMS. Since the first group of algorithms outperforms the second, we can state that there is strong evidence of better forecasting capabilities when those models are applied to this problem. With regard to the remaining group, and applying the logical relation between the combination of the pairwise comparison proposed by [37], we can say that if the aforementioned outperforming group predicts better than the Naive group, it is not possible that the outperforming algorithms perform as good as the remaining and the remaining algorithms do not differ, in terms of RMSE, from the Naive algorithms. Consequently, it can be stated that MetaWAvg, MetaAvg, Deep2L, Deep3L and BiLSTM perform better than the other proposals.

Ultimately, we can conclude that, when taking into account computational resources and forecasting power of our model, the most satisfying algorithms for practical matters are ANN with up to three hidden layers.

Table 3 Test set evaluation results of the considered models for different forecasting horizons

Model	Forecasting horizon					
	1	2	6	12	24	48
RMSE						
Naive	13.01	20.64	33.91	29.69	25.53	29.18
CAMS	25.39	–	–	–	–	28.95
LinReg	10.79	15.37	18.44	17.79	17.95	21.42
ANN	10.39	13.41	15.23	15.54	15.90	19.22
Deep2L	10.41	13.48	14.94	15.27	15.56	19.06
Deep3L	10.41	13.47	14.86	15.10	15.51	18.84
Deep10L	10.48	13.57	15.23	15.48	15.84	19.69
Deep20L	10.48	13.79	15.21	16.69	15.75	19.56
LSTM	10.24	13.46	14.97	15.31	15.95	19.75
DeepLSTM	10.22	13.36	15.05	15.27	15.70	19.15
BiLSTM	10.43	13.44	14.97	15.27	15.94	19.94
CNN	10.36	13.40	15.29	15.65	15.96	19.83
MetaAvg	**10.07**	13.03	14.53	14.75	15.09	18.51
MetaWAvg	**10.07**	**13.01**	**14.52**	**14.74**	**15.08**	**18.50**
Mean bias						
Naive	− **0.01**	− **0.01**	− **0.01**	− 0.02	− **0.01**	− **0.02**
CAMS	− 13.88	–	–	–	–	− 13.94
LinReg	− 0.29	− 0.52	− 0.41	− 0.27	− 0.34	− 0.40
ANN	− 0.13	− 0.05	− 0.09	− **0.01**	0.04	0.15
Deep2L	0.05	0.03	− 0.14	− 0.03	− 0.27	− 0.19
Deep3L	**0.01**	0.08	− **0.01**	0.06	− 0.36	− 0.06
Deep10L	− 0.34	− 0.48	0.69	− 0.41	− 0.62	0.15
Deep20L	0.05	0.12	0.15	− 0.03	0.34	0.19
LSTM	− 0.21	0.15	− 0.79	− 0.73	− 2.75	− 1.06
DeepLSTM	0.08	0.42	0.52	− 0.32	− 2.54	− 0.001
BiLSTM	− 1.36	− 0.97	− 1.52	0.09	0.05	0.58
CNN	0.08	− 0.36	0.02	0.71	0.52	− 0.26
MetaAvg	0.35	0.44	0.32	0.38	− 0.23	0.001
MetaWAvg	0.34	0.44	0.31	0.37	− 0.24	− 0.005

In boldface, best results for each horizon

4.8 Reproducibility

The source code implementing the models described and the data frames used in experiments are available in [38].

5 Summary and conclusion

In this paper, we presented the core of a NO_2 forecasting system currently being developed for the Municipality of Madrid. Its main purpose is to assist in the decision-making process about the introduction of traffic restrictions in order to fight the air pollution problem. We presented results from a set of predictive models on a selected subset of forecasting horizons from 1 to 48 h. We also studied the relative importance of the predictors to determine those that contribute the most in reducing the generalisation error.

Our experiments show that, for 1 h ahead forecasting, linear models are comparable to more advanced nonlinear ones. However, as h grows and the forecasting problem becomes more difficult, neural networks outperform the other considered learners. Furthermore, our experiments show that increasing the depth of neural networks above three hidden layers does not seem to help in a decisive manner. The computation takes much longer, and it does not provide a consistently better generalisation error. Neural networks with LSTM layers gave us results comparable to feed-forward ANN, while providing a more elegant way to deal with time series albeit suffering from higher bias. The meta-learning approach of ensembling single learners proved to be another promising approach,

Table 4 Average Friedman ranking of the considered models with $F = 65.10 \sim \chi^2_{13}$ and a p value of 6.34E−9 at α = 0.05

Algorithm	Ranking
MetaWAvg	13.91 (1)
MetaAvg	13.08 (2)
Deep3L	10.41 (3)
BiLSTM	10.33 (4)
Deep2L	9.41 (5)
LSTM	7.91 (6)
ANN	7.75 (7)
DeepLSTM	7.41 (8)
CNN	6.5 (9)
Deep10L	6.16 (10)
Deep20L	6.08 (11)
LinReg	3.00 (12)
CAMS	1.66 (13)
Naive	1.33 (14)

Note that Friedman test was done using RMSE implying higher error, higher rank and consequently worse. In parenthesis, ranks based on performance and Friedman ranks

Table 5 Pairwise rejected hypothesis at $\alpha = 0.05$ with unadjusted p value and adjusted Shaffer p values

i	Hypothesis	Unadjusted p	Shaffer p
1	Naive versus MetaWAvg	1.88E−7	1.71E−5
2	CAMS versus MetaWAvg	3.93E−7	3.07E−5
3	Naive versus MetaAvg	1.14E−6	8.92E−5
4	CAMS versus MetaAvg	2.27E−6	1.77E−4
5	LinReg versus MetaWAvg	6.18E−6	4.82E−4
6	LinReg versus MetaAvg	2.98E−5	2.32E−3
7	Naive versus Deep3L	1.69E−4	0.01
8	Naive versus BiLSTM	1.94E−4	0.01
9	CAMS versus Deep3L	2.91E−4	0.02
10	CAMS versus BiLSTM	3.32E−4	0.02
11	Naive versus Deep2L	8.17E−4	0.06

boosting the performance of our system up to an overall 2% in RMSE.

We also showed how the relative importance of the predictors differs for three selected forecasting horizons: 1, 12 and 48 h and implemented an algorithm for backward feature removal. For all the forecasting horizons we studied, the CAMS numerical pollution predictions tend to be one of the most important features included. Indeed, removing these inputs from the feature set of a neural network resulted in an increase in the RMSE of about 5–10% on average. The comparison of the neural network models with the raw CAMS predictions shows how coupling global forecasts with the parameters describing the local meteorological and traffic conditions can definitely improve forecasting accuracy.

Finally, our results indicate the importance of data engineering, especially in projects where a very large set of data is not available. In other words, when there is a very large training data set, focus can be placed on developing complex deep learning systems. However, when the data are limited, it is data curation and preprocessing rather than increasing the complexity and depth what can lead to improvements in the results.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. World Health Organization (2014) 7 million premature deaths annually linked to air pollution. http://www.who.int/mediacentre/news/releases/2014/air-pollution/en/. Accessed 2 Feb 2018
2. Sellier Y, Galineau J, Hulin A, Caini F, Marquis N, Navel V, Bottagisi S, Giorgis-Allemand L, Jacquier C, Slama R, Lepeule J (2014) Health effects of ambient air pollution: do different methods for estimating exposure lead to different results? Environ Int 66:165–173. https://doi.org/10.1016/j.envint.2014.02.001
3. Arroyo V, Díaz J, Carmona R, Ortiz C, Linares C (2016) Impact of air pollution and temperature on adverse birth outcomes: Madrid, 2001–2009. Environ Pollut 218:1154–1161. https://doi.org/10.1016/j.envpol.2016.08.069
4. Díaz J, Ortiz C, Falcón I, Salvador C, Linares C (2018) Short-term effect of tropospheric ozone on daily mortality in Spain. Atmos Environ 187:107–116. https://doi.org/10.1016/j.atmosenv.2018.05.059
5. Madrid City Council (2016) Protocolo de medidas a adoptar durante episodios de alta contaminación por dióxido de Nitrógeno. http://www.mambiente.munimadrid.es/opencms/opencms/calaire/ServCiudadanos/ProtocoloNO2.html?CSRF_TOKEN=daaf25dfdd39bc9d881dffc264e11515bc9344fc. Accessed 2 Feb 2018
6. Goodfellow I, Bengio Y, Courville A (2016) Deep learning. MIT Press. http://www.deeplearningbook.org. Accessed 1 Aug 2018
7. Hrust L, Klaić ZB, Križan J, Antonić O, Hercog P (2009) Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations. Atmos Environ 43(35):5588–5596. https://doi.org/10.1016/j.atmosenv.2009.07.048
8. Gardner M, Dorling S (1999) Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. Atmos Environ 33(5):709–719. https://doi.org/10.1016/S1352-2310(98)00230-1
9. Perez P, Reyes J (2006) An integrated neural network model for PM10 forecasting. Atmos Environ 40(16):2845–2851. https://doi.org/10.1016/j.atmosenv.2006.01.010
10. Elangasinghe MA, Singhal N, Dirks KN, Salmond JA (2014) Development of an ANN—based air pollution forecasting system

with explicit knowledge through sensitivity analysis. Atmos Pollut Res 5(4):696–708. https://doi.org/10.5094/APR.2014.079
11. Gong B, Ordieres-Meré J (2016) Prediction of daily maximum ozone threshold exceedances by preprocessing and ensemble artificial intelligence techniques: case study of Hong Kong. Environ Model Softw 84(Supplement C):290–303. https://doi.org/10.1016/j.envsoft.2016.06.020
12. Kukkonen J, Partanen L, Karppinen A, Ruuskanen J, Junninen H, Kolehmainen M, Niska H, Dorling S, Chatterton T, Foxall R, Cawley G (2003) Extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modelling system and measurements in central Helsinki. Atmos Environ 37(32):4539–4550. https://doi.org/10.1016/S1352-2310(03)00583-1
13. Siwek K, Osowski S (2012) Improving the accuracy of prediction of PM10 pollution by the wavelet transformation and an ensemble of neural predictors. Eng Appl Artif Intell 25(6):1246–1258. https://doi.org/10.1016/j.engappai.2011.10.013
14. Salazar L, Nicolis O, Ruggeri F, Kisel'ák J, Stehlík M (2018) Predicting hourly ozone concentrations using wavelets and ARIMA models. Neural Comput Appl. https://doi.org/10.1007/s00521-018-3345-0
15. Aznarte JL (2017) Probabilistic forecasting for extreme NO2 pollution episodes. Environ Pollut 229(Supplement C):321–328. https://doi.org/10.1016/j.envpol.2017.05.079
16. Ayturan Y, Ayturan Z, Altun H (2018) Air pollution modelling with deep learning: a review. Int J Environ Pollut Environ Model 1(3):58–62
17. Winkler RL (1989) Combining forecasts: a philosophical basis and some current issues. Int J Forecast 5(4):605–609. https://doi.org/10.1016/0169-2070(89)90018-6
18. MACC-III monitoring atmospheric composition and climate. http://www.gmes-atmosphere.eu/. Accessed 28 Jan 2018
19. Marécal V, Peuch V-H, Andersson C, Andersson S, Arteta J, Beekmann M, Benedictow A, Bergström R, Bessagnet B, Cansado A, Chéroux F, Colette A, Coman A, Curier RL, Denier van der Gon HAC, Drouin A, Elbern H, Emili E, Engelen RJ, Eskes HJ, Foret G, Friese E, Gauss M, Giannaros C, Guth J, Joly M, Jaumouillé E, Josse B, Kadygrov N, Kaiser JW, Krajsek K, Kuenen J, Kumar U, Liora N, Lopez E, Malherbe L, Martinez I, Melas D, Meleux F, Menut L, Moinat P, Morales T, Parmentier J, Piacentini A, Plu M, Poupkou A, Queguiner S, Robertson L, Rouïl L, Schaap M, Segers A, Sofiev M, Tarasson L, Thomas M, Timmermans R, Valdebenito A, van Velthoven P, van Versendaal R, Vira J, Ung A (2015) A regional air quality forecasting system over Europe: the MACC-II daily ensemble production. Geosci Model Dev 8(9):2777–2813. https://doi.org/10.5194/gmd-8-2777-2015
20. Madrid City Council, catalogue of open data. https://datos.madrid.es/portal/site/egob/. Accessed 15 Jan 2018
21. Zhang Y, Bocquet M, Mallet V, Seigneur C, Baklanov A (2012) Real-time air quality forecasting, part I: history, techniques, and current status. Atmos Environ 60(Supplement C):632–655. https://doi.org/10.1016/j.atmosenv.2012.06.031
22. Zhang Y, Bocquet M, Mallet V, Seigneur C, Baklanov A (2012) Real-time air quality forecasting, part II: State of the science, current research needs, and future prospects. Atmos Environ 60(Supplement C):656–676. https://doi.org/10.1016/j.atmosenv.2012.02.041
23. Grabczewski K, Jankowski N (2005) Feature selection with decision tree criterion. In: Fifth international conference on hybrid intelligent systems (HIS'05). IEEE, p 6. https://doi.org/10.1109/ICHIS.2005.43
24. Chandrashekar G, Sahin F (2014) A survey on feature selection methods. Comput Electr Eng 40(1):16–28. https://doi.org/10.1016/j.compeleceng.2013.11.024**(40th-year commemorative issue)**
25. Tharwat A (2016) Principal component analysis—a tutorial. Int J Appl Pattern Recognit 3:197. https://doi.org/10.1504/IJAPR.2016.079733
26. Bergmeir C, Benítez JM (2012) On the use of cross-validation for time series predictor evaluation. Inf Sci 191:192–213
27. James G, Witten D, Hastie T, Tibshirani R (2014) An introduction to statistical learning: with applications in R. Springer, Berlin
28. Kingma DP, Ba J (2014) Adam: a method for stochastic optimization, CoRR. arXiv:abs/1412.6980
29. Ruder S (2016) An overview of gradient descent optimization algorithms, CoRR. arXiv:abs/1609.04747
30. Chollet F et al (2015) Keras. https://github.com/keras-team/keras. Accessed 1 Aug 2018
31. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
32. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324. https://doi.org/10.1109/5.726791
33. Polikar R (2006) Ensemble based systems in decision making. IEEE Circuits Syst Mag 6(3):21–45. https://doi.org/10.1109/MCAS.2006.1688199
34. Geman S, Bienenstock E, Doursat R (1992) Neural networks and the bias/variance dilemma. Neural Comput 4(1):1–58. https://doi.org/10.1162/neco.1992.4.1.1
35. Friedman M (1937) The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J Am Stat Assoc 32:674–701
36. Navares R, Aznarte J (2016) What are the most important variables for Poaceae airborne pollen forecasting? Sci Total Environ 579:1161–1169
37. Shaffer J (1986) Modified sequentially rejective multiple test procedures. J Am Stat Assoc 81:826–831
38. Valput D, Aznarte JL (2018) Air pollution forecasting system, Madrid: Source code. https://github.com/dvalps/Air-quality-forecasting-Madrid. Accessed 1 Aug 2018

Chapter 8

Predicting Air Quality with Deep Learning LSTM: Towards Comprehensive Models

Type:	Published Article
Title:	*Predicting Air Quality with Deep Learning LSTM: Towards Comprehensive Models*
Journal:	Ecological Informatics
Authors:	Ricardo Navares & José Luis Aznarte
Published:	October 2019
Impact Factor:	2.310
Quartile:	Q2
DOI	10.1016/j.ecoinf.2019.101019

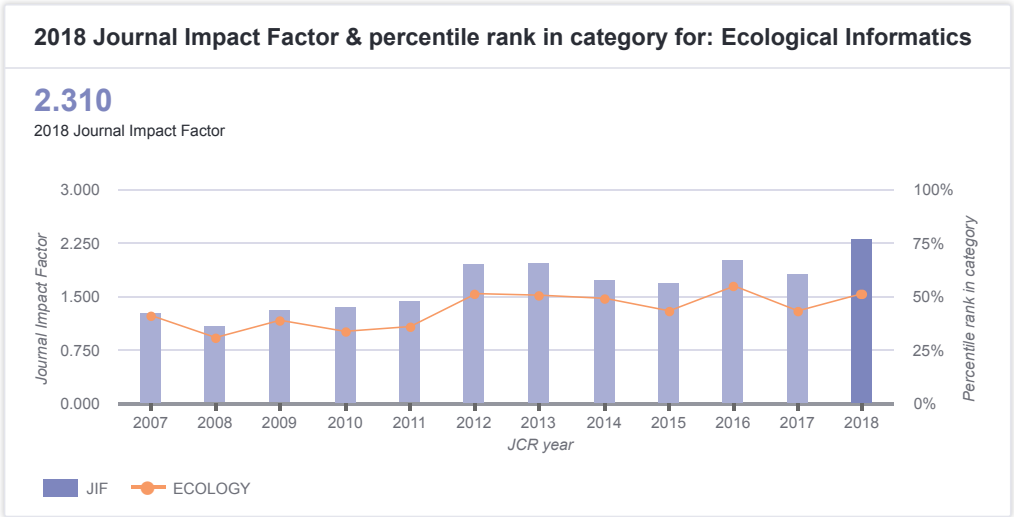


FIGURE 8.1: Impact factor Ecological Informatics

Ecological Informatics 55 (2020) 101019

Contents lists available at ScienceDirect

Ecological Informatics

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

Predicting air quality with deep learning LSTM: Towards comprehensive models

Ricardo Navares, José L. Aznarte*

Department of Artificial Intelligence, UNED, Juan del Rosal, 16, 28040 Madrid, Spain

ARTICLE INFO

Keywords:
Air quality
Forecast
Neural networks
Deep learning

ABSTRACT

In this paper we approach the problem of predicting air quality in the region of Madrid using long short term memory recurrent artificial neural networks. Air quality, in this study, is represented by the concentrations of a series of air pollutants which are proved as risky for human health such as CO, NO_2, O_3, PM10, SO_2 and airborne pollen concentrations of two genus (Plantago and Poaceae). These concentrations are sampled in a set of locations in the city of Madrid. Instead of training an array of models, one per location and pollutant, several comprehensive deep network configurations are compared to identify those which are able to better extract relevant information out of the set of time series in order to predict one day-ahead air quality. The results, supported by statistical evidence, indicate that a single comprehensive model might be a better option than multiple individual models. Such comprehensive models represent a successful tool which can provide useful forecasts that can be thus applied, for example, in managerial environments by clinical institutions to optimize resources in expectation of an increment of the number of patients due to the exposure to low air quality levels.

1. Introduction

In the last decades, air quality has been gaining attention due to the health threats produced by high levels of environmental pollution (Ozkaynak et al., 2009). Within the context of this study, air quality is related to both chemical pollutants and biotic factors present in the environment. Concretely, chemical pollutants are considered as the agents released in the environment which disrupt ecosystems such as CO, O_3, NO_2, SO_2 and PM10 which are also considered as the main air chemical pollutants in the studied region (Querol et al., 2012). On the other hand, biotic factors refer to airborne pollen concentrations of the Plantago and Poaceae genus which are the most common and aggressive in terms of allergic and respiratory disorders (Subiza et al., 1995).

Air quality information systems are increasingly used to predict future air pollution levels, which allows for alerting about peaks in admissions in clinical institutions, traffic and environmental management in urban areas or minimizing the exposure for patients in order to prevent adverse effects (Abraham et al., 2009; González et al., 2001; Linares and Díaz, 2008; Ozkaynak et al., 2009).

Field experts have been employing observation-based models to relate records of pollutants to one or more variables which can be measured of predicted, usually meteorological data (Navares and Aznarte, 2016; Sabariego et al., 2012; Schaber and Badeck, 2003; Silva-Palacios et al., 2016; Smith and Emberlin, 2006). Despite the extensive literature, few consider the problem of taking into account both types of pollutants altogether as they are inherently different problems: atmospheric pollen concentrations depend on plant development during previous seasons which, at the same time, depends on the climatological conditions during plant evolution (Cannell and Smith, 1983; Smith and Emberlin, 2006). This implies long and mid-term relations between past atmospheric conditions and current plant status. Contrarily, chemical air pollutant levels are related to recent past atmospheric conditions (Navares et al., 2018). Both pollutants show influence on the development and attack of, for instance, allergic respiratory diseases (D'Amato et al., 2011).

Neural network models have been successfully applied to environmental modeling (Gardner and Dorling, 1998) and air quality problems (Castellano-Méndez et al., 2005; Chaloulakou et al., 1998; Chelani et al., 2002; Grivas and Chaloulakou, 2006; Iglesias-Otero et al., 2015). However, given the nature of the problem (short term influential variables for chemical pollutants and mid-long term influential variables for pollen) this approach requires a thorough research and selection of relevant variables based on expert knowledge (Andersen, 1991; Catalano et al., 2016; Navares and Aznarte, 2016). In addition to the temporal dimension, it is important to take into account the spatial

* Corresponding author.
E-mail address: jlaznarte@dia.uned.es (J.L. Aznarte).

https://doi.org/10.1016/j.ecoinf.2019.101019
Received 22 August 2019; Received in revised form 23 September 2019; Accepted 4 October 2019
Available online 24 October 2019

interactions between observation stations as they might be implicitly related. These approaches imply a new research process which might add a new set of influential variables every time a new kind of pollutant or a new genus of pollen is taken into consideration by the system.

In this paper we propose several long short term memory (LSTM) network setups (Hochreiter and Schmidhuber, 1997) to gain insights on how influential is network design when dealing with interrelated time series of different nature. The study compares network topologies in order to find the most suitable configuration to solve the problem, exposing the advantages and disadvantages of each one. The objective is twofold: on the one hand, we show how to avoid thorough preprocessing steps to find influential features (both long and short term, and with differences at each location as a result of particular environmental conditions of the areas where the observation stations are located) by letting the network extract them regardless of the pollutant type. Such a unified approach avoids manually fitting one specific model per pair of location and pollutant, saving human resources and increasing the scalability of the system. On the other hand, we provide a convenient network topology for accurate forecasts at each location which is able to obtain relevant information from data both temporal and spatial. The problem chosen to prove the validity of the proposal is the prediction of air quality over a dataset which consist of a dozen of time series with different characteristics and regimes, sampled over 13 adjacent locations.

2. Materials and methods

2.1. Data description

Chemical air pollutants were measured using the gravimetric method or an equivalent method (β-attenuation) and were provided by the Madrid Municipal Air Quality Monitoring Grid (http://www.mambiente.munimadrid.es/). The grid consists of a network of 24 urban background stations spread across the city, which capture chemical air pollutants in real time. Hourly data was aggregated to obtain daily mean levels of chemical air pollutants for the study period from 01 to 01-2001 to 31-12-2013. Daily mean concentrations ($\mu g/m^3$) of particulate matter 10 μm in diameter (PM10), carbon monoxide, surface ozone (O_3) and nitrogen dioxide (NO_2) in $\mu g/m^3$.

Pollen observations correspond to daily grains per cubic meter of Poaceae and Plantago pollen registered at Complutense University of Madrid (Pharmacy Faculty). Pollen counts followed the standard methodology of the Spanish Aerobiological Network (Galán Soldevilla et al., 2007) and were provided by Red Palinológica de la Comunidad de Madrid.

Weather observations consist of average daily temperature in Celsius degrees, wind speed measured in *m/s*, daily rainfall in *mm/h*, pressure in *hPa* and degree of humidity in percentage. Data sets for locations (Table 1) consist of observations from 01 to 01-2001 to 31-12-2013. In the presence of missing hourly data, when these observations are less 20% within a day, averages and aggregations were calculated ignoring missing values, otherwise, the daily data is considered as missing. This approach led to complete time series during the study period.

Different pollutants have different nature of the time series having significant seasonal patterns. Some of them show higher variability around the seasonal component as in the case of O_3 and PM10, while others present highers peaks and less variability during the rest of the year as in the pollen counts. It is noticeable that SO_2 and CO concentrations dramatically decreased when compared to early observations due to the progressive transition from coal-powered heating systems to natural gas and the progressive upgrade of the urban car fleet (Díaz et al., 2007).

2.2. Methodology

Compared to traditional neural networks, recurrent neural networks (RNN) are implemented with loops or connections between units allowing information persistence from one step of the network to the next. The ability to map input sequences to output sequences by incorporating past context into their internal state makes them especially promising for tasks that require to learn how to use past information such as time series analysis. RNNs can be thought of as multiple copies of the same neural network, each transferring information to its successor and forming a chain-like architecture which is naturally related to sequences.

RNNs might be able to look at recent information to perform a present task which makes them suitable for time series predictions. However, relevant information might appear further in the past and, as the time gap grows, RNNs are unable to connect the information.

Long short-term memory networks (LSTM) were first introduced in 1997 by (Hochreiter and Schmidhuber, 1997) and improved in 2000 by (Gers et al., 2000). They are a variation of RNNs capable of learning long-term dependencies by including in the architecture special units called memory blocks. In addition, multiplicative units called gates control the flow of information from a LSTM unit to another.

An LSTM unit performs self-loops which enable the flow of the gradient for long durations, enabling it to deal with the vanishing gradient problem. Together with an input gate, an output gate and a forget gate, this architecture models the short-term memory that allows the network to learn over many time steps. For this reason, LSTM had been shown to outperform more traditional recurrent networks on several temporal processing tasks (Gers et al., 2000).

The common scoring rule root mean squared error (RMSE) will be used to measure the average magnitude of the error of several network configurations:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{1}$$

where y_i is the observed i^{th} data point, $\hat{y}_i$ the predicted and n the total number of data points in the test set. As benchmark models the LSTMs are compared to the traditional linear regression (LinReg) and computational intelligence technique Random Forest (RF) proposed by (Breiman, 2001). In order to evaluate the results obtained by the algorithms, we use a nonparametric Friedman test (Friedman, 1937) test with a post-hoc procedure as described in (Navares and Aznarte, 2016) to determine, in significance terms, which algorithms are considered the best performers based on their RMSE.

The Friedman test (Friedman, 1937) is a non-parametric analogue of the parametric two-way ANOVA. The objective of the application of the test is to determine if there is a difference among model performances and consequently, whether one (or more) is consistently better than the others. Given that non-parametric hypothesis tests are applied to nominal or ordinal data, the original computed RMSE for each location is converted to its correspondent rank within the set and combined by averaging: $R_j = \frac{1}{n}\sum_i r_i^j$ (where j denotes the model, i refers to each location and n is the total number of pairs {model, location}). Since the error is being used to compare the models, the highest rank 1 will be assigned to the highest error, thus the worst performer. The null hypothesis of equality of medians is tested by the F-statistic

$$F = \frac{12n}{k(k+1)}\left[\sum_j R_j^2 - \frac{k(k+1)^2}{4}\right], \tag{2}$$

where k is the number of algorithms and $F \sim {\chi_{k-1}}^2$. Still, this test is not sufficient as it only indicates the presence of significant differences in the whole variable space. A ranking conversion is computed to obtain the p-value of each pair (Conover, 1999). The former is a valid procedure to compare two models but is not suitable for multiple comparison

Table 1
Availability of variables and locations.

	Long.	Lat.	CO	NO_2	O_3	Plantago	Poaceae	PM10	SO_2	Pr	R	Hum	T	W
ArturoSoria	3° 38′ W	40° 26′ N	*	*	*						*			
BarrioPilar	3° 42′ W	40° 28′ N	*	*	*						*			
CasaCampo	3° 44′ W	40° 25′ N	*	*	*			*	*	*	*	*	*	*
CuatroCaminos	3° 42′ W	40° 26′ N		*					*		*		*	
Farmacia	3° 45′ W	40° 27′ N				*	*							
Farolillo	3° 43′ W	40° 23′ N	*	*	*				*		*		*	
Moratalaz	3° 38′ W	40° 24′ N	*	*				*	*					
PlazaEspana	3° 42′ W	40° 25′ N	*	*					*	*	*	*	*	*
PzadelCarmen	3° 42′ W	40° 25′ N	*	*	*				*					
PzaLadreda	3° 43′ W	40° 23′ N								*	*	*	*	*
RamonyCajal	3° 40′ W	40° 27′ N		*							*			
StaEugenia	3° 36′ W	40° 22′ N									*		*	*
Vallecas	3° 39′ W	40° 23′ N		*				*	*		*			

Pr: Pressure.
R: Rain.
Hum: Hummidity.
W: Wind speed.
T: Average Temperature.
* represent the presence of data in the corresponding location and for the corresponding pollutant

as there is no control of error propagation (Type I errors) when making more than one comparison.

Thus, once the existence of significant differences in the group of models is evidenced, a post-hoc test adjusts the value of the significance level α at each pairwise comparison to allow multiple comparisons. (Holm, 1979) proposed the adjustment by selecting the p-values of each test, starting with the most significant p_i, and test the hypothesis of $H_i : p_i > \alpha/(k - i)$, being k the total number of models in our proposal. If H_i is rejected then allows to test H_{i+1}, being p_{i+1} the next most significant p-value and so on. An extension of this step-down method was proposed by (Shaffer, 1986), which uses a logical relation between the combination of the hypotheses of all pairwise comparisons. For instance, if a model a_1 is better/worse than a_2, it is not possible that a_1 is as good/bad as a_3 and a_2 has the same performance as a_3. Based on this argument and following Holm's method, instead of rejecting $H_i : p_i \leq \alpha/(k - i)$, rejects $H_i \leq \alpha/t_i$, being t_i the maximum number of hypotheses which can be true given the number of false hypotheses in $j \in \{1, \ldots, i\}$.

2.3. Experimental design

The aim is to provide the best one day-ahead forecast in terms of accuracy, and consequently to see if LSTMs are able to efficiently store relevant information over time to position themselves as a strong candidate when deciding which technique to use in such problems. In order to do so, the full historic data set was split into a training set consisting of the period between 01 and 01-2001 and 31-12-2012, leaving the last period (01-01-2013 to 31-12-2013) as a test set.

Selecting the topology of LSTM networks depends on the application domain and there is no general rule of thumb for the amount of hidden nodes that should be used. It has to be figured out case-specifically by trial and error. Thus, starting with the simplest network of one memory cell, the architecture is extended by including units in the layer. The stop condition is given by the validation error from the random selected 10-fold cross-validation on the training set. Even though time series show inherent serial correlation and potential non-stationarity of the data, (Bergmeir et al., 2018) proves that cross-validation empirically compares to out-of-sample or other time-series-specific valuation techniques.

Different architectures are proposed to check the convenience of each one to solve the problem. The first one consists on a fully connected LSTM (FC-LSTM), which is the first and most common approach, where all the input variables are parsed through a 500 LSTM hidden units layer to test the capability of the LSTM to obtain and discriminate relevant information. This layer is connected to a layer of 100 sigmoid units to obtain higher order relations among the locations and pollutants (Fig. 1(a)). The idea is to see if the first layer captures the temporal dependencies while the subsequent dense layer deals with location dimensions which will be then transferred to the output layer.

The second configuration is tailored to ease the discrimination of information by forcing the LSTM units to target different groups of pollutants (GP-LSTM). The core idea is to force each LSTM group to focus on its correspondent pollutant outputs assuming that the same pollutants behave similarly across locations and that the LSTM units discriminate input information to subsequently obtain the relations of the locations only per pollutant group. Thus, the LSTM layer consists of 100 units per group which receive the full set of 1-day lagged input variables and it is connected only to a number of outputs equal to the number of observation stations of that group (Fig. 1(b)). The intention is to facilitate that LSTM units extract the information of each individual group of pollutants and then obtain the relations among groups in the output layer. For instance, in the case of carbon monoxide, the 57 input variables (Table 1) feed 100 LSTM units which are fully connected to 7 outputs, one per station where CO is available. This configuration totals 700 LSTM units, 100 per pollutant.

An alternative setup with a similar configuration and the same aim of easing the discrimination process is also proposed. In this case, the network is assisted by only using as input for each group of LSTM units individual groups of pollutants (IGP-LSTM) and the meteorological variables as seen in Fig. 1(c). Each group consists of 100 LSTM units with a total of 700 LSTM units in the layer, one per group of pollutants, which are fully connected to the output layer. Since the inputs are split by pollutant, the reason behind including a fully connected output layer is to include the interactions among pollutants.

Finally, a more simple version was taken into consideration by training the network using the full set of inputs to feed 7 groups of 66 LSTM units (462 in total), one group per single combination pollutant-station (SP-LSTM, Fig. 1(d)) which outputs one single target variable. This is equivalent to training 31 individual networks (one per target pollutant and station) with a set up 57–66-1 where 57 is the number of 1-day lagged input variables (Table 1), 66 LSTM units and 1 is the output. Fig. 1 summarizes the four different network topologies which were tested against the benchmark algorithms.

In order to obtain the relationship between the inputs and outputs, the well-known *backpropagation* algorithm (Rumelhart et al., 1986) was proposed, employing as loss function the mean squared error since it is a regression problem. As an alternative of the classic Stochastic Gradient Descent (SGD) optimization model to fit network weights, the *Adam* algorithm proposed by (Kingma and Ba, 2019) was used with a

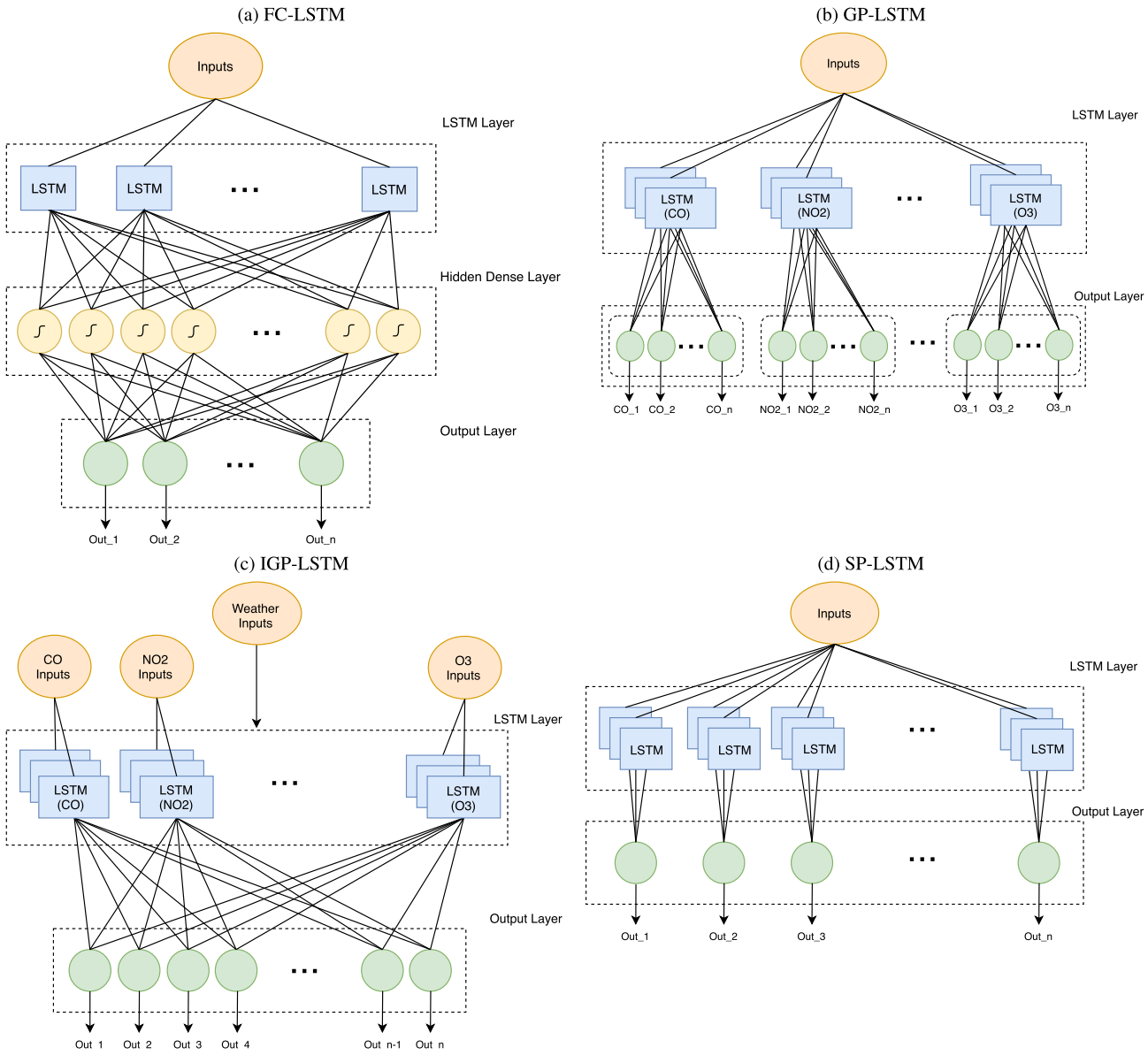


Fig. 1. LSTM architectures: (a) Fully connected LSTM (FC-LSTM), (b) LSTM grouped by pollutant class (GP-LSTM), (c) LSTM fed by pollutant variable (IGP-LSTM), (d) One variable LSTM (SP-LSTM).

learning rate $\alpha = 0.001$, an exponential decays of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ were used as suggested by (Ruder, 2016). As opposed to SGD which maintains a single learning rate α for all weight updates, the method computes individual adaptive learning rates from the estimates of first and second moments of the gradients (Kingma and Ba, 2019). Specifically, the algorithm uses an exponential moving average of the gradient and its square using the parameters β_1 and β_2 to control the decay rates of these moving averages. Performances are compared to the traditional and commonly used linear regression enabling the identification and the characterization of relationships among each pollutant and meteorological variables and its 1-day lagged observation. Therefore, one linear model is fitted per pollutant. As an extension, an identical setup is used to train random forests to compare with another family of computational intelligence models.

3. Results

Table 2 shows prediction errors for each pollutant at each location, while Table 2 shows the average prediction errors for each pollutant. Linear regression obtains an average RMSE of 0.107 across all location for carbon monoxide (CO) while Random Forest manages to diminish this error to 0.086. All LSTM configurations outperform Random Forest results except SP-LSTM which results in an average RMSE of 0.093 mainly due to the error at Farolillo where it underperforms with an RMSE of 0.127. GP-LSTM is the most accurate in forecasting CO levels with an average RMSE of 0.083. In Fig. 3 can be seen the percentage improvement of each algorithm with respect to LinReg. There is an average improvement around 20% with all computational intelligence based algorithms for this pollutant, except for the aforementioned SP-LSTM, which performs poorly (-20.27%) at the observation station of Farolillo.

With regard to Nitrogen dioxide, an average RMSE of 12.07 is shown for LinReg followed by RF and SP-LSTM with an error of 10.60 and 11.00 respectively. The remaining LSTM configurations reduce this error to RMSE levels lower than 9.70 implying an increase in accuracy of around 20% when compared to LinReg as shown in Fig. 3.

Results for ozone are paired and oscillate around an error of 11 $\mu g/m^3$ being RF and FC-LSTM the worst and best performer in average

Table 2
RMSE per location and variable. Highlighted in bold best average performer for each pollutant.

Pollutant	Station	LinReg	RF	SP-LSTM	IGP-LSTM	GP-LSTM	FC-LSTM
CO ($\mu g/m^3$)	ArturoSoria	0.087	0.072	0.086	0.080	0.080	0.087
	BarrioPilar	0.137	0.106	0.120	0.119	0.111	0.114
	CasaCampo	0.072	0.055	0.055	0.054	0.059	0.057
	Farolillo	0.106	0.100	0.127	0.083	0.082	0.077
	Moratalaz	0.110	0.087	0.079	0.081	0.075	0.078
	PzaCarmen	0.124	0.087	0.089	0.080	0.082	0.083
	PzaEspana	0.116	0.093	0.096	0.097	0.091	0.091
Average		0.107	0.086	0.093	0.085	**0.083**	0.084
NO_2 ($\mu g/m^3$)	ArturoSoria	12.310	11.817	13.052	10.103	10.312	10.514
	BarrioPilar	13.685	11.623	11.933	10.560	10.427	10.479
	CuatroCaminos	11.885	10.290	10.422	10.442	9.471	9.474
	CasaCampo	10.047	8.232	7.826	7.515	7.439	7.632
	Farolillo	10.667	8.551	8.966	8.114	8.020	7.862
	Moratalaz	13.163	12.325	12.312	10.607	9.799	9.932
	PzaCarmen	11.002	9.710	9.448	8.300	8.004	8.288
	PzaEspana	12.041	11.454	11.561	10.339	10.005	10.875
	RamonyCajal	13.629	12.515	13.248	11.588	11.197	10.731
	Vallecas	12.327	9.550	11.273	8.983	9.739	8.592
Average		12.076	10.607	11.004	9.655	9.442	**9.438**
O_3 ($\mu g/m^3$)	ArturoSoria	10.501	10.945	10.957	10.836	11.795	10.139
	BarrioPilar	11.442	11.471	11.484	10.796	11.343	10.506
	CasaCampo	12.392	13.809	13.866	12.042	14.108	12.526
	Farolillo	11.471	12.268	12.039	11.550	11.981	10.872
	PzaCarmen	10.063	11.507	9.730	9.742	9.536	9.840
Average		11.174	12.000	11.615	10.993	11.753	**10.777**
PM10 ($\mu g/m^3$)	CasaCampo	8.094	6.314	6.784	5.323	5.350	5.700
	Moratalaz	8.244	6.458	5.969	6.147	5.965	6.382
	Vallecas.	9.005	6.672	6.102	5.856	5.847	5.548
Average		8.447	6.481	6.285	5.776	**5.721**	5.877
SO_2 ($\mu g/m^3$)	CuatroCaminos	2.213	1.815	1.787	1.571	1.695	1.727
	CasaCampo	0.831	1.033	1.188	0.613	0.665	0.742
	Farolillo	1.253	1.149	1.366	0.832	0.809	0.855
	Moratalaz	4.186	4.965	4.396	4.192	4.165	4.065
	PzaCarmen	1.869	1.709	1.779	1.730	1.605	1.748
	PzaEspana	1.336	1.218	1.631	1.122	1.159	1.202
	Vallecas	1.455	1.342	1.206	1.272	1.268	1.422
Average		1.877	1.890	1.908	**1.619**	1.624	1.680
Plantago (grains/m^3)	Farmacia	6.948	9.927	7.927	6.938	6.875	7.296
Poaceae (grains/m^3)	Farmacia	24.344	24.220	25.738	23.686	25.268	24.933
Average		15.646	17.074	16.833	**15.312**	16.072	16.114

respectively. It can be seen in Fig. 3 that RF and SP-LSTM underperform LinReg while the other networks show mixed results depending on the location.

On the other hand, computational intelligence models clearly overcome, in terms of accuracy, linear regression when predicting particulate matter (PM10), having an average improvement between 25% and 30% except for RF and SP-LSTM which perform 23.16% and 6.18% better respectively. Similar situation occurs when using GP-LSTM and FC-LSTM to predict SO_2 averaging a RMSE of 1.624 and 1.68 respectively. Although the best average performer for this pollutant is IGP-LSTM, it fails to improve LinReg results at Moratalaz where the error is 0.16% higher. For the pollen series, IGP-LSTM is the only model which improves LinReg results in both series by 0.15% and 2.70% in Plantago and Poaceae respectively.

In general, both LSTM and RF perform better than linear regression (as seen in the left part of Fig. 2). Table 2 shows that LSTM architectures achieved lower errors. Figs. 2 and 3 clearly show these improvements are higher for IGP-LSTM, GP-LSTM and FC-LSTMS than the other models. Regarding bias, shown in the right part of Fig. 2, we can see how GP.LSTM and IGP.LSTM show zero bias in median while the former produces predictions with very low bias also in mean.

In light of the combined error/bias results, we would be tempted to choose GP.LSTM as the best overall model. In order to investigate this and provide statistical evidence, a Friedman rank test was performed over the errors shown in Table 2. A Friedman statistic of $F = 72.96$, distributed according a χ^2 with 5 degrees of freedom obtains a p-value of 4.34e-11 with $\alpha = 0.05$, which provides strong evidence of the existence of significant difference between the algorithms ranked GP-LSTM as the best performer (as expected), followed by FC-LSTM and IGP-LSTM and RF, SP-LSTM and linear regression with ranks 4, 5 and 6 respectively.

Since Friedman's null hypothesis was rejected, a post-hoc pairwise comparison was carried out to check the differences between the proposed algorithms. Table 3 shows there is strong evidence of differences between GP-LSTM, IGP-LSTM, FC-LSTM and linear regression (hypotheses 1, 2 and 3) and SP-LSTM (hypotheses 4, 5 and 6) which implies, given Friedman ranks, that GP-LSTM, IGP-LSTM and FC-LSTM perform better than the benchmark LinReg and the SP-LSTM configuration. Hypotheses 7, 8 and 9 provide statistical evidence of GP-LSTM, IGP-LSTM and FC-LSTM performing better than RF. Lastly, there is no difference in terms of error between SP-LSTM and the benchmark algorithms.

4. Discussion

As we have seen in Section 3 there is statistical evidence of the outperformance of GP-LSTM, IGP-LSTM and FC-LSTM with respect to the other proposed methods. This situation is clear for CO, NO_2 and PM10 where there is an error reduction higher than 10% at all locations except for Arturo Soria when forecasting CO. With a close look at this location, we have seen a yearly average concentrations of 0.40 $\mu g/m^3$ with a standard deviation of 0.24 $\mu g/m^3$ while this average goes over 0.47 $\mu g/m^3$ with a standard deviation of at least 0.32 $\mu g/m^3$ for the remaining locations, suggesting a lower improvement when the

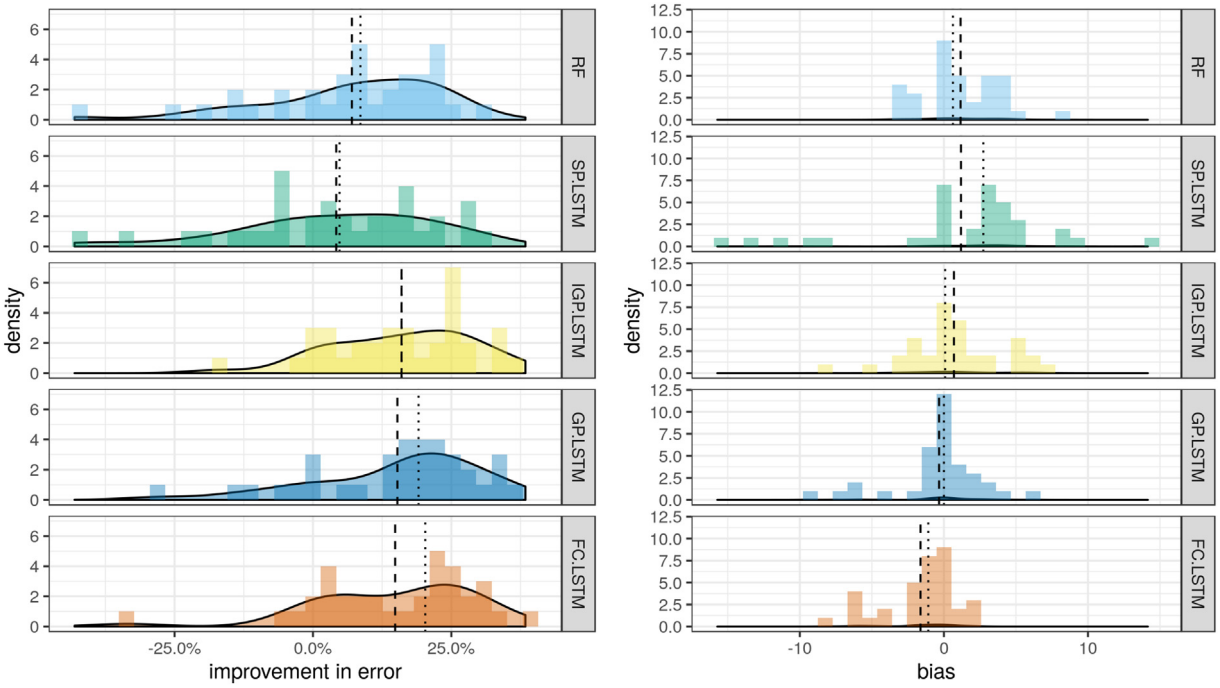


Fig. 2. Relative improvement with respect to LinReg of each of the methods applied (left) and bias (right). Dashed vertical line represents the mean, dotted vertical line represents median.

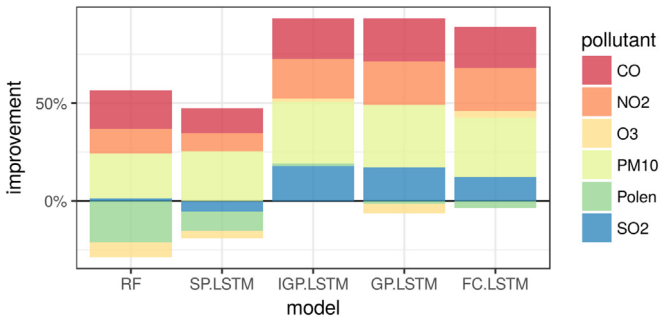


Fig. 3. Average improvement over LinReg per pollutant.

Table 3
Adjusted Holm and Schaffer *p*-values with pairwise rejected hypothesis at $\alpha = 0.05$.

i	Hypothesis	$p_{unadj.}$	p_{Holm}	p_{Shaf}
1	Linear vs GP-LSTM	5.171E-10	7.756E-9	7.756E-9
2	Linear vs IGP-LSTM	9.131E-9	1.369E-7	9.131E-8
3	Linear vs FC-LSTM	9.131E-9	1.369E-7	9.131E-8
4	SP-LSTM vs GP-LSTM	3.815E-7	5.722E-6	3.815E-6
5	SP-LSTM vs IGP-LSTM	4.021E-6	4.021E-5	4.0215E-5
6	SP-LSTM vs FC-LSTM	4.021E-6	4.423E-5	4.021E-5
7	RF vs GP-LSTM	1.398E-4	0.001	9.787E-4
8	RF vs IGP-LSTM	8.354E-4	0.006	0.005
9	RF vs FC-LSTM	8.354E-4	0.006	0.005
10	Linear vs RF	0.016	0.096	0.096
11	RF vs SP-LSTM	0.204	1.021	0.817
12	Linear vs SP-LSTM	0.256	1.024	1.024
13	IGP-LSTM vs GP-LSTM	0.639	1.919	1.919
14	GP-LSTM vs FC-LSTM	0.639	1.919	1.919
15	IGP-LSTM vs FC-LSTM	1.0	1.919	1.919

variability reduces. Nevertheless, most of the improvements are over 20%.

It is to note the low performance of SP-LSTM when forecasting CO at Farolillo (-20.27%). A simple analysis of the series shows that the standard deviation taken by years goes from 0.73 $\mu g/m^3$ in 2001 to 0.12 $\mu g/m^3$ in 2012 while this change in the behavior of the series is more constant and smoother at other locations. Not including a fully connected layer, either a dense sigmoid or output, weights more past information from the only target series (CO at Farolillo) incurring in overfitting.

The best performing configurations, GP-LSTM, IGP-LSTM and FC-LSTM, manage to obtain good improvements in the case of SO_2 even though, there are dramatic changes on series patterns with respect to past years due to the progressive transition from coal-powered heating to natural gas. An exception is the results at Moratalaz where the error improvement decreases for all models with respect to their performances at other locations. The particular location of this station makes the observed levels of SO_2 behave differently when compared to other areas. It is located between the A3 and R3 highways, which are the two main access to Madrid from the East, and M30 and M40 which are city main beltways. Consequently, SO_2 levels remain high due to the dense traffic. While SO_2 mean values for the most recent years of this study stay around 3.15 $\mu g/m^3$ with average maximums of 18.90 $\mu g/m^3$, at Moratalaz these means double up to 6.43 $\mu g/m^3$ with maximums of 34.66 $\mu g/m^3$.

Random Forest does not improve LinReg when predicting O_3 at all studied locations. A similar situation occurs with SP-LSTM with the exception of Plaza del Carmen, suggesting these two proposals are not able to capture the strong seasonal pattern of this pollutant. These patters are captured by FC-LSTM improving LinReg average performance by 3.5% although this does not apply for the observations at Casa de Campo. The reason is because tropospheric ozone behaves opposite to other pollutants as its concentration levels are higher outside urban centers where the air quality is assumed as clean in general. Not only is it formed due to directly traffic or industrial emissions but also when combined these with airborne pollutants and solar radiation (Sharma et al., 2016), having Casa de Campo all the conditions to concentrate high levels compared to other locations as it is the largest public park in Madrid.

With respect to airborne pollen concentrations, IGP-LSTM is the only model which managed to improve LinReg results for both pollen

genus. Pollen series are particularly characteristic since they show very low concentrations (or almost none) during the calendar year until the main pollination season where the high peaks appear. Nevertheless, IGP-LSTM is able to identify those peaks, specially in the case of Plantago where the test set includes the highest peak (around 120 grains/m^3) among all observed train data.

Fig. 2 (left hand side) shows that IGP-LSTM presents a higher consistency of results as the median improvement with respect to LinReg does not differ much from the average. On top of that, IGP-LSTM presents the shorter negative tail among all the models, while GP-LSTM and FC-LSTM show a heavy-tailed distribution of improvements suggesting these configurations are not as good as IGP-LSTM when obtaining the characteristics of some time series. In fact, Fig. 3 shows that, on average, they do not manage to improve LinReg results for O_3 and Pollen respectively. However, when considering the bias of the predictions (right hand side of Fig. 2) we see how the predictions of GP-LSTM are, in general, slightly better than the rest.

5. Conclusions

This paper presents a comparison study of different LSTM configurations in order to obtain the most suitable to forecast air quality in the region of Madrid. Several pollutants showing very different behaviors were taken into consideration. In addition to the intrinsic differences in the behavior among pollutant types, each pollutant behaves differently at each location given the conditions of the zones studied due to several factors such as traffic congestion or green areas. This adds an extra dimension to the complexity problem which let test the capability of the proposed models to obtain relevant information for forecasting.

We have seen that there is statistical evidence that the LSTM grouped by pollutant class (GP-LSTM) and the LSTM with individual groups of pollutants as inputs (IGP-LSTM) outperform benchmark algorithms and the other two proposals. Furthermore, the GP-LSTM shows smaller bias, but evidence also shows that performing a discriminative input of the groups of pollutants as in IGP-LSTM eases the network to focus on the relevant information and provides more stable results across locations.

By including in the configuration fully connected layers, either as an output or hidden layer, the networks are able to better identify the relations among pollutants with no data preprocessing. However, we have seen that there is still room for improvement as the LSTMs struggle to identify the presence of sudden high peaks as past information weights on the predictions. This situation can be mitigated by capping pollutant observation levels to thresholds over which it implies risk for human health.

References

Abraham, G., Byrnes, G., Bain, C., 2009. Short-term forecasting of emergency inpatient flow. Inf. Technol. Biomed. 13, 380–388.

Andersen, T.B., 1991. A model to predict the beginning of the pollen season. Grana 30, 269–275.

Bergmeir, C., Hyndman, R., Koo, B., 2018. A note on the validity of cross-validation for evaluating autoregressive time series prediction. Comput. Stat. Data Anal. 120, 70–83. https://doi.org/10.1016/j.csda.2017.11.003.

Breiman, L., 2001. Random forests. Machine Learning 45 (1), 5–32. https://doi.org/10.1023/A:1010933404324.

Cannell, M., Smith, R., 1983. Thermal time, chill days and prediction of budburst in *Picea sitchensis*. J. Appl. Ecol. 20, 269–275.

Castellano-Méndez, M., Aira, M.J., Iglesias, I., Jato, V., González-Manteiga, W., 2005. Artificial neural networks as a useful tool to predict the risk level of Betula pollen in the air. Int. J. Biometeorol. 49, 310–316.

Catalano, M., Galatioto, F., Bell, M., Namdeo, A., Bergantino, A.S., 2016. Improving the prediction of air pollution peak episodes generated by urban transport networks. Environ. Sci. Pol. 60, 69–83.

Chaloulakou, A., Saisana, M., Spyrellis, N., 1998. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. Sci. Total Environ. 313, 1–13.

Chelani, A., Rao, C., Phadke, K., Hasan, M., 2002. Prediction of sulphur dioxide concentration using artificial neural networks. Environ. Model. Softw. 17, 161–168.

Conover, W.J., 1999. Nonparametric methods. In: Wiley, B., O'Sullivan, M. (Eds.), Practical Nonparametric Statistics. John Wiley and Sons, pp. 233–305. https://doi.org/10.1002/bimj.19730150311.

D'Amato, G., Rottem, M., Dahl, R., Blaiss, M., Ridolo, E., Cecchi, L., Rosario, N., Motala, C., Ansotegui, I., Annesi-Maesano, I., 2011. For the WAO Special Committee on Climate Change, Allergy, climate change, migration, and allergic respiratory diseases: an update for the allergist. World Allergy Organ. J. 4, 120–125.

Díaz, J., Linares, C., Tobías, A., 2007. Short term effects of pollen species on hospital admissions in the city of Madrid in terms of specific causes and age. Aerobiologia 23, 231–238.

Friedman, M., 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. J. Am. Stat. Assoc. 32, 674–701.

Galán Soldevilla, C., Cariñanos González, P., Alcázar Teno, P., Domínguez Vílches, E., 2007. Manual de Calidad y Gestión de la Red Española de Aerobiología. Universidad de CÃ³rdoba.

Gardner, M., Dorling, S., 1998. Artificial neural networks (the multilayer perceptron) a review of applications in the atmospheric sciences. Atmos. Environ. 32, 2627–2636.

Gers, F., Schmidhuber, J., Cummins, F., 2000. Learning to forget: continual prediction with LSTM. Neural Comput. 12, 2451–2471.

González, S., Díaz, J., Pajares, M., Alberdi, J., López, C., Otero, A., 2001. Relationship between atmospheric pressure and mortality in the Madrid autonumus region : a time series study. Int. J. Biometeorol. 45, 34–40.

Grivas, G., Chaloulakou, A., 2006. Artificial neural network models for prediction of pm10 hourly concentrations, in the greater area of Athens, Greece. Atmos. Environ. 40, 1216–1229.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scand. J. Stat. 6, 65–70.

Iglesias-Otero, M.A., Fernández-González, M., Rodríguez-Caride, D., Astray, G., Mejuto, J.C., Rodríguez-Rajo, F.J., 2015. A model to forecast the risk periods of Plantago pollen allergy by using ANN methodology. Aerobiologia 31, 201–211.

Kingma, D.P., Ba, J., 2019. Adam: a method for stochastic optimization. In: CoRR abs/1412.6980, URL. http://arxiv.org/abs/1412.6980.

Linares, C., Díaz, J., 2008. Impact of high temperatures on hospital admissions: comparative analysis with previous studies about mortality (Madrid). Eur. J. Pub. Health 18, 318–322.

Navares, R., Aznarte, J., 2016. What are the most important variables for poaceae airborne pollen forecasting? Sci. Total Environ. 579, 1161–1169.

Navares, R., Aznarte, J., 2016. Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features. Int. J. Biometeorol. https://doi.org/10.1007/s00484-016-1242-8.

Navares, R., Díaz, J., Linares, C., Aznarte, J., 2018. Comparing Arima and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in Madrid. Stoch. Env. Res. Risk A. 1–11. https://doi.org/10.1007/s00477-018-1519-z.

Ozkaynak, H., Qualters, B.G. nd J.R., Strosnider, H., McGeehin, M., Zenick, H., 2009. Summary and findings of the epa and cdc symposium on air pollution exposure and health. J. Expo. Sci. Environ. Epidemiol. 19, 19–29.

Querol, X., Viana, M., Moreno, T., Alastuey, A., 2012. Bases científico-técnicas para un plan nacional de mejora de la calidad del aire. Informes CSIC 1, 19–25.

Ruder, S., 2016. An overview of gradient descent optimization algorithms. In: arXiv (Ed.), CoRR abs/1609.04747. arXiv URL. http://arxiv.org/abs/1609.04747.

Rumelhart, D.E., Hinton, G.E., Ronald, R.J., 1986. Learning representations by back-propagating errors. Nature 323, 533–536.

Sabariego, S., Cuesta, P., Fernández-González, F., Pérez-Badia, R., 2012. Models for forecasting airborne cupressaceae pollen levels in Central Spain. Int. J. Biometeorol. 56, 253–258.

Schaber, J., Badeck, F.-W., 2003. Physiology-based phenology models for forest tree species in Germany. Int. J. Biometeorol. 47, 193–201.

Shaffer, J., 1986. Modified sequentially rejective multiple test procedures. J. Am. Stat. Assoc. 81, 826–831.

Sharma, S., Sharma, P., Khare, M., Kwatra, S., 2016. Statistical behavior of ozone in urban environment. Sustain. Eviron. Res. 142–148. https://doi.org/10.1016/j.serj.2016.04.006.

Silva-Palacios, I., Fernández-Rodríguez, S., Durán-Barroso, P., Tormo-Molina, R., Maya-Manzano, J., Gonzalo-Garijo, A., 2016. Temporal modelling and forecasting of the airborne pollen of cupressaceae on the southwestern Iberian peninsula. Int. J. Biometeorol. 60, 1509–1517.

Smith, M., Emberlin, J., 2006. A 30-day-ahead forecast model for grass pollen in North London, UK. Int. J. Biometeorol. 50, 233–242.

Subiza, J., Jerez, M., Jiménez, J., Narganes, M., Cabrera, M., Varela, S., Subiza, E., 1995. Allergenic pollen pollinosis in madrid. J. Allergy Clin. Immunol. 96, 15–23.

Chapter 9

Comparing ARIMA and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in Madrid

Type:	Published Article
Title:	*Comparing ARIMA and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in Madrid*
Journal:	Stochastic Environmental Research and Risk Assessment
Authors:	Ricardo Navares & Julio Díaz & Cristina Linares & José Luis Aznarte
Published:	March 2018
Impact Factor:	2.807
Quartile:	Q1
DOI	10.1007/s00477-018-1519-z

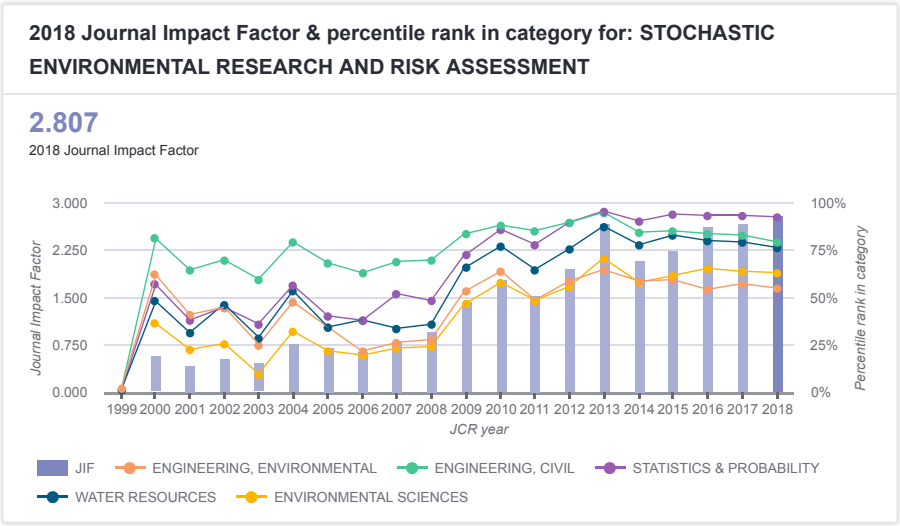


FIGURE 9.1: Impact factor Stochastic Environmental Research and Risk Assessment

Stochastic Environmental Research and Risk Assessment
https://doi.org/10.1007/s00477-018-1519-z

ORIGINAL PAPER

Comparing ARIMA and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in Madrid

Ricardo Navares[1] · Julio Díaz[2,3] · Cristina Linares[2,3] · José L. Aznarte[1,3]

Abstract
Anticipating future workloads in a hospital may be of capital importance in order to distribute resources and improve patient attention. In this paper, we tackle the problem of predicting daily hospital admissions in Madrid due to circulatory and respiratory cases based on biometeorological indicators. A range of forecasting algorithms were proposed covering four model families: ensemble methods, boosting methods, artificial neural networks and ARIMA. Experiments show how the last two obtain better results in average, demonstrating that the problem can be properly solved with both approaches. Furthermore, a recently proposed technique known as stacked generalization was also used to dynamically combine the predictions from the four models, finally improving the performance with respect to the individual models.

Keywords Forecasting · Emergency hospital admissions · ARIMA · Neural networks · Random forests · Gradient boosting machines · Stacked generalization

1 Introduction

During the Seventies George E. P. Box y Gwilym Jenkins proposed the autoregressive integrated moving average models (ARIMA) to predict and analize time series related to economic variables (Box and Jenkins 1976). Since then, this sort of stochastic models has been applied to forecast the evolution of time series in different fields such as environmental atmospheric (Díaz et al. 1999; Kumar and Goyal 2011; Olsen et al. 2016) or biological pollution (Rodríguez-Rajo et al. 2006). Not only are ARIMA models used in forecasting, but also in determining through statistical significance the influence of independent variables on the behavior of certain dependent variable. Given the epidemiological meaning of the value of the estimators, those which are statistically significant for the ARIMA model, they are used to quantify the impact of independent environmental variables on health indicators. Thus, ARIMA models have been used to evaluate the influence of atmospheric pollution on patients morbility (Díaz et al. 1999), the relation between biotic factors and hospital admissions (Díaz et al. 2007), meteorological event such as heat and cold waves and daily mortality (Alberdi et al. 1998; Díaz et al. 2002, 2005; Montero et al. 2012; Linares et al. 2016; Roldán et al. 2016) or daily emergency admissions (Linares and Díaz 2008).

One of the biggest contributions of ARIMA models to the clinical field is the forecast of the behavior of different pathologies, concretely those of infectious nature, in places where it is critical to know in advance the development of certain diseases in order to apply preventive measures and plan medical resources (Anwar et al. 2016; Kumar et al. 2014; Nsoesie et al. 2014; Luque et al. 2009). Also, economic implications motivated the use of ARIMA models to predict sudden emergency admission and consequently optimize resources, easing clinical institutions management (Zhu et al.

✉ José L. Aznarte
jlaznarte@dia.uned.es

[1] Department of Artificial Intelligence, Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal, 16, 28040 Madrid, Spain

[2] Department of Epidemiology and Biostatistic, National School of Public Health, Carlos III Institute of Health, Avda. Monforte de Lemos, 5, 28029 Madrid, Spain

[3] Instituto Mixto de Investigación ENS-UNED (IMIENS), Madrid, Spain

Published online: 03 March 2018

2015; Dominak et al. 2015; McWilliams et al. 2014; Abraham et al. 2009; Earnest et al. 2005; Díaz et al. 2001).

Compared to traditional time series analysis techniques, computational intelligence techniques have been gaining popularity as effective approaches for predicting environmental and biometeorological conditions (Castellano-Méndez et al. 2005; Aznarte et al. 2007; Navares and Aznarte 2016a, b, 2017). Among these techniques we propose to use Random Forests, Gradient Boosting Machines and Artificial Neural Networks trying to cover three main techniques in the computational intelligence field such as ensemble learning, boosting algorithms and bio-inspired learning. Thus, the scope of this study is double: on the one hand forecasting the daily number of emergency hospital admissions due to circulatory and respiratory causes for the target group of patients older than 65 years. On the other, compare different approaches and try to establish a framework in order to exploit the advantages of each one.

2 Materials and methods

2.1 Data description

2.1.1 Target variables

Target variables consist on daily hospital admissions to emergency services due to either circulatory or respiratory cases. The target group is patients older than 65 years recorded in hospitals across the region of Madrid from the first of January 2005 to the eleventh of June 2010 (Fig. 1).

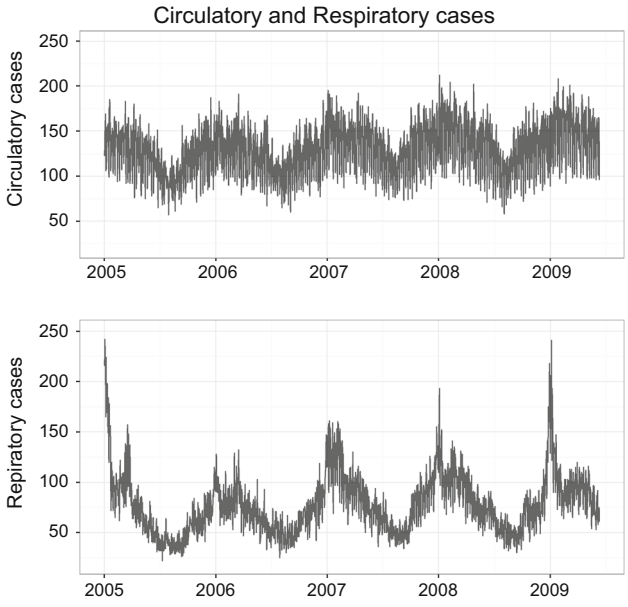


Fig. 1 Emergency hospital admissions recorded in Madrid due to circulatory (top chart) and respiratory (bottom chart) cases in Madrid

In order to preserve the confidentiality of the data, daily hospital admissions were provided as the total aggregated emergency cases (circulatory and respiratory) in Madrid.

Figure 1 shows that both time series have a significant seasonality. Although it can be clearly seen that winter months concentrate the larger number of admission, respiratory cases present bigger peaks and less variablity during the rest of the year. Conversly, circulatory-related admissions show more variablility around the seasonal component. Several factors can explain this seasonal behavior. Some of them relate to the intrinsic evolution of the diseases and the also seasonal behavior of the atmospheric conditions (cold waves, pollution, pollen...) which will be analysed.

2.1.2 Independent variables

Chemical air pollutants Daily mean concentrations ($\mu g/m^3$) of particulate matter less than 2.5 and 10 μm in diameter (PM2.5 and PM10), surface ozone (O_3) and nitrogen dioxide (NO_2). All measurements were made using the gravimetric method or an equivalent method (β-attenuation). Hourly data was aggregated to obtain daily mean levels of chemical air pollutants. We used the data supplied by the Madrid Municipal Air Quality Monitoring Grid (http://www.mambiente.munimadrid.es/), a network that consists of 27 urban background stations spread across the city, which capture chemical air pollutants in real time (Fig. 2). No specific validation was performed within the project to assess the representativeness of spatial variability

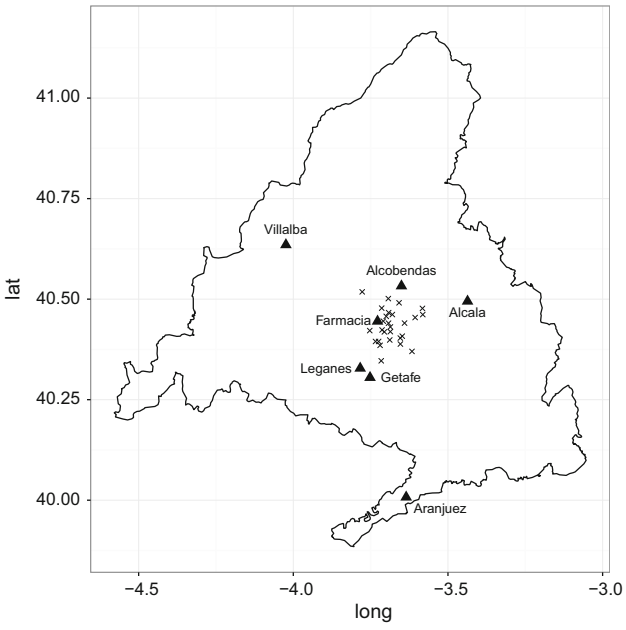


Fig. 2 Location of sources of data: weather and pollen stations are represented by triangles and stations from the Madrid's air quality monitoring grid are represented by crosses

in air pollutants; an ecological exposure was used, as is common in most time-series studies. Since the number of hospital admissions were provided at a city aggregated level, the average of the concentrations for all locations was used per chemical pollutant. In case of missing data at one location, that location was removed in the averaging process.

Studies prior to 2000 show a relevant relationship between Sulfur dioxide (SO_2) and mortality and morbidity in Madrid (Díaz et al. 1999). However, from year 2000, SO_2 concentrations have damatically decreased due to the progressive transition from coal-powered heating to natural gas. As a result, this pollutant has lost influence in hospital admissions due to respiratory and circulatory cases (Díaz et al. 2007). Consequently, SO_2 was not taken into consideration in this study.

With respect to Carbon monoxide (CO), it has little impact on health in Madrid. In fact, it is not contemplated among the pollutants considered to measure air quality in Spain (Quero et al. 2012). Thus, CO was also discarded for this study.

Biotic factors Pollen observations correspond to daily grains per cubic meter of Poaceae and Plantago pollen registered across the region of Madrid: Alcalá de Henares, Alcobendas, Aranjuez, Complutense University of Madrid (Farmacia), Getafe, Leganés and Villalba (Fig. 2). Pollen counts followed the standard methodology of the Spanish Aerobiological Network (Galán Soldevilla et al. 2007) and were provided by Red Palinológica de la Comunidad de Madrid. The nature of pollen time series (very low concentrations across the year with sudden high peaks during the pollination season) demands special attention to the missing data during the critical months of February, March and April. The weighted interpolation proposed by Navares and Aznarte (2016a) was used in order to deal with the missing data points, which account for less than 3% of the total.

2.1.3 Control variables

Meteorological variables Meteorological observations were obtained from sensors placed in the close surroundings of the pollen stations, consisting of daily maximum ($T_{\max}$)and minimum ($T_{\min}$) temperature in Celsius degrees, pressure in mbar and degree of humidity in percentage. Other studies in the region of Madrid show that wind speed and direction are not significant in this setup (Alberdi et al. 1998; Díaz et al. 2001, 2007). Consequently, they were not taken into consideration for the model. Very few missing data points appear in meteorological time series (less than 1% of the total). Given the sparsity of the presence of these missing data points, they were estimated through linear interpolation using the precedent and subsequent observation.

The relationship establish between maximum temperature and morbility in Madrid was known that has an "V" shape (Alberdi et al. 1998). To control for the possible effect of temperature on the dependent variable considered, we defined the variable: T_{heat} as follows, considering the threshold of 34 °C above which a heat wave is defined in Madrid (Díaz et al. 2015):

$$T_{\text{heat}}(i) = \begin{cases} 0 & if\, T_{\max}(i) \leq 34\,^{\circ}\text{C} \\ T_{\max}(i) - 34 & if\, T_{\max}(i) > 34\,^{\circ}\text{C} \end{cases}, \quad (1)$$

where T_{max} is the daily maximum temperature. Similarly, a cold wave is defined in Madrid when daily minimum temperature is below − 2 °C (Carmona et al. 2016). In order to take into account the effect of low temperatures T_{cold} is defined as follows:

$$T_{\text{cold}}(i) = \begin{cases} 0 & if\, T_{\min}(i) \geq -2\,^{\circ}\text{C} \\ -T_{\min}(i) - 2 & if\, T_{\min}(i) < -2\,^{\circ}\text{C} \end{cases}, \quad (2)$$

Additionally, in order to consider a synoptic scale of meteorological situations changes a variable $diff(P) = P_t - P_{t-1}$ is defined. Being P_t the daily average pressure at time t and P_{t-1} previous day daily pressure.

According to previous studies on the effect of atmospheric pollution on morbility in Madrid (Díaz et al. 1999), there is a lineal relation between hospital admissions and the primary pollutants NO_2, PM10 y PM 2.5 and quadratic in the case of ozone, with a concentration of 45 $\mu g/m^3$ as the minimum value of this parabola (Díaz et al. 2001). Thus, an additional variable O_{3a} was defined to establish this relation as follows:

$$O_{3a}(i) = \begin{cases} 0 & if\, O_3(i) \leq 45\,\mu g/m^3 \\ O_3 - 45\,\mu g/m^3 & if\, O_3(i) > 45\,\mu g/m^3 \end{cases}, \quad (3)$$

The effects of meteorological and atmospheric conditions on hospital admissions can be either immediate or might take one or several days. Hence, including lagged variables derived from all the set of independent variables (chemical air pollutants, biotic factors and meteorological variables) is interesting to model these situations.

In the related literature there are several studies for the region of Madrid that suggest the use of a 5-days lag for the pollutants NO_2, PM10, PM2.5 (Jiménez et al. 2010); 8 days for O_3 (Díaz et al. 1999); 4 days for heat temperatures (Díaz et al. 2015); 14 for cold waves temperatures (Carmona et al. 2016); 14 days for relative humidity (Carmona et al. 2016; Díaz et al. 2015) and 5 for *diff*(*P*) (González et al. 2001).

2.2 Methodologies

Arima The ARIMA forecasting equation for a stationary time series is a linear (i.e., regression-type) equation in which the predictors consist of lags of the dependent variable and/or lags of the forecast errors. That is:

Predicted value of Y equals a constant (μ) and/or a weighted sum of one or more recent values of Y $(Y_{t-1},\ldots,Y_{t-p})$ and a weighted sum of recent values of the errors $(e_{t-1},\ldots,e_{t-q})$. In terms of y, the general forecasting equation is:

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \cdots - \theta_p e_{t-p}, \tag{4}$$

where ϕ_i is the coefficient of the autoregressive (AR) term i and θ_i represents the coefficient of the moving average (MA) term i.

Apart from the corresponding lags of the series p (Y_{t-p}), the errors (e) and its lags (e_{t-q}), exogenous variables $(X,\ldots,Z)$, which represent the environmental independent variables, were include along with their corresponding lags up to $t-s$ and $t-m$, resulting in:

$$\begin{aligned}\hat{y}_t = {} & \mu + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} \\ & - \theta_1 e_{t-1} - \cdots - \theta_p e_{t-p} \\ & + \beta_0 X_t + \beta_1 X_{t-1} + \cdots + \beta_s X_{t-s} + \cdots \\ & + \gamma_0 Z_t + \gamma_1 Z_{t-1} + \cdots + \gamma_m Z_{t-m}\end{aligned} \tag{5}$$

The value of the estimator $\beta_0,\beta_1\ldots\gamma_0,\gamma_1\ldots$ of the variables that are significant at $p<0.05$ (p value proportionated by SPSS v15) indicating increased Y to an increment by one unit of the each independent variable $(X,\ldots,Z)$ respectively.

The model's goodness-of-fit was obtained by analysis of residuals (AIC, BIC, ACF, Box-Ljung).

Random forest Ensemble learning has been gaining attention during the last decade and it consists in leveraging the performance of many weak learners by combining them to form a strong learner. Breiman (1996a) proposed a method called *bagging* also known as bootstrap aggregating. The procedure fits one model using each bootstrap sample and combines them by averaging. Breiman (2001) adds an extra layer of randomness to *bagging* by using decision trees in order to construct a collection of trees (forest) with controlled variance which improves stability and accuracy. Thus, *bagging* intervenes at two levels, in data selection and subsequently in variable selection. The combination of random trees by averaging provides robustness against *overfitting* as well as against the presence of outliers.

Gradient boosting machines In addition to average the combination of multiple learners, another popular ensemble technique is *boosting*. The principle behind is starting with a weak learner and turn it into a strong learner. This process is also known as *additive training*. Proposed by Friedman (2001) GBM adds sequentially new models (trees) to the ensemble, which is represented by an error function of the previous iteration fitted model. Thus, each new tree is trained with respect to the error of the whole ensemble so far. In the regression problem the error function is the classic square error (SE) which conforms the objective function to optimize.

An important concern in computational intelligence is the generalization capabilities of the models which might suffer from a non proper learning scheme and, resulting in overfitting. In order to mitigate the effects of *overfitting*, Friedman (2001) proposes a technique known as *shrinkage* to control the complexity of the model. *Shrinkage* is a common regularization approach which shrinks regression coefficients to zero and consequently, reduces the impact of unstable coefficients. In the context of GBM, *shrinkage* penalizes (reduces) the importance of each tree at each consecutive step. Hence, the final objective consists of two terms, a training loss function represented by SE and the regularization which measures the complexity of the model.

Artificial neural networks ANNs are a tool for modelling nonlinear processes based on the information collected by a vector named input layer, through which the information is propagated layer by layer establishing the relations between the inputs and the final layer called output layer. Intermediate or hidden layers consist of one or more units called neurons which are interconnected to the neurons of the previous and subsequent layers. The number of hidden layers and the number of neurons of each one define the *topology* of the network.

Each neuron generates an excitatory answer to signals received through an activation function which can be selected among the different functions available but, following recommendations from the literature the sigmoidal activation function was chosen (Bishop 1995; Haykin 1999).

The learning of the network is based on obtaining the relationship between the input and the output layer by comparing, via root square mean error (RMSE), network outputs with the actual values through the well-known *backpropagation* algorithm (Rumelhart et al. 1986)

The aim is to find the network *topology* which minimizes the error. This procedure is based on a trial and error approach which, starting from a simple network of one hidden layer with few neurons, consists on increasing the capacity of the network (sequentially incrementing the number of neurons in a hidden layer as well as the number of layers) to optimize the results.

Stacked generalization Proposed by Wolpert (1992), stacked generalization or stacking is an ensemble technique that uses a new model to learn how best combine the predictions from two or more models with the aim of reducing error generalization. Opposed to more traditional approaches of ensemble learning such as voting or averaging, as in the case of RF, which are winner-takes-all ways of combining (Wolpert 1992), using a meta-learner to ensemble allows to identify the circumstances under which the pooled predictions shall gain or lose weight in the final forecast.

The idea behind of stacking is splitting the training set in two subsets train_a and train_b. A first stage trains the pool of selected models on train_a to create predictions for train_b and repeat using train_b for training to generate train_a predictions. As a final step of this first stage the pool of models are trained over the full training set to create predictions for the test set. The second stage consists of training the meta-learner using the training set, which contains the predictions of the models from stage one, and create the final predictions for the test set.

In order to check the performance of the forecasts the common quadratic scoring rule root mean squared error (RMSE), defined in (6), was used to measure the average magnitude of the error. The goodness of fit of the set of predictions to the observed values is represented by R^2 along with RMSE.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{6}$$

where y_i is the observed ith data point, $\hat{y}_i$ the predicted and n the total number of data points in the test set. Also, as it is intended to compare the models given two time series of different nature (Fig. 1), we propose to use a normalized error measure know as mean absolute percentage error defined as:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{7}$$

where y_i is the observed ith data point, $\hat{y}_i$ the predicted and n the total number of data points in the test set.

It is important to note that our objective is to compare how different forecasting approaches (not models) perform when applied to the same problem. Therefore, although they all start with the same dataset, the data selection process is not necessarily the same in all cases.

2.3 Experimental design

The aim is to provide the most suitable technique, in terms of accuracy, to forecast the one day-ahead cases of admissions due to circulatory and respiratory diseases. A first approach is to test the accuracy of each proposed model when applied in isolation to check which one is more suitable for each serie. In order to do so, the full set is split in a training set, which covers from 16-01-2005 to 11-06-2009, and leaving the last year as a test set for prediction. The best performer among the models will be used as a benchmark for the subsequent experiments. Secondly, this study focuses in researching different stacking configurations and models to minimize the generalization error (also known as out-of-sample error) of the previous models tested.

With regards to the computational intelligence models, a common problem is the parameterization, as the performance of each model heavily relies on this step. For the tree-based algorithms a grid search procedure was performed using a 10-fold cross validation on the training set. Given a set of values for each parameter, grid search explores each combination and selects the one which minimizes the prediction error on the validation. The same set of input variables was used for all the computational intelligence models. This set tries to represent the short-term interactions by including lags up to seven days of the variables, as well as the long-term influence by including the cumulative sum of the fortnight before the forecast time (Table 1). Additionally, in order to capture seasonality, the variables Julian day number (doy), the month of the year ($\{\text{Mon}, \ldots, \text{Sun}\} \rightarrow \{1, \ldots, 7\}$) and the season ($\{\text{Winter}, \text{Summer}, \text{Spring}, \text{Autumn}\} \rightarrow \{1, 2, 3, 4\}$) were also included.

Artificial neural networks add an additional research step which involves finding the right architecture or topology of the network for the study problem. Using the same validation scheme as before, the capacity of the network was incrementally increased until the validation error increases with respect to the previous configuration tested. It is reasonable to start with the simplest network of one hidden layer with one hidden unit and increment the number of units in the layer, up to the boundary $N_h = 0.5 * N_i$. where the number of hidden units N_h is the half of the number of inputs N_i. Once the convenient number of units for the first layer is selected, an extra layer is included and the process is repeated.

To train the network the backpropagation algorithm (Rumelhart et al. 1986) was used on the sigmoid activation function. As backpropagation algorithm applies the gradient descent optimization, it might get stuck in local minima leading to sub-optimal solutions. To avoid this situation a momentum term was used in the objective function which increases step size towards the minimum by trying to jump from local minima leading the weight updates as,

$$\Delta\omega_i(t+1) = -\eta\frac{\partial E}{\partial \omega_i} + \alpha\Delta\omega_i(t), \tag{8}$$

Table 1 Input variables for the computational intelligence models

	t − 1	t − 2	t − 3	t − 4	t − 5	t − 6	t − 7	$\sum_{t-1}^{t-16}$
Circulatory cases	✓	✓	✓	✓	✓	✓	✓	✓
Respiratory cases	✓	✓	✓	✓	✓	✓	✓	✓
Air pollutants								
NO	✓	✓	✓	✓	✓	✓	✓	✓
NO_2	✓	✓	✓	✓	✓	✓	✓	✓
O_3	✓	✓	✓	✓	✓	✓	✓	✓
PM10	✓	✓	✓	✓	✓	✓	✓	✓
PM2.5	✓	✓	✓	✓	✓	✓	✓	✓
Biotic factors								
Poaceae	✓	✓	✓	✓	✓	✓	✓	✓
Plantago	✓	✓	✓	✓	✓	✓	✓	✓
Meteorological								
Pressure	✓	✓	✓	✓	✓	✓	✓	✓
Rainfall	✓	✓	✓	✓	✓	✓	✓	✓
Relative Humidity	✓	✓	✓	✓	✓	✓	✓	✓
Temperature	✓	✓	✓	✓	✓	✓	✓	✓

where E is the error function represented in this study by RMSE, η the learning rate and α the momentum. After the preliminary research, we found that a network with 1 hidden layer of 5 units and a $\eta = 0.005$ and $\alpha = 0.6$ shows as the best candidate configuration, in terms of accuracy, for this problem.

As mentioned before, once the models are set up, we test their forecast accuracy by training them on the full training set which consists on observations from the 16th of January 2005 to the 11th of June 2009, leaving the last 365 observations for testing. Once performances are compared, the principal idea behind stacking is to find a model and setup which minimizes the generalization error assuming the meta-learner (or algorithm used to combine previous models) might be able to select and apply the most convenient of the previous base models to each present situation in the series, in terms of environmental conditions.

A simple approach for stacking was taken in order to reduce full system complexity. The original training set (16-01-2005 to 11-06-2009) is split in two subsets consisting on leaving out the last 365 observations. Thus, the base models (ARIMA, RF, GBM and ANN) are trained on the new training set (16-01-2005 to 11-06-2008) to create predictions for the left out period (12-06-2008 to 11-06-2009), henceforth *stack set*. On the stack set, base models predictions are combined though a meta-learner which is trained using a 10-fold cross-validation for parametrization. Several candidates selected from the base models were tested as meta-learners, being GBM the one with better results in the cross-validation.

As a final step, the base models are trained using the full training set to create predictions for the test set. These predictions are used by the meta-learner (GBM) to generate the final forecast for the test set. In order to help the model to decide under which circumstances which model to apply, three dummy variables were included to represent the day of the year, the day of the week and the season. This process is summarize in Fig. 3.

3 Results

The study is based on daily morbility cases, due to respiratory and circulatory diseases, recorded from 01-01-2005 to 11-06-2010 in Madrid. As shown in Fig. 1, the characteristics of the time series differ, having the number of respiratory cases a more volatile behavior than the circulatory, which shows a more constant pattern. Although, both show a strong seasonal pattern.

The first experiment consists on testing each of the proposed models (ARIMA, RF, GBM and ANN) in isolation by training them on the full training set (01-01-2005 to 11-06-2009) and leaving the last 365 observations for testing. The explanatory variables found for the ARIMA model in the circulatory and respiratory time series are shown in Tables 2 and 3, respectively. In Table 4 the results for each model for the respiratory cases are shown. Tree-based models perform poorly when compared to ARIMA and the neural network, obtaining higher RMSE, being the Random Forest the worst performer. In this case, ARIMA overperforms the other models managing to explain the 91% of the variance of the respiratory cases.

Table 5 shows that the artificial neural network obtains the best results for circulatory cases. ANN obtains a RMSE

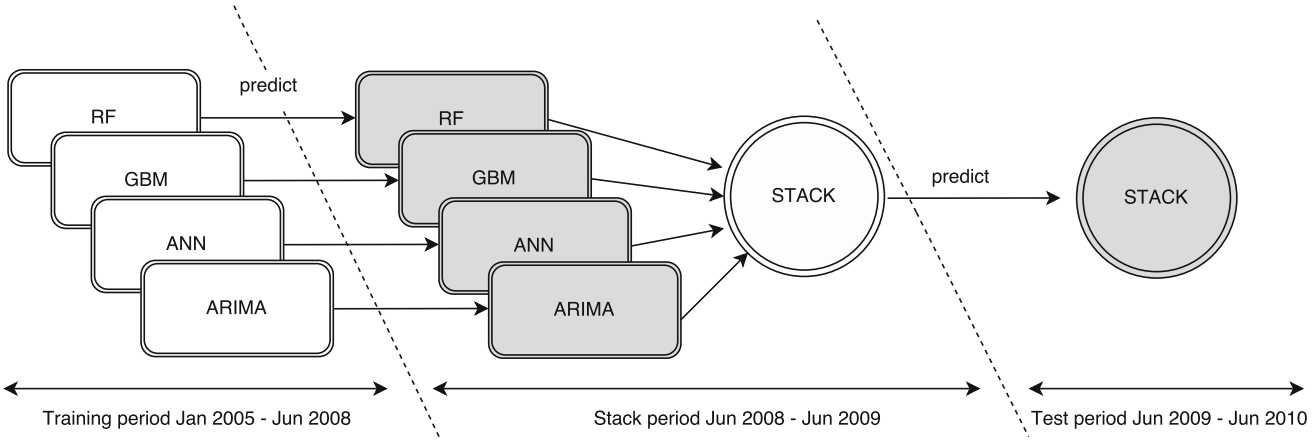


Fig. 3 Diagram of the architecture of the stack. Shapes shaded in gray represent forecast for the correspondent period, non-shaded shapes represent training processes

Table 2 Explanatory variables obtained by the ARIMA model for the circulatory-related admissions

	Estimations	SE
Non seasonal		
AR1	− 0.55	0.15
AR2	0.18	0.03
MA1	− 0.70	0.15
Seasonal*		
Seasonal AR1	0.12	0.03
Reg. coefficients		
N1(trend)	0.01	0.00
Sine 365 days	11.01	0.89
Cosine 365 days	10.24	1.30
Sine 180 days	− 2.95	0.91
Monday	39.54	1.75
Tuesday	42.91	1.86
Wednesday	43.89	1.93
Thursday	40.06	1.94
Friday	46.94	1.88
Saturday	8.50	1.73
NO_2	0.18	0.04
Rel. Humidity	0.09	0.04
LAGS (*diff*(*P*), 1)	− 0.37	0.14
Constant	74.09	3.03

* Season = 7 days

of 13.69 compared to the 17.11 of the ARIMA with a mean absolute percentage error of 8.39 and 10.66% respectively. Again, the ensemble tree-based models poorly perform compared to the other two approaches.

As a second part, we split the full set in three subsets, a new training set consisting on the period from 01-01-2005 to 11-06-2008 to train the models, the subsequent 365 observations to stack the predictions of the base models through GBM and again, the last 365 observations to test. Tables 4 and 5 show the benefit of stacking, which reduces previous best performer error and consequently, increases R2 for respiratory cases.

4 Discussion

We have seen in Sect. 3 the benefits, in terms of error reduction, of using the stack prediction technique.

Figure 4 shows a comparison of the best performer (ARIMA) versus the stacked predictor on the respiratory test set and the monthly distribution of errors. In general, the distribution of the errors is similar for both approaches, being the stacked errors slightly more concentrated around the median error value as the boxes are slightly shorter. It is noticeable the special characteristics of this series in the months of December and January where the stack predictor has slightly more difficulties in capturing the sudden peaks compared to the ARIMA. Since the stacking is based on assigning weights to each algorithm, in this case it might be lowering the influence of the ARIMA model during those sudden peak periods because this improves the prediction of the entire test period.

On the other hand, when predicting the time series of circulatory cases, the stack technique manages to reduce the RMSE with respect to the ANN which was the best non-aggregated performer. Figure 5 shows the comparison of the monthly distribution of error for both algorithms. This time, it can be clearly seen that in general the distribution of errors for the stacked predictor show higher concentrations around the median since the boxes are shorter.

Table 3 Explanatory variables obtained by the ARIMA model for the respiratory-related admissions

	Estimations	SE
Non seasonal		
AR1	− 0.03	0.04
AR2	0.92	0.03
MA1	− 0.24	0.04
MA2	0.73	0.03
Seasonal*		
Seasonal AR1	− 0.06	0.09
Seasonal AR2	0.85	0.09
Seasonal MA1	− 0.15	0.11
Seasonal MA2	0.75	0.10
Reg. coefficients		
N1 (trend)	0.02	0.01
Sine 365 days	16.27	3.28
Cosine 365 days	24.62	3.29
Cosine 90 days	4.41	1.55
Monday	18.58	2.32
Tuesday	16.74	2.35
Wednesday	14.93	2.37
Thursday	14.84	2.37
Friday	21.99	2.36
Saturday	5.35	2.32
LAGS (PM2.5,5)	0.15	0.07
LAGS (PM10,1)	0.10	0.03
LAGS (NO2,4)	0.43	0.17
LAGS (T_{cold},4)	5.97	2.99
LAGS (T_{cold},14)	7.04	2.98
LAGS (Rel. Hum., 11)	− 0.09	0.03
diff(*P*)	− 0.27	0.11
LAGS (*diff*(*P*), 2)	− 0.22	0.11
Constant	48.06	6.62

* Season = 7 days

Table 4 Forecast results for respiratory-related admissions

Model	RMSE (%)	MAPE	R2
RF	22.64	18.61	0.74
GBM	16.87	14.45	0.85
ANN	15.06	13.44	0.89
ARIMA	13.15	12.65	0.91
STACK	**13.04**	**12.32**	**0.92**

Bold values indicate the best result for every error measure

If we compare the results on both series, we can see how ARIMA obtains a mean absolute percentage error (MAPE) of 12.65% for the respiratory time series and a 10.66% for the circulatory, while the ANN achieves a 13.44 and a 8.39% respectively. A similar pattern is found in the stacked predictor. This difference in error can be explained by the nature of the respiratory time series, which shows higher changes in standard deviation between the spring-summer season and the autumn-winter. Admissions due to circulatory cases show an average deviation of 27 patients regardless the seasonality while the deviation of the respiratory cases increases to an average of 29 patients during winter season from 15 patients during the summer. Additionally, ARIMA models with exogenous variables (5) require the prediction of these independent variables at time t in order to estimate the dependent variable at the same time. This situation leads to an error propagation from the models used to predict the exogenous variables and consequently, the prediction error of the target variable increases.

Table 5 Forecast results for circulatory-related admissions

Model	RMSE (%)	MAPE	R2
RF	22.96	14.17	0.77
GBM	18.62	10.81	0.83
ANN	13.69	8.39	**0.90**
ARIMA	17.11	10.66	0.84
STACK	**13.24**	**8.09**	**0.90**

Bold values indicate the best result for every error measure

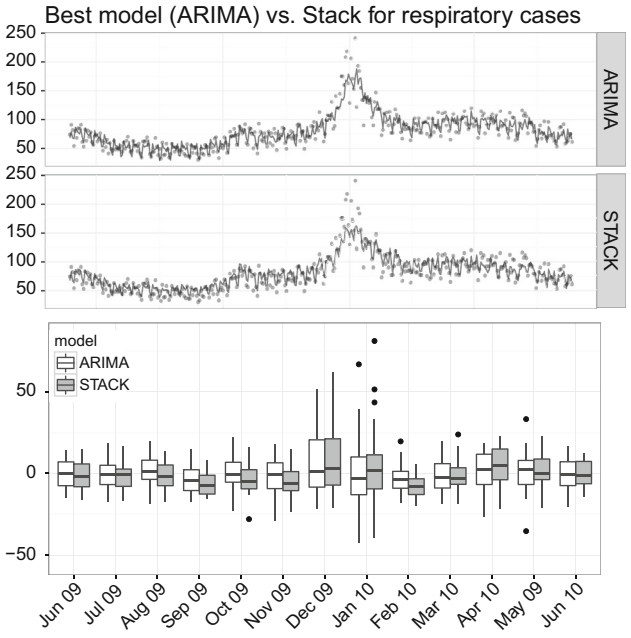


Fig. 4 Observed (dots) versus Predicted of best model compared to the stack for respiratory cases. Bottom chart shows the monthly boxplots of the error per month

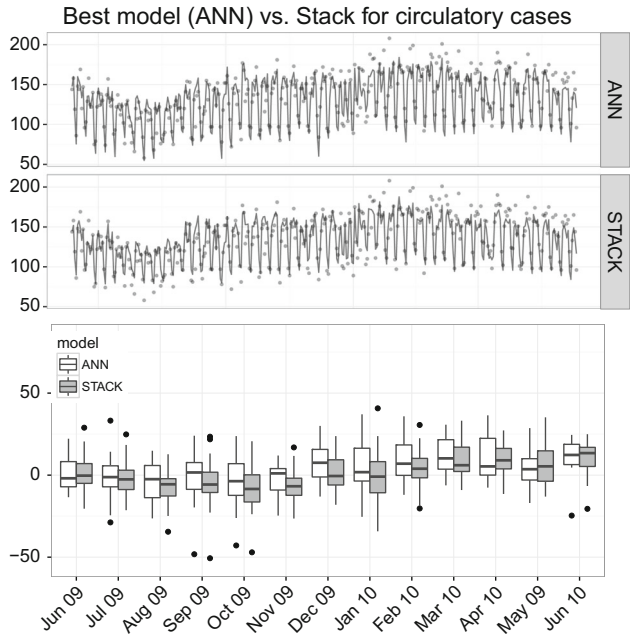


Fig. 5 Observed (dots) versus Predicted of best model compared to the stack for respiratory cases. Bottom chart shows the monthly boxplots of the error per month

However, in both cases, the ARIMA model and a simple ANN with one hidden layer clearly obtain better results when compared to more complex approaches. This suggests that the problem might be mostly linear in nature. Notice that a single-layer ANN captures linear relations among the inputs, its output being the sum of the weighted inputs. Furthermore, this possible linearity of the series does not seem to be parallel to the axes since both, RF and GBM, divide the feature space into hyper-rectangles. Thus, both algorithms decision boundaries are parallel to the axes requiring a higher number of trees to adapt the diagonal shape. As a consequence, this situation requires more complex tree-based algorithms which not always obtain a solution as optimal as models of other nature such as the ARIMA and the ANN. Among computational intelligence models, tree-based approaches provide an advantage with respect to ANNs which is that importance for the different variables can be easily obtained. GBM and RF were selected to test the bias-variance trade off. Bias-variance decomposition is a way to analyse algorithm performance error which is explained by the difference of the observed value and the predicted (bias) and the variability of the prediction given a data point (variance). GBM reduces error by reducing the bias (also to some extent variance by aggregating the output of many models) while RF reduces variance as trees are made uncorrelated to maximize the decrease in variance. RF is practically tuning-free, saving research time while GBM have few hyper-parameters to tune making this algorithm performance highly dependent on tuning (Caruana and Niculescu-Mizil 2006).

As opposed to ARIMA models, ANNs are not constrained by any predefined mathematical relationship between the dependent and independent variables as they have the ability to capture these relationships during the training phase (White 1989). ANNs are known as a black-box approach meaning that, although they might provide good results, they are not readily interpretable and thus it is harder to use them for diagnose as opposed to ARIMA models.

To sum up, we have seen the benefits of stacking, which performs as expected by minimizing the generalization error (or out-of-sample error). Stacking increases the percentage of variance explained by the forecast in both series capturing and combining the best predictive capabilities of each model to configure a better solution. Although this improvement is marginal in the problems presented (around 1% in the respiratory cases and 3% in the circulatory cases with respect to performance of the best model), this technique is a useful tool for automatic model selection under different circumstances when accuracy is the priority in detriment of model complexity. This marginal improvement suggests there are strong similarities between the performance of the best algorithms (ANN and ARIMA) since the biggest improvement occurs when stacking together more dissimilar predictors (Breiman 1996b). Providing additional variables, such as seasonal ones, might help the stacking method (which is a tree-based algorithm) to adapt the importance of each underlying algorithm according to different situations. For instance, during the summer, one of the algorithms participating in the stacking might gain importance with respect to another, while the situation might reverse the rest of the year. Table 6 shows the relative importance given to each of the proposed models along with the dummy variables which represent the day of the year (*doy*), the day of the week (*dow*) and the season (*sea*).

Table 6 Relative importance of each variable in the Stack

Model	Respiratory model (%)	Circulatory model (%)
RF	13.29	12.44
GBM	15.88	17.44
ANN	17.86	24.47
ARIMA	21.51	20.72
doy	23.88	14.26
dow	5.13	9.53
sea	2.45	1.14

doy : Julian Day Number, $dow : \{Mon, \ldots, Sun\} \rightarrow \{1, \ldots, 7\}$, $sea : \{Winter, Summer, Spring, Autumn\} \rightarrow \{1, 2, 3, 4\}$

In that table, it can be seen how the stacked methods favor the models with lower error in the corresponding case, having ARIMA a relative importance of 21.51% for the respiratory time series and the ANN 24.47% for the circulatory time series. The day of the year (*doy*) obtains high relevance when deciding which model should be applied under which situation. This is specially true in the case of the respiratory series, which is coherent with the fact that it has a marked seasonal pattern. Knowing the day of the year might allow the stacked predictor to choose which model to use, especially in the peak winter season.

5 Conclusions

In this study we tackled the problem of forecasting the number of daily hospital admissions to emergency services due to circulatory and respiratory cases, based on air quality indicators. We conducted a comparison of the forecast accuracy of four predictive models of different nature on these two problems. Among these models, we tested two tree-based approaches which represent ensemble (random forests) and boosting methods (generalized boosted models). In addition to the tree-based models, we also tested artificial neural networks and, in order to allow for a comparison with a classic approach, the well-known traditional ARIMA model.

Amongst all the considered models, we found that the ARIMA and the ANN overperform, in terms of accuracy, random forests and gradient boosting machines on both time series.

In addition, we also proposed an aggregative technique which recently has become popular known as stacked generalization. We found that this technique, which is based on dynamically combining the prediction of other models, performs better when used to combine the forecasts of the four aforementioned models.

The results show that it is possible to forecast the respiratory and circulatory emergency hospital admissions by using air quality indicators, and that computational intelligence methods are suited for the task. However, there is still room for improvement by, for example testing other different models, different cross-validation techniques (randomly dividing the dataset in chunks, instead of validating against a single year) or including other exogenous variables into the stack, which might help the algorithm to better decide when to use one of the models or other.

Acknowledgements This work has been partially funded by Ministerio de Economía y Competitividad, Gobierno de España, through a *Ramón y Cajal* Grant (RYC-2012-11984) and by Instituto Mixto de Investigación ENS-UNED (IMIENS) through a research project awarded to Dr. Aznarte, Dr. Linares and Dr. Díaz in its 2017 call.

References

Abraham G, Byrnes G, Bain C (2009) Short-term forecasting of emergency inpatient flow. Inf Technol Biomed 13:380–388

Alberdi JC, Díaz J, Montero JC, Mirón IJ (1998) Daily mortality in madrid community (spain) 1986–1991: relationship with atmospheric variables. Eur J Epidemiol 14:571–578

Anwar M, Lewnard J, Parikh S, Pitzer V (2016) Time series analysis of malaria in afghanistan: using arima models to predict future trends in incidence. Malar J 15:566

Aznarte JL, Benítez Sánchez JM, Lugilde DN, de Linares Fernández C, de la Guardia CD, Sánchez FA (2007) Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models. Exp Syst Appl 32(4):1218–1225

Bishop CM (1995) Neural networks for pattern recognition. Oxford University Press, Oxford

Box GEP, Jenkins GM (1976) Time series analysis: forecasting and control. Holden-Day, South Windsor

Breiman L (1996a) Bagging predictiors. Mach Learn 25:123–140

Breiman L (1996b) Stacked regressions. Mach Learn 24:49–64

Breiman L (2001) Random forest. Mach Learn 45:5–32

Carmona R, Díaz J, Mirón J, Ortiz C, León I, Linares C (2016) Geographical variation in relative risks associated with cold waves in spain: the need for a cold wave prevention plan. Environ Int 88:103–111

Caruana R, Niculescu-Mizil A (2006) An empirical comparison of supervised learning algorithms. In: Proceedings of the 23rd international conference on machine learning, ICML'06. ACM, New York, pp 161–168, https://doi.org/10.1145/1143844.1143865,

Castellano-Méndez M, Aira MJ, Iglesias I, Jato V, González-Manteiga W (2005) Artificial neural networks as a useful tool to predict the risk level of Betula pollen in the air. Int J Biometeorol 49:310–316

Díaz J, García R, Ribera P, Alberdi JC, Hernández E, Pajares MS (1999) Modeling of air pollution and its relationship with mortality and morbidity in madrid (spain). Int Arch Occup Environ Health 72:366–376

Díaz J, Alberdi JC, Pajares MS, López R, López C, Otero A (2001) A model for forecasting emergency hospital admissions: effect of environmental variables. J Environ Health 64:9–15

Díaz J, López C, Jordán A, Alberdi J, García R, Hernández E, Otero A (2002) Heat waves in madrid, 1986–1997: effects on the health of the elderly. Int Arch Occup Environ Health 75:163–170

Díaz J, García R, López C, Linares C (2005) Mortality impact of extreme winter temperatures. Int J Biometeorol 49:179–183

Díaz J, Linares C, Tobías A (2007) Short term effects of pollen species on hospital admissions in the city of madrid in terms of specific causes and age. Aerobiologia 23:231–238

Díaz J, Carmona R, Mirón J, Ortiz C, León I, Linares C (2015) Geographical variation in relative risks associated with heat: update of spains heat wave prevention plan. Environ Int 85:273–283

Dominak M, Swiecicki L, Rybakowski J (2015) Psychiatric hospitalizations for affective disorders in warsaw, poland: effect of season and intensity of sunlight. Phychiatry Res 229:289–294

Earnest A, Chen M, Ng D, Sin L (2005) Using autoregressive integrated moving average (arima) models to predict and monitor the number of beds occupied during a sars outbreak in a tertiary hospital in singapore. BMC Health Serv Res 5:36

Friedman JH (2001) Greedy function approximation: a gradient boosting machine. Ann Stat 29:1189–1232

Galán Soldevilla C, Cariñanos González P, Alcázar Teno P, Domínguez Vílches E (2007) Manual de Calidad y Gestión de

la Red Española de Aerobiología. Universidad de Crdoba, Crdoba

González S, Díaz J, Pajares M, Alberdi J, López C, Otero A (2001) Relationship between atmospheric pressure and mortality in the madrid autonumus region: a time series study. Int J Biometeorol 45:34–40

Haykin S (1999) Neural networks and learning machines. Pearson Prentice Hall, Upper Saddle River

Jiménez E, Linares C, Matínez D, Díaz J (2010) Role of saharan dust in the relationship between particulate matter and short-term daily mortality among the elderly in madrid (spain). Sci Total Environ 408:5729–5736

Kumar A, Goyal P (2011) Forecasting of daily air quality index in delhi. Sci Total Environ 409:5517–23

Kumar V, Mangal A, Panesar S, Yadav G, nd D, Raut RT, Singh S (2014) Forecasting malaria cases using climatic factors in delhi, india: a time series analysis. Malar Res Treat 482:851

Linares C, Díaz J (2008) Impact of high temperatures on hospital admissions: comparative analysis with previous studies about mortality (madrid). Eur J Public Health 18:318–322

Linares C, Mirón I, Sánchez R, Carmona R, Díaz J (2016) Time trend in natural-cause, circulatory-cause and respiratory-cause mortality associated with cold waves in spain, 1975–2008. Stoch Res Risk Assess 30:1565–1574

Luque M, Bauerfiend A, Díaz J, Linares C, Omeire N, Herrera D (2009) Influence of temperature and rainfall on the evolution of cholera epidemics in lusaka, zambia 2003–2006: analysis of a time series. Trans R Soc Trop Med Hyg 103:137–143

McWilliams S, Kinsella A, O'Callaghan E (2014) Daily weather variables and affective disorder admissions to psychiatric hospitals. Int J Biometeorol 58:2045–57

Montero J, Mirón I, Criado-Álvarez J, Linares C, Díaz J (2012) Relationship between mortality and heat waves in castile-la mancha (1975–2003): influence of local factors. Sci Total Environ 414:73–78

Navares R, Aznarte J (2016a) Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features. Int J Biometeorol. https://doi.org/10.1007/s00484-016-1242-8

Navares R, Aznarte J (2016b) What are the most important variables for poaceae airborne pollen forecasting? Sci Total Environ 579:1161–1169

Navares R, Aznarte JL (2017) Forecasting the start and end of pollen season in madrid. In: Rojas I, Pomares H, Valenzuela O (eds) Advances in time series analysis and forecasting. ITISE 2016. Contributions to Statistics. Springer, Cham

Nsoesie E, Mekaru S, Ramakrishnan N, Marathe M, Brownstein J (2014) Modeling to predict cases of hantavirus pulmonary syndrome in chile. PLoS Negl Trop Dis 8:e2779

Olsen J, Mitchell R, Mackay D, Humphreys D, Ogilvie D, Team MS (2016) Effects of new urban motorway infrastructure on road traffic accidents in the local area: a retrospective longitudinal study in scotland. J Epidemiol Community Health 70:1088–1095

Quero X, Viana M, Moreno T, Alastuey A (2012) Bases científico-técnicas para un plan nacional de mejora de la calidad del aire. Informes CSIC

Rodríguez-Rajo F, Valencia-Barrera R, Vega-Maray A, Suárez F, Fernández-González D, Jato V (2006) Prediction of airborne alnus pollen concentration by using arima models. Ann Agric Environ Med 13:25–32

Roldán E, Gómez M, Pino M, Pórtoles J, Linares C, Díaz J (2016) The effect of climate-change-related heat waves on mortality in spain: uncertainties in health on a local scale. Stoch Res Risk Assess 30:831–839

Rumelhart DE, Hinton GE, Ronald RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536

White H (1989) Learning in artificial neural networks: a statistical perspective. Neural Comput 1:425–464

Wolpert DH (1992) Stacked generalization. Neural Netw 5:241–259

Zhu T, Luo L, Zhang X, Shi Y, Shen W (2015) Time series approaches for forecasting the number of hospital daily discharged inpatients. IEEE J Biomed Health Inform. https://doi.org/10.1109/JBHI.2015.2511820

Chapter 10

Deep learning architecture to predict daily hospital admissions.

Type:	Published Article
Title:	*Deep learning architecture to predict daily hospital admissions*
Journal:	Neural Computing and Applicationas
Authors:	Ricardo Navares & José Luis Aznarte
Published:	March 2020
Impact Factor:	4.664
Quartile:	Q1
DOI	10.1007/s00521-020-04840-8

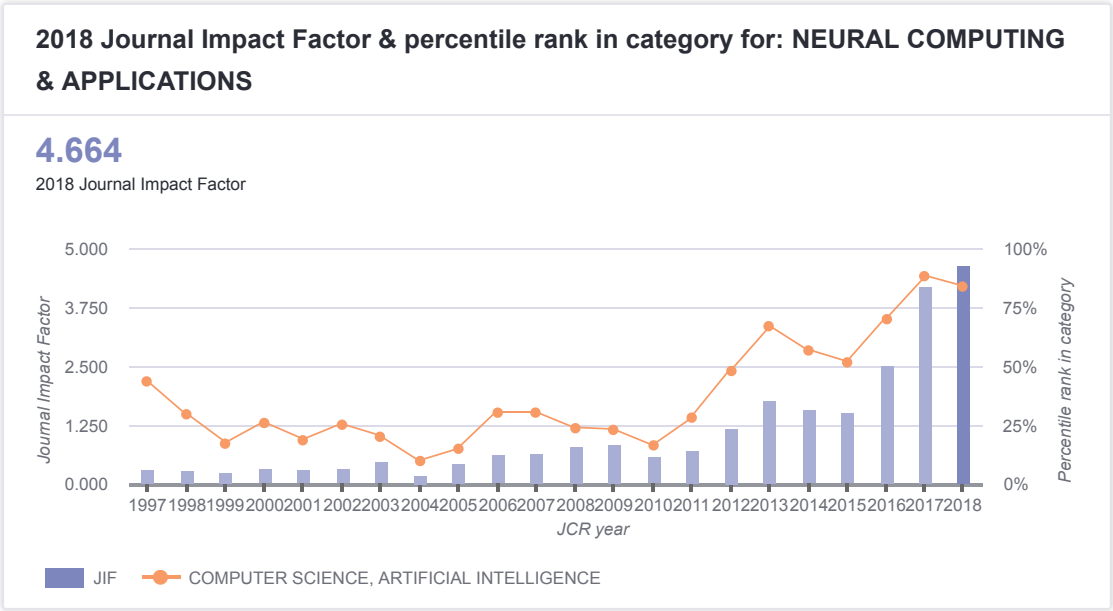


FIGURE 10.1: Impact factor Neural Computing and Applications

Neural Computing and Applications
https://doi.org/10.1007/s00521-020-04840-8

REVIEW ARTICLE

Deep learning architecture to predict daily hospital admissions

Ricardo Navares[1] · José L. Aznarte[1]

Received: 2 August 2019 / Accepted: 5 March 2020

Abstract
Air pollution and airborne pollen play a key role in respiratory and circulatory disorders and thus have a direct relation to hospital admissions for these causes. Knowing in advance the influx of patients to emergency services allows clinical institutions to optimize resources and to improve their service. Since the variables influencing respiratory and circulatory-related hospital admissions belong to fields such aerobiology or meteorology, we aim for a data-based system which is able to predict admissions without a priori assumptions. Given the number and distribution of observation stations (meteorological, pollen and chemical pollution stations and hospital), previous approaches generate many model-dependent systems that need to be combined in order to obtain the full representation of future environmental conditions. A unified approach able to extract all temporal dynamics as well as all spatial relations would allow a better representation of the aforementioned conditions and consequently a more precise hospital admissions forecast. The proposed system is based on a specific neural network topology of long short-term memories and convolutional neural networks to obtain the spatio-temporal relations between all independent and target variables. It was applied to forecast daily hospital admissions due to respiratory- and circulatory-related disorders. The proposal outperforms the benchmark approaches by reducing as an average the prediction error by 28% and 20% for the circulatory and respiratory cases, respectively. Consequently, the system extracts all relevant information without specific field knowledge and provides accurate hospital admissions forecasts.

Keywords Convolution · Neural networks · Forecasting · Hospital admissions

1 Introduction

During the last few decades, air pollution and allergens have been consistently linked with mortality [3, 10, 13, 37, 41, 47] due to its known effects on patients with respiratory and circulatory disorders [11, 12]. Forecasting these pathologies is a critical issue in order to apply preventive measures and to plan medical resources [4, 34] in order to avoid congestions and overcrowding emergency departments in hospitals [59]. Knowing in advance the influx of patients emergency admissions eases clinical institutions management in optimizing resources with the consequent economic implication [2, 8, 14, 16–19, 25, 39, 60]. Furthermore, improving the efficiency of resources is directly related to the improvement in patient care [46].

Air pollution and allergens are one of many environmental factors which play a causative role in the incidence of respiratory and circulatory diseases [30]. With this in mind, within the context of this study the term air pollution is used in a wide sense, referring to both chemical air pollutants and airborne pollen concentrations of Plantago and Poaceae, which are considered two of the most aggressive genus in relation to respiratory disorder [55]. Traditional observation-based models employ a number of different methods to relate records of air concentrations (either chemical or biological) to one or more variables that can be measured or predicted, usually meteorological data. Examples include regression models [50, 54], time-series models [53], computational intelligence techniques [5, 20, 21, 40, 43, 52] and, in the case of pollen concentrations, process-based phenological models [51] or source-based (such as traffic, heating systems) models for the chemical pollutants [29]. Given the rich availability of

✉ Ricardo Navares
rnavares2@alumno.uned.es

[1] Department of Artificial Intelligence, UNED, Juan del Rosal, 16, 28040 Madrid, Spain

Published online: 16 March 2020

techniques and methods, clinical research directly includes current variables levels to predict current cases of diseases related to air quality conditions [8, 9, 12, 45]. One of the main drawbacks of such models is that they are generally specific to a particular site or a particular pollutant or pollen genus.

Additionally, the nature of the independent variables implies several fields of expertise such as meteorology, biology and environmental sciences among others. Predicting chemical pollutants and airborne pollen concentrations based on meteorological conditions is inherently different problems: Atmospheric pollen concentrations depend on plant development during previous seasons which, in turn, depends on the climatological conditions during plant evolution [7, 54]. This implies long and mid-term relations between past atmospheric conditions and current plant status. Contrarily, chemical air pollutant levels are related to recent past atmospheric conditions [45]. These differences suggest different approaches when predicting each problem if more traditional methods are used. In addition to the temporal dimension, it is important to take into account the spatial interactions between observation stations as they are implicitly related. The access to these expertise fields resources is not always available in research departments; neither is it the time to deepen into each specific model driver. Therefore, a system which can tackle the problem from a pure data-based point of view would be of high interest when this lack of resources occurs. Furthermore, due to data protection laws, patient information is usually available at aggregated levels to avoid determining their precedence and consequently it is not possible to assign them to their exposure area.

Soft computing techniques have been gaining importance due to satisfactory results when applied to real-world problems [20, 52]. Specifically, long short-term networks are applied to identify temporal structures in time series [20, 57] and convolutional neural networks to extract local spatial patterns [52]. Given the different behaviors of the variables involved (Fig. 1), this research combines both methods to pose that if there exists mutual influential relations between the variables as well as shared implicit information (both temporal and spatial), the information and the relations can be automatically extracted in order to obtain an optimal solution of the prediction model that can be used to forecast its impact on hospital admissions.

The objective of this research is to develop a method able to deal with the previously introduced problems and limitations of current approaches. A system able to efficiently predict air quality allows to establish future environmental scenarios out of which an improved and realistic family of clinical models, among other applications, can be derived to forecast urgency admissions due to respiratory and circulatory causes. The proposed system based on long short-term memory networks [24, 28] and convolutional neural networks [35] automatically extracts the relevant information and their interactions given all data available, managing to filter and transform the information based on its forecasting capability on hospital admissions. The proposal will be compared with already published related studies showing its benefits in terms of accuracy and cost and research time saving as no feature engineering and biometeorological background is needed.

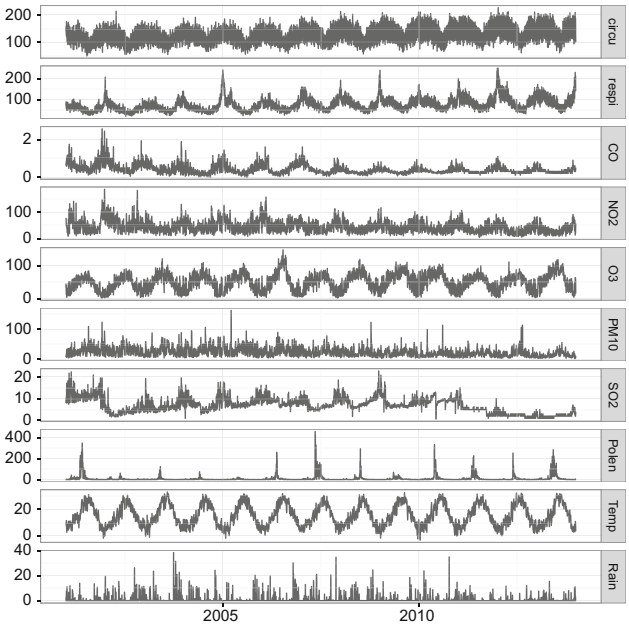


Fig. 1 Sample observations of pollutants

2 Materials and methods

2.1 Data description

Target Variables consist of daily hospital admissions of patients older than 65 years who were recorded by the medical system between January 1, 2001, and December 31, 2013, as emergency cases due to either circulatory or respiratory cases. This demographic range represents as an average around 90% of the mortality cases for each disorder according to the World Health Organization.[1] Due to data protection and confidentiality policies, the data were provided at an aggregated level across the region of Madrid preventing any kind of spatial analysis.

Chemical air pollutants consist of daily mean concentrations (in $\mu g/m^3$) of particulate matter of 10 μm in diameter (PM10), carbon monoxide (CO), sulfur dioxide (SO_2),

[1] https://www.who.int/healthinfo/global_burden_disease/estimates/en/index1.html.

Table 1 Availability of variables and locations

	Long.	Lat.	CO	NO_2	O_3	Plantago	Poaceae	PM10	SO_2	Pr	R	Hum	*T*	*W*
ArturoSoria	3° 38′ W	40° 26′ N	*	*	*						*			
BarrioPilar	3° 42′ W	40° 28′ N	*	*	*						*			
CasaCampo	3° 44′ W	40° 25′ N	*	*	*			*	*	*	*	*	*	*
CuatroCaminos	3° 42′ W	40° 26′ N		*					*		*		*	
Farmacia	3° 45′ W	40° 27′ N				*	*							
Farolillo	3° 43′ W	40° 23′ N	*	*	*				*		*		*	
Moratalaz	3° 38′ W	40° 24′ N	*	*				*	*					
PlazaEspana	3° 42′ W	40° 25′ N	*	*					*	*	*	*	*	*
PzadelCarmen	3° 42′ W	40° 25′ N	*	*	*				*					
PzaLadreda	3° 43′ W	40° 23′ N								*	*	*	*	*
RamonyCajal	3° 40′ W	40° 27′ N		*							*			
StaEugenia	3° 36′ W	40° 22′ N									*		*	*
Vallecas	3° 39′ W	40° 23′ N		*				*	*		*			

* represents data availability

Pr pressure, *R* rain, *Hum* humidity, *W* wind speed, *T* average temperature

ozone (O_3) and nitrogen dioxide (NO_2) measured by Madrid's Municipal Air Quality Monitoring Network (http://www.mambiente.munimadrid.es/). This network records hourly values for these air pollutants, which are then aggregated to daily mean levels, at 24 urban stations (Table 1). Given maintenance issues and network updates, not all data points are available at all locations so only those with a limited amount of missing data points were taken into consideration (Table 1). Sparse missing data were linearly interpolated using the precedent and the subsequent available data points.

These pollutants are the most problematic in terms of air quality (with European regulations fixing yearly thresholds for each of them) and are considered the major primary pollutants in Madrid, including ozone which shows its influence in respiratory cases [8].

Weather observations are daily temperature in Celsius degrees, wind speed measured in m/s, daily rainfall in mm/h, pressure in mbar and degree of humidity in percentage. Data sets for locations (Table 1) consist of observations from January 1, 2001, to December 31, 2013, and located as shown in Table 1.

Pollen observations correspond to daily grains per cubic meter of Poaceae and Plantago pollen registered at Complutense University of Madrid (Pharmacy Faculty). Pollen counts followed the standard methodology of the Spanish Aerobiological Network [56] and were provided by Red Palinológica de la Comunidad de Madrid. The nature of pollen time series (very low concentrations across the year with sudden high peaks during the pollination season) demands special attention to the missing data during the critical months of February, March and April. The weighted interpolation proposed by [42] was used in order to deal with the missing data points, which account for less than 3% of the total. These pollen genus are considered among the most aggressive in relation to health disorder in Madrid [55].

Figure 1 shows the target time series (in the first two rows) along with a sample of the pollutants considered in this study. In the case of respiratory admissions, it can be seen that most of the cases were recorded during winter months while circulatory admissions show a larger number of records across the year around the seasonal component. With respect to chemical and biotic pollutants, we can find a similar mixed behavior with higher variability around the seasonal component as in the case of O_3 or the presence of higher peaks with very low levels out of the pollination season for the pollen airborne concentrations.

2.2 Methodology

Long short-term memory networks (LSTM) were first proposed by [28] and improved in 2000 by [24]. They are a variation of recurrent neural networks (RNNs) capable of learning long-term dependencies by including in the architecture special units called memory blocks which aim to overcome the issue of the vanishing gradient [27].

An LSTM unit performs self-loops which enable the flow of the gradient for long durations, enabling it to deal with the vanishing gradient problem. Together with an input gate, an output gate and a forget gate, this architecture models the short-term memory that allows the network to learn over many time steps. For this reason, LSTM had been shown to outperform more traditional recurrent networks on several temporal processing tasks [24].

The learn gate. The learn gate takes the short-term memory (STM) and the input event and combines them. Actually, after combining the event and the STM, it ignores redundant information. Mathematically, the learn gate obtains as an input the short-term memory STM_{t-1} and the event E_t and puts them into a linear function which consists of joining the vectors, multiplying it by the weight matrix W_n, adding a bias b_n and squeeze the result with a tanh activation function:

$$N_t = \tanh(W_n \cdot [\text{STM}_{t-1}, E_t] + b_n). \tag{1}$$

The new information N_t passes through the gate but still needs to ignore the information which is not relevant. In order to do so, N_t is multiplied by the ignore vector i_t. This ignore vector is calculated via a simple small neural network whose inputs are again the STM and the event and uses the sigmoid (σ) activation function to squeeze the information:

$$i_t = \sigma(W_i \cdot [\text{STM}_{t-1}, E_t] + b_i), \tag{2}$$

being the learn gate represented as $N_t \times i_t$.

The forget gate. Takes the long-term memory (LTM) and decides which parts to keep and to forget. The LTM at $t-1$ is multiplied by a forget factor f_t which is calculated through a one-layer neural network with a linear function, which uses the STM at $t-1$ and the event E_t, and combines it with a sigmoid activation:

$$f_t = \sigma\big(W_f \cdot [\text{STM}_{t-1}, E_t] + b_f\big) \tag{3}$$

being b_f the bias and W_f the weight matrix. The forget gate can be expressed as $\text{LTM}_{t-1} \times f_t$.

The remember gate. Takes the output from the forget gate and the output from the learn gate and adds them to obtain the new LTM:

$$\text{LTM}_t = \text{LTM}_{t-1} \cdot f_t + N_t \cdot i_t \tag{4}$$

The use gate. It combines the LTM that just came out from the forget gate and the STM that came out from the learn gate to come out with a new STM and an output. In order to do so, it applies a small neural network on the output of the forget gate using the tanh activation function (5) and another neural network on the STM and the events using the sigmoid function (6):

$$U_t = \tanh(W_u \cdot \text{LTM}_{t-1} \cdot f_t + b_u) \tag{5}$$

$$V_t = \sigma(W_v[\text{STM}_{t-1}, E_t] + b_v) \tag{6}$$

As a final step, the network multiplies (5) and (6) to obtain the new output $STM_t = U_t \times V_t$ which also works as a new STM.

Convolutional Neural Networks (CNN) [35] have been successfully applied in several domains such as image recognition [33] or linguistics [23]. Based on their success, researchers have started to use them for time-series analysis [22]. CNNs differ from feed-forward neural networks mainly by the existence of convolutional layers, which are hidden layers that utilize the power of mathematical convolution to transform inputs. Convolution allows for the encoding of the local properties of the input in such a way that propagates the information in a more efficient manner. CNN filters, obtained by the convolution of inputs and weights, are local in input space and are thus to exploit the strong, spatially local correlation present in the time series. That means that they work well for identifying simple patterns within local regions of the data (subset of features) which then will be used by subsequent layers to form more complex patterns. One-dimensional CNNs share the same characteristics with the most commonly used two-dimensional ones differing only in the dimensionality of the input and how the filter slides across the data.

Several studies have proved the advantages of one-dimensional CNN [1, 32] for certain applications when compared to more complex, deeper and higher-dimensional CNNs. Instead of matrix or tensor operations, the convolution operation requires only array operations (Eq. 7) to define the feature map $s[n]$

$$s[n] = (f * g)[n] = \sum_{m=-\inf}^{\inf} f[m]g[n-m] \tag{7}$$

where g is the convolutional filter, f the input space and m is the size of the filter. This low computational provides a major advantage in forward and backpropagation operations as they can effectively be executed in parallel.

The main difference between LSTM units and CNNs is that the latest only considers current inputs, exploiting input local correlations and properties present. This means that CNNs are able to identify patterns within local regions of the data (features). On the other hand, LSTMs also consider previously input signals taking advantage of the seasonal information present in the time series (Fig. 1).

The majority of the studies found in the literature combine, in this precise order, CNNs and LSTMs for sequence prediction problems with spatial inputs as can be the case of images or videos [15, 58]. Placing CNN before allows to capture the spatial structure of one variable. An alternative is to transform the one-dimensional time series into two-dimensional matrix as an input of the CNN to capture temporal patterns. In this problem, we have temporal and spatial dimensions for 60 variables (Table 1). In order to deal with this multivariate problem, the novelty presented in this study is to reverse the order in order to achieve patient influx forecasts via LSTMs which then are

parsed to the CNN which acts as an error-corrector model. Furthermore, the characteristic topology of the LSTM module (grouped by pollutant type) eases the discrimination of non-relevant information for predicting.

In order to compare the results with similar research studies, the widely used [16–18, 25] scoring rule root mean squared error (RMSE) will be used to measure the average magnitude of the error along with the coefficient of determination R^2:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2}, \tag{8}$$

where y_i is the observed ith data point, $\hat{y_i}$ the predicted and n the total number of data points in the test set. RMSE provides error magnitudes directly comparable to the target time series which eases its interpretability. The results will be directly compared with traditional time-series techniques and computational intelligence models proposed by [45].

2.3 Experimental design

The aim is to provide the most accurate forecast of one-day-ahead cases of hospital admissions due to circulatory and respiratory cases through algorithm selection, configuration and parametrization. Thus, avoiding any kind of dimensionality reduction, either through variable selection or feature engineering. Consequently, this approach saves human resources in terms of expertise in the different fields the variables involved such as meteorology, botany or medical fields.

In order to do so, a k-fold cross-validation (CV) is performed over the full historical data set (January 1, 2001, to December 31, 2013) using the LSTM neural network to generate forecasts. These forecasts are then parsed through a three-layer one-dimensional CNN, along with previous-day observation of circulatory and respiratory cases, to perform the final prediction (Fig. 2). The technique consists of leaving one year out for testing and using the remaining for training. This is repeated per year (13-fold CV) in the study period resulting in a full new set of one-day-ahead forecasts between January 1, 2001, and December 31, 2013. Given the serial correlation and the non stationarity of the involved time series, this application is usually avoided by practitioners, but it was favorably compared to traditional out-of-sample evaluation by [6]. Furthermore, CV method includes each data point as test set once so no decision needs to be made by an expert in order to define out-of-sample periods. The underlying idea is letting the first block of the network to obtain and store the relevant spatial and temporal relations to forecast the target variables to subsequently allow the CNN block to serve as a error correction model.

As mentioned in Sect. 2.1, hospital admissions were provided at aggregated level due to confidentiality issues. As a result, it is not known beforehand which local pollutants and observation stations are the best candidates to directly influence patient admissions. LSTM network topology is aimed to ease the discrimination process by using separated groups of LSTM units based on the type of pollutant plus two extra unit groups to extract the temporal relations from the circulatory and respiratory observations, respectively. Each group consists of 100 LSTM units, yielding a total of 900 units in the layer (five chemical pollutants, two air allergens and two target variables), which are fully connected to a hidden layer of 100 rectified linear units (ReLU) in order to obtain the spatial relations as well as the interactions among pollutants and hospital admissions. For instance, in the case of carbon monoxide, 7 one-day-lagged input variables, one per observation station (Table 1), are input to one group of 100 LSTM units which are then fully connected to the subsequent hidden linear layer. The hidden ReLU layer is connected to 2 ReLU neurons which generate the one-day-ahead estimations for hospital admissions due to circulatory and respiratory cases. The results from these 2 output neurons will be used to feed the one-dimensional convolutional neural network.

Then, the CNN receives as an input the estimated values of the LSTM module along with previous-day circulatory and respiratory observations, and thus the input space width of the CNN is of size 4. The input is parsed to a three-layer of 16 convolutional filters of size 2 each. Usually, in order to reduce the dimensionality of parameters of a CNN, a summarization is performed through a pooling layer which was not found, during the study, beneficial for this problem when accuracy is taken into account. Finally, the CNN block is fully connected to 50 ReLU neurons that transfer the information to a 2-neurons output layer which generates the final predictions, one per target variable in order to form the full system LSTM-CNN.

The relationships between the inputs and the outputs are obtained through the well-known *backpropagation* algorithm proposed by [49], employing the mean squared error as the loss function, which heavily penalizes large errors when compared to RMSE used as a report metric in this study. In order to avoid the vanishing gradient problem in recurrent neural networks, the *Adam* algorithm proposed by [31] was used as optimization model to fit network weights. Compared to the classic stochastic gradient descent (SGD) which maintains a single learning rate α for all weight updates, the method computes individual adaptive learning rates from the estimates of first and second moments of the gradients [31]. Specifically, the algorithm uses an

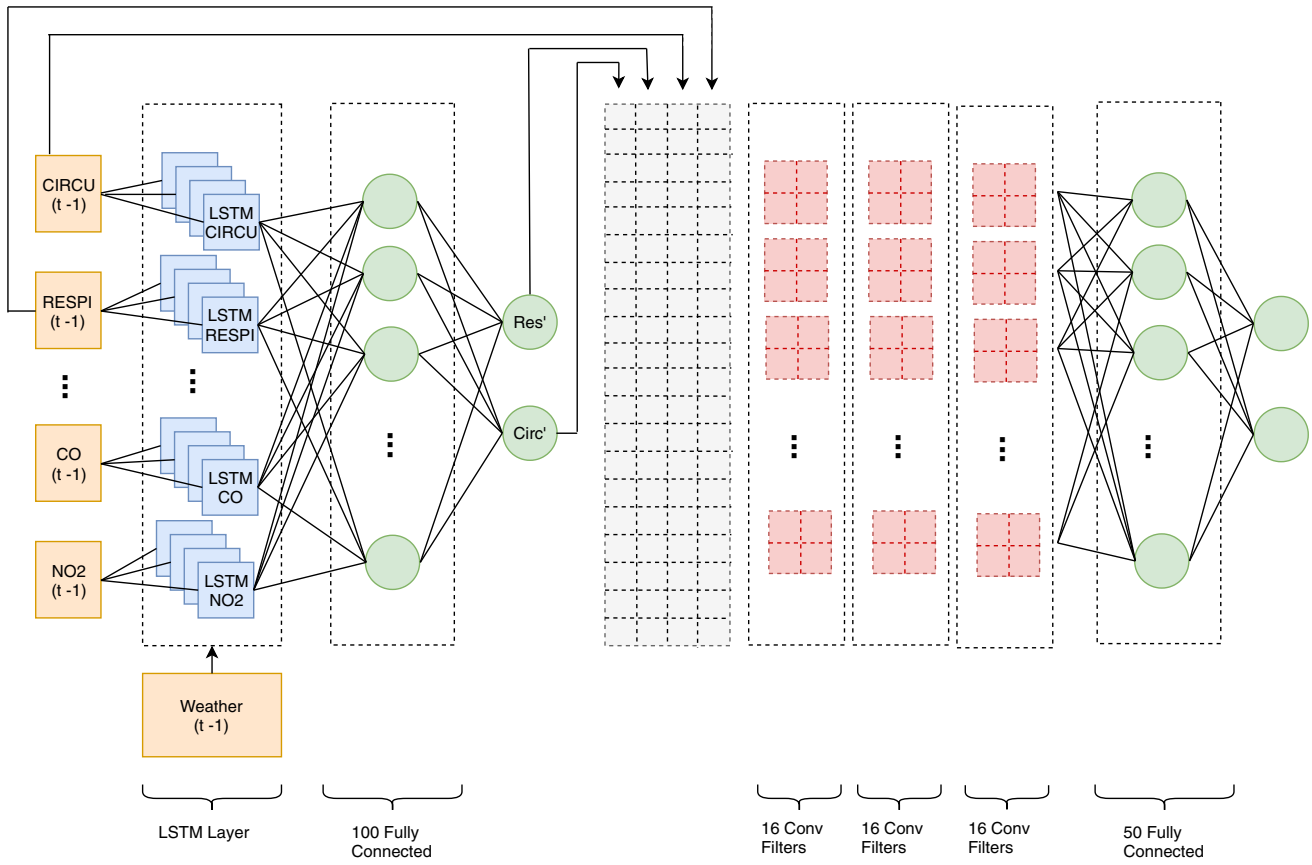


Fig. 2 Network architecture

exponential moving average of the gradient and its square using the parameters β_1 and β_2 to control the decay rates of these moving averages. This bias-correction helps Adam slightly outperform the alternative RMSprop toward the end of optimization as gradients become sparser. In order to obtain the optimal set of values, a grid search was performed over all possible combinations of parameter values based on the cross-validation results. A learning rate $\alpha = 0.001$, and exponential decays of $\beta_1 = 0.9$ and $\beta_2 = 0.999$ were used as suggested by [48]. The full network is trained using batches of 50 observations over 1000 epochs.

3 Results

As stated above, the study period consists of daily morbility cases due to circulatory and respiratory cases from January 1, 2001, to December 31, 2013. As a first step of the experiment, only the LSTM module was iteratively evaluated leaving one full year out and using the remaining years as a train set as suggested by [6]. After 13 rounds (one per year), we obtain one-day-ahead forecasts for the full study period with RMSEs detailed in Table 2.

On average, an RMSE of 13.98 is achieved for circulatory cases and an RMSE of 15.06 in the case of the admissions due to respiratory disorders. The error is consistently lower across most years for the circulatory cases as the time series show less variance around the seasonal pattern when compared to the respiratory cases which are characterized by the presence of higher spikes during winters, as shown in Figs. 3 and 4. A R^2 of 0.88 and 0.84 was obtained for circulatory and respiratory cases, respectively (Table 2).

The second part of the experiment intends to improve over the errors produced by the LSTM module. In order to do so, its outputs are used to feed the CNN module along with the previous-day circulatory and respiratory cases observations to compile the full system shown in Fig. 2. Figure 3 shows the comparison between the predicted values of the LSTM and the full setup for circulatory cases. It can be clearly seen that including the CNN module helps to better extract the variance of the observations driving the average RMSE down from 13.98 to 11.21 (Table 2) and increasing the coefficient of determination R^2 from 0.88 to 0.93. Figure 4 shows a similar behavior of the CNN for respiratory cases achieving an RMSE of 11.76 with a $R^2$of

Table 2 Results for the evaluation set

	Year	Circulatory		Respiratory	
		LSTM-CNN	LSTM	LSTM-CNN	LSTM
RMSE	2001	11.57	12.37	10.46	11.51
	2002	11.84	15.19	11.24	15.56
	2003	11.16	13.97	12.56	16.20
	2004	12.24	14.90	9.96	16.78
	2005	9.64	11.94	12.15	15.23
	2006	12.03	13.24	10.93	11.48
	2007	9.47	14.49	11.84	16.35
	2008	12.00	14.94	13.82	16.32
	2009	12.43	15.15	13.44	15.86
	2010	9.93	14.59	10.86	15.49
	2011	9.72	13.77	11.68	15.88
	2012	11.69	13.97	11.23	15.28
	2013	11.99	13.16	12.69	13.84
	Average	11.21	13.98	11.76	15.06
R^2	2001	0.91	0.88	0.84	0.81
	2002	0.91	0.86	0.89	0.79
	2003	0.89	0.83	0.82	0.73
	2004	0.88	0.80	0.95	0.85
	2005	0.93	0.90	0.94	0.92
	2006	0.92	0.87	0.86	0.85
	2007	0.94	0.86	0.91	0.83
	2008	0.93	0.88	0.86	0.85
	2009	0.91	0.87	0.87	0.81
	2010	0.95	0.90	0.89	0.81
	2011	0.95	0.91	0.94	0.88
	2012	0.95	0.91	0.96	0.93
	2013	0.97	0.96	0.92	0.91
	Average	0.93	0.88	0.90	0.84

0.90 compared to the RMSE of 15.06 obtained in the first experiment.

Table 2 shows how consistently LSTM-CNN improves the metrics for the circulatory and the respiratory cases across all years. Not only the error is consistently lower for this model on every year and both series, but also the variance of the error is reduced in most of the cases. As we can see in the circulatory cases (Fig. 3), the LSTM configuration underestimates higher and lower observed values around the seasonal component. This problem is overcome by the CNN correcting this situation and showing more variance. Similarly, the respiratory cases show heavy-tailed distributions due to high peaks during winter periods. The LSTM configuration tends to underestimate winters with the presence of high peaks such as 2004, with a maximum of 242 patients with a yearly average of 63, and 2008 with

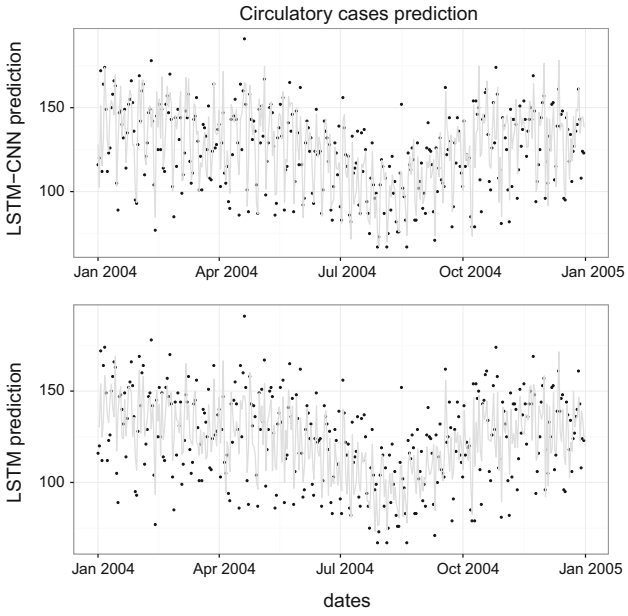


Fig. 3 Circulatory predictions (solid line) and observed values (dots) comparison between LSTM module (bottom chart) and full CNN network (top chart)

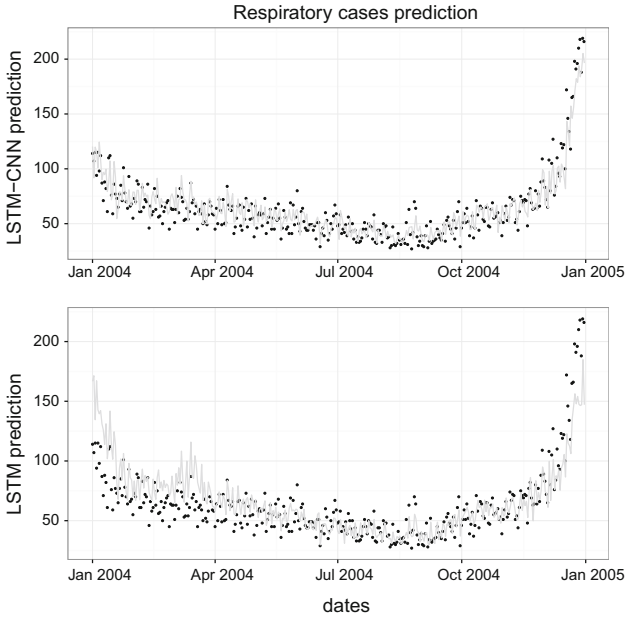


Fig. 4 Respiratory predictions (solid line) and observed values (dots) comparison between LSTM module (bottom chart) and full CNN network (top chart)

218 maximum admissions with a yearly average of 79, obtaining an RMSE of 14.90 and 14.94, respectively. Also it overestimates years which have low influx of winter admissions when compared to other years such as 2003, with a maximum of 140 and a yearly average of 64, and 2010 with a maximum of 173 with a yearly average of 81 patients, obtaining above the average errors in both cases.

On the other hand, LSTM-CNN setup describes better these situations and consequently drives the errors on these years closer or below its average.

4 Discussion

As we have seen, the proposed approach of stacking the LSTM results through a convolutional neural network outperforms the first experiment based only in LSTMs. The full system manages to obtain relevant information free from any feature engineering or synthetic input variable formation. Consequently, it has immediate practical application since no field specific research is needed. Nevertheless, the forecasting capabilities of the proposal are expected to improve if input variables are derived based on field knowledge such as biology or environmental sciences. Variable selection and feature engineering improve model performance [44] as they ease the model in extracting relevant information since it was preprocessed by a human expert. Still, the initial assumption of this study was the lack of these resources.

The study was limited by the amount of data available. Deep architectures generally require amounts of data in order to be trained, an improvement in the results is expected by increasing the study period. Furthermore, this limitation in the availability of the variables made PM2.5 pollutant to be excluded from this study which shows direct influence in respiratory cases [8]. Also, Olea pollen concentrations would contribute to draw a better picture in terms of air quality representations as it has more effects on health than Plantago [55].

Even though the results are satisfactory, the appearance of high peaks during winter in the case of respiratory disorders increases the error as these high accumulations of admissions occur during a short period and, therefore, the number of observations is limited in order to learn the pattern. However, this circumstance is also common in other approaches.

The results presented in this study are directly comparable in terms of accuracy to previous studies such as [45] (Table 3), which test a traditional time-series approach (ARIMA) and several computational intelligence models. The aforementioned study reports an RMSE of 13.04 in the case of respiratory-related admissions. For this variable, the pure LSTM approach does not manage to outperform the ARIMA model, but it does when compared to the artificial neural network and the tree-based algorithms. However, LSTM-CNN reduces the error to 11.75. On the other hand, [45] shows an RMSE of 13.24 for the circulatory admission, which means a close performance when compared to the LSTM approach while LSTM-CNN outperforms with an RMSE = 11.20. Besides the improvement in accuracy, the proposal saves the research step of finding relevant independent variables for the model as LSTM-CNN manages to extract this information from the algorithmic point of view. As a consequence, no expert human input and specific research is needed when compared to the referred research.

Table 3 Global results (the first five rows are taken from [45])

Model	Circulatory		Respiratory	
	RMSE	R^2	RMSE	R^2
RF	22.96	0.77	22.64	0.74
GBM	18.62	0.83	16.87	0.85
ANN	13.69	0.90	15.06	0.89
ARIMA	17.11	0.84	13.15	0.91
STACK	13.24	0.90	13.04	0.92
$LSTM_{avg}$	13.98	0.88	15.06	0.84
$LSTM\text{-}CNN_{worst}$	12.43	0.88	12.56	0.84
$LSTM\text{-}CNN_{best}$	9.47	0.97	9.96	0.96
$LSTM\text{-}CNN_{avg}$	11.21	0.93	11.76	0.90

In terms of execution time, long short-term memory networks require high computational cost, especially when the number of dimensions of memory cell is high [38] given the direct relation between the number of weights and the dimensions of the memory cells. Several proposals tackled this increase in complexity using lighter architectures [36]; however, [26] shows that default topologies do not suffer from major loss in accuracy. Nevertheless, this situation is overcome by the initial hypothesis of this research as the proposal can be input with any other variable the researcher might consider influential such as a different location, another pollen genus or additional pollutants regardless of the intrinsic nature of the time series as, for example, seasonal patterns, spikes or trends.

Finally, the usefulness of the proposal is shown by its performance when compared to the benchmark research [45] and its flexibility to add or remove input variables without the need of doing extra research or having field expert input as it was tested without any feature generation. Consequently, it shows a potential save of costs when resources are not available.

5 Conclusions

The present study proposes an algorithmic approach for forecasting daily hospital admissions due to respiratory and circulatory disorders based on environmental indicators. The initial hypothesis was to apply artificial intelligence methodologies to avoid feature engineering as not always

the resources and research time of the independent input variables are available. Therefore, the algorithm architecture needs to be able to automatically extract and filter relevant information for prediction. We have seen that CNN-LSTM allows not only to solve the problem but also to obtain better results, in terms of accuracy, when compared to already published approaches which include the mentioned independent variable preprocess and research. However, we believe that including feature engineering and selection preprocess will improve even further the results.

One of the major limitations of deep architectures is the interpretability of the results and the identification of the contribution of each variable to the prediction. This situation makes deep architectures weak in terms of diagnosis as their explainability is limited. Notwithstanding, several techniques are proposed such as Shapley additive explanations in order to cope with these limitations.

Even though LSTMs have been proved successful in time-series applications, we showed the benefits of stacking their predictions through one-dimensional convolutional neural networks. CNNs are being widely used in several study fields, but still the number of applications in time-series forecasting problems is fairly limited when compared to other areas. Despite the fact the results are promising and outperform previous studies, there is still a need for further research in network topologies in order to improve performance and reduce model complexity. Complexity would be definitely reduced by using Gated Recurrent Units as they use two gates instead of the three used in the LSTMs. However, there is a potential loss of long-term information which might influence forecast accuracy.

Acknowledgements This work was only possible thanks to the ongoing fruitful collaboration with Julio Díaz and Cristina Linares, from the Carlos III National Institute of Health, Madrid, Spain. References [16–18, 25] were added upon request by Reviewer 3.

Compliance with ethical standards

Conflict of interest The authors declare no conflict of interests.

References

1. Abdeljaber O, Avci O, Kiranyaz S, Boashash B, Sodano H, Inman D (2017) 1-D CNNs for structural damage detection: verification on a structural health monitoring benchmark data. Neurocomputing 275:1308–1317
2. Abraham G, Byrnes GB, Bain CA (2009) Short-term forecasting of emergency inpatient flow. Inf Technol Biomed 13:380–388
3. Alberdi JC, Díaz J, Montero JC, Mirón IJ (1998) Daily mortality in madrid community (Spain) 1986–1991: relationship with atmospheric variables. Eur J Epidemiol 14:571–578
4. Anwar MY, Lewnard JA, Parikh S, Pitzer VE (2016) Time series analysis of malaria in Afghanistan: using arima models to predict future trends in incidence. Malar J 15:566
5. Baghban A, Jalali A, Shafiee M, Ahmadi M (2018) Developing an anfis based swarm concept model for estimating relative viscosity of nanofluids. Eng Appl Comput Fluid Mech 13:08
6. Bergmeir C, Hyndman RJ, Koo B (2018) A note on the validity of cross-validation for evaluating autoregressive time series prediction. Comput Stat Data Anal 120:70–83
7. Cannell MGR, Smith RI (1983) Thermal time, chill days and prediction of budburst in Picea sitchensis. J Appl Ecol 20:269–275
8. Díaz J, Alberdi JC, Pajares MS, López R, López C, Otero A (2001) A model for forecasting emergency hospital admissions: effect of environmental variables. J Environ Health 64:9–15
9. Díaz J, Carmona R, Mirón JL, Ortiz C, León I, Linares C (2015) Geographical variation in relative risks associated with heat: update of Spain's heat wave prevention plan. Environ Int 85:273–283
10. Díaz J, García R, López C, Linares C (2005) Mortality impact of extreme winter temperatures. Int J Biometeorol 49:179–183
11. Díaz J, García R, Ribera P, Alberdi JC, Hernández E, Pajares MS (1999) Modeling of air pollution and its relationship with mortality and morbidity in madrid (Spain). Int Arch Occup Environ Health 75:366–376
12. Díaz J, Linares C, Tobías A (2007) Short term effects of pollen species on hospital admissions in the city of madrid in terms of specific causes and age. Aerobiologia 23:231–238
13. Díaz J, López C, Jordán A, Alberdi JC, García R, Hernández E, Otero A (2002) Heat waves in Madrid, 1986–1997: effects on the health of the elderly. Int Arch Occup Environ Health 75:163–170
14. Dominak M, Swiecicki L, Rybakowski J (2015) Psychiatric hospitalizations for affective disorders in Warsaw, Poland: effect of season and intensity of sunlight. Phychiatry Res 229:289–294
15. Donahue J, Anne Hendricks L, Rohrbach M, Venugopalan S, Guadarrama S, Saenko K, Darrell T (2014) Long-term recurrent convolutional networks for visual recognition and description. arXiv eprint. arXiv:1411.4389
16. de Jesus Rubio J (2009) SOFMLS: online self-organizing fuzzy modified least-squares network. IEEE Trans Fuzzy Syst 17:1296–1309
17. de Jesus Rubio J, Cruz D, Elias Barrón I, Ochoa G, Balcazarand Ricardo, Aguilar Arturo (2019) ANFIS system for classification of brain signals. J Intell Fuzzy Syst 37:4033–4041
18. de Jesus Rubio J, García-Trinidad E, Ochoa G, Elias Barrón I, Cruz D, Balcazar R, Lopez-Gomez J, Novoa J (2019) Unscented kalman filter for learning of a solar dryer and a greenhouse. J Intell Fuzzy Syst 37:6731–6741
19. Earnest A, Chen MI, Ng D, Sin LY (2005) Using autoregressive integrated moving average (arima) models to predict and monitor the number of beds occupied during a sars outbreak in a tertiary hospital in Singapore. BMC Health Serv Res 5:36
20. Faizollahzadeh Ardabili S, Najafi B, Shamshirband S, Minaei Bidgoli B, Deo RC, Chau KW (2018) Computational intelligence approach for modeling hydrogen production: a review. Eng Appl Comput Fluid Mech 12(1):438–458
21. Fotovatikhah F, Herrera M, Shamshirband S, Chau KW, Faizollahzadeh Ardabili S, Piran MJ (2018) Survey of computational intelligence as basis to big flood management: challenges, research directions and future work. Eng Appl Comput Fluid Mech 12(1):411–437
22. Gamboa JCB (2017) Deep learning for time-series analysis. *CoRR*. arXiv:1701.01887
23. Gehring J, Auli M, Grangier D, Yarats D, Dauphin YN (2017) Convolutional sequence to sequence learning. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on

machine learning, volume 70 of proceedings of machine learning research. International Convention Centre, Sydney, Australia, 06–11 Aug 2017. PMLR, pp 1243–1252
24. Gers FA, Schmidhuber J, Cummins F (2000) Learning to forget: continual prediction with LSTM. Neural Comput 12:2451–2471
25. Giap CN, Son LH, Chiclana F (2018) Dynamic structural neural network. J Intell Fuzzy Syst 34:2479–2490
26. Greff K, Srivastava RK, Koutník J, Steunebrink BR, Schmidhuber J (2015) LSTM: a search space odyssey. *CoRR*. arXiv:1503.04069
27. Hochreiter S, Bengio Y, Frasconi P, Schmidhuber J (2001) Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In: Kremer SC, Kolen JF (eds) A field guide to dynamical recurrent neural networks. IEEE Press, New Jersey
28. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9:1735–1780
29. Hu X, Xu D, Wan Q (2018) Short-term trend forecast of different traffic pollutants in minnesota based on spot velocity conversion. Int J Environ Res Public Health 15:1925
30. Kelly FJ, Fussell JC (2015) Air pollution and public health: emerging hazards and improved understanding of risk. Environ Geochem Health 37:631–649
31. Kingma D, Ba J (2014) Adam: a method for stochastic optimization. *CoRR*. arXiv:1412.6980
32. Kiranyaz S, Ince T, Gabbouj M (2015) Real-time patient-specific ECG classification by 1D convolutional neural networks. IEEE Trans Bio-Med Eng 63:08
33. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems 25. Curran Associates, Inc., New York, pp 1097–1105
34. Kumar V, Mangal A, Panesar S, Yadav G, Talwar R, Raut D, Singh S (2014) Forecasting malaria cases using climatic factors in Delhi, India: a time series analysis. Malar Res Treat 2014:482851
35. Lecun Y, Bottou L, Bengio Y, Haffner P (1998) Gradient-based learning applied to document recognition. Proc IEEE 86(11):2278–2324
36. Li X, Qin T, Yang J, Liu T-Y (2016) LightRNN: memory and computation-efficient recurrent neural networks. arXiv eprint. arXiv:1610.09893
37. Linares C, Mirón IJ, Sánchez R, Carmona R, Díaz J (2016) Time trend in natural-cause, circulatory-cause and respiratory-cause mortality associated with cold waves in Spain, 1975–2008. Stoch Res Risk Assess 30:1565–1574
38. Masuko T (2017) Computational cost reduction of long short-term memory based on simultaneous compression of input and hidden state. In: 2017 IEEE automatic speech recognition and understanding workshop (ASRU), pp 126–133
39. McWilliams S, Kinsella A, O'Callaghan E (2014) Daily weather variables and affective disorder admissions to psychiatric hospitals. Int J Biometeorol 58:2045–57
40. Moazenzadeh R, Mohammadi B, Shamshirband S, Chau KW (2018) Coupling a firefly algorithm with support vector regression to predict evaporation in Northern Iran. Eng Appl Comput Fluid Mech 12(1):584–597
41. Montero JC, Mirón IJ, Criado-Álvarez JJ, Linares C, Díaz J (2012) Relationship between mortality and heat waves in Castile-la Mancha (1975–2003): influence of local factors. Sci Total Environ 414:73–78
42. Navares R, Aznarte JL (2016) Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features. Int J Biometeorol. https://doi.org/10.1007/s00484-016-1242-8
43. Navares R, Aznarte JL (2017) Forecasting the start and end of pollen season in Madrid. Springer, Berlin
44. Navares R, Aznarte JL (2019) Forecasting plantago pollen: improving feature selection through random forests, clustering, and friedman tests. Theor Appl Climatol 139:08
45. Navares R, Díaz J, Linares C, Aznarte JL (2018) Comparing arima and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in Madrid. Stoch Environ Res Risk Assess 32:2849–2859
46. Obermeyer Z, Emanuel EJ (2016) Predicting the future—big data, machine learning, and clinical medicine. N Engl J Med 375:1216–1219
47. Roldán E, Gómez M, Pino MR, Pórtoles J, Linares C, Díaz J (2016) The effect of climate-change-related heat waves on mortality in Spain: uncertainties in health on a local scale. Stoch Res Risk Assess 30:831–839
48. Ruder S (2016) An overview of gradient descent optimization algorithms. *CoRR*. arXiv:1609.04747
49. Rumelhart DE, Hinton GE, Ronald RJ (1986) Learning representations by back-propagating errors. Nature 323:533–536
50. Sabariego S, Cuesta P, Fernández-González F, Pérez-Badia R (2012) Models for forecasting airborne cupressaceae pollen levels in central Spain. Int J Biometeorol 56:253–258
51. Schaber J, Badeck F-W (2003) Physiology-based phenology models for forest tree species in Germany. Int J Biometeorol 47:193–201
52. Shamshirband S, Rabczuk T, Chau K-W (2019) A survey of deep learning techniques: application in wind and solar energy resources. IEEE Access 7:164650–164666
53. Silva-Palacios I, Fernández-Rodríguez S, Durán-Barroso P, Tormo-Molina R, Maya-Manzano JM, Gonzalo-Garijo A (2016) Temporal modelling and forecasting of the airborne pollen of Cupressaceae on the southwestern Iberian peninsula. Int J Biometeorol 60:1509–1517
54. Smith M, Emberlin J (2006) A 30-day-ahead forecast model for grass pollen in north London, UK. Int J Biometeorol 50:233–242
55. Subiza J, Jerez M, Jiménez JA, Narganes MJ, Cabrera M, Varela S, Subiza E (1995) Allergenic pollen pollinosis in Madrid. J Allergy Clin Immunol 96:15–23
56. Soldevilla CG, González PC, Teno PA, Vílches ED (2007) Manual de Calidad y Gestión de la Red Española de Aerobiología. Universidad de Córdoba, Córdoba
57. Valput D, Navares R, Aznarte JL (2019) Forecasting hourly NO_2 concentrations by ensembling neural networks and mesoscale models. Neural Comput Applic. https://doi.org/10.1007/s00521-019-04442-z
58. Vinyals O, Toshev A, Bengio S, Erhan D (2014) Show and tell: a neural image caption generator. *CoRR*. arXiv:1411.4555
59. Yousefi M, Yousefi M, Ferreira R Poley Martins, Kim JH, Fogliatto FS (2018) Chaotic genetic algorithm and adaboost ensemble metamodeling approach for optimum resource planning in emergency departments. Artif Intell Med 84:23–33
60. Zhu T, Luo L, Zhang X, Shi Y, Shen W (2015) Time series approaches for forecasting the number of hospital daily discharged inpatients. IEEE J Biomed Health Inform 21:515–526

Chapter 11

Side project I: Direct assessment of health impacts on hospital admission from traffic intensity in Madrid

Type:	Published Article
Title:	*Direct assessment of health impacts on hospital admission from traffic intensity in Madrid*
Journal:	Environmental Research
Authors:	Ricardo Navares & Julio Díaz & José Luis Aznarte & Cristina Linares
Published:	May 2020
Impact Factor:	5.026
Quartile:	Q1
DOI	10.1016/j.envres.2020.109254

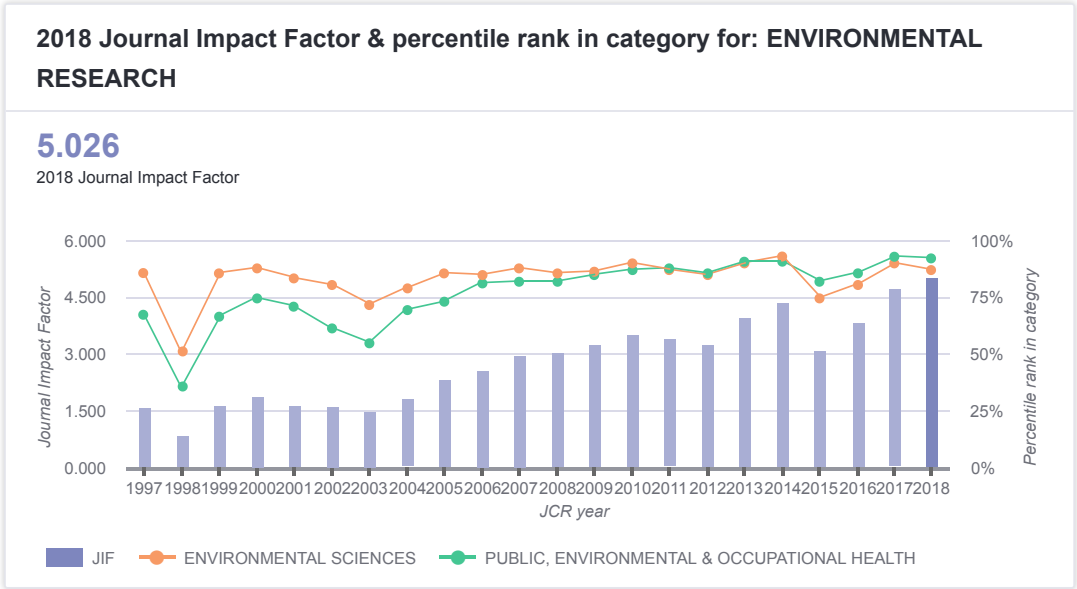


FIGURE 11.1: Impact factor Environmental Research

Environmental Research 184 (2020) 109254

Contents lists available at ScienceDirect

Environmental Research

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

Direct assessment of health impacts on hospital admission from traffic intensity in Madrid

Ricardo Navares[a], Julio Diaz[b,*], Jose L. Aznarte[a], Cristina Linares[b]

[a] Department of Artificial Inteligence, UNED, Madrid, Spain
[b] National School of Public Health, Carlos III Institute of Health, Madrid, Spain

ARTICLE INFO

Keywords:
Traffic intensity
Hospital admissions
Attributable risk
ARIMA
Poisson regression

ABSTRACT

In this paper we establish the attributable risk on respiratory and cardiovascular disorders related to traffic intensity in Madrid. In contrast to previous related studies, the proposed approach directly associates road traffic counts to patient emergency admission rates instead of using primary air pollutants. By applying Shapley values over gradient boosting machines, a first selection step is performed among all traffic observation points based on their influence on patient emergency admissions at Gregorio Marañon hospital. A subsequent quantification of the relative risk associated to traffic intensity of the selected point is calculated via ARIMA and log-linear Poisson regression models. The results obtained show that 13% of respiratory cases are related to traffic intensity while, in the case of cardiovascular disorders, the percentage increases to 39%.

1. Introduction

The impact of atmospheric pollution on hospital admissions in Madrid has been extensively researched over the last decade, both related to chemical air pollutants (Linares and Díaz, 2010a,b; de Miguel-Díez et al., 2019; Marques-Mejías et al., 2018) and to environmental noise levels (Tobías et al., 2001; Linares and Díaz, 2010a; Díaz et al., 2020; Carmona et al., 2018). All these research studies establish the relations between levels of pollutants measured at different observation stations in Madrid, mainly NO^2, PM^{10}, $PM^{2.5}$ and noise levels, and the correspondent health indicators.

In urban areas, over 55% of particulate matter (PM) is directly related to road traffic as well as about 70% of NO^2 emissions (Quero et al., 2012). With respect to environmental noise levels, the percentage associated with road traffic surpasses 70% (Recio et al., 2016). Even though road traffic is the major source of pollution in big cities, there are no previous studies which relate the main cause (road traffic) with the effect (health indicators).

The main objective of this paper is to analyze the association between traffic intensity and hospital admissions. This differs from previous research proposals since it directly analyses the impact of the daily number of vehicles on hospital admissions due to respiratory and cardiovascular disorders. As a consequence of the results obtained, it is also intended to increase the comprehension of the effects of intense road traffic in major cities.

Exposure to transport-related air pollution increases the risk of premature death due to respiratory and cardiovascular causes (Krzyzanowski et al., 2005; WHO Regional Office for Europe, 2013; Burns et al., 2020; EEA, 2020; Mannucci et al., 2019). Knowing the attributable risk associated with road traffic not only enables traffic control policies to local authorities, but also increases the awareness of the aftereffect to its exposure.

Among all traffic observation points surrounding hospital Gregorio Marañon in Madrid (Fig. 2), the proposal selects the one which has the most impact on the number of emergency admissions due to respiratory and cardiovascular cases recorded. Gradient boosting machines (Friedman, 2001) were used to perform variable selection. Tree-based models are a popular and effective method for feature selection (Xu et al., 2019), however its interpretation might vary depending on the assumptions taken over the metric used to calculate variable relative importance. Therefore the approach proposed by Lundberg and Lee (2017), which is based on Shapley values (Shapley, 1953) was used to provide a more comprehensive analysis.

In order to estimate the impact of road traffic on hospital admissions two approaches were taken: autoregressive integrated moving average models (ARIMA) and log-linear Poisson regression models. Both models have been extensively used not only for forecasting the evolution of time series in a wide range of fields such as environmental atmospheric (Díaz, García, Ribera, Alberdi, Hernández and Pajares, 1999; Navares et al., 2018) but also in studying the eventuality of epidemiological

* Corresponding author. Avda. Monforte de Lemos 5, 28029, Madrid, Spain.
E-mail address: J.diaz@isciii.es (J. Diaz).

https://doi.org/10.1016/j.envres.2020.109254
Received 13 January 2020; Received in revised form 10 February 2020; Accepted 12 February 2020
Available online 25 February 2020
0013-9351/

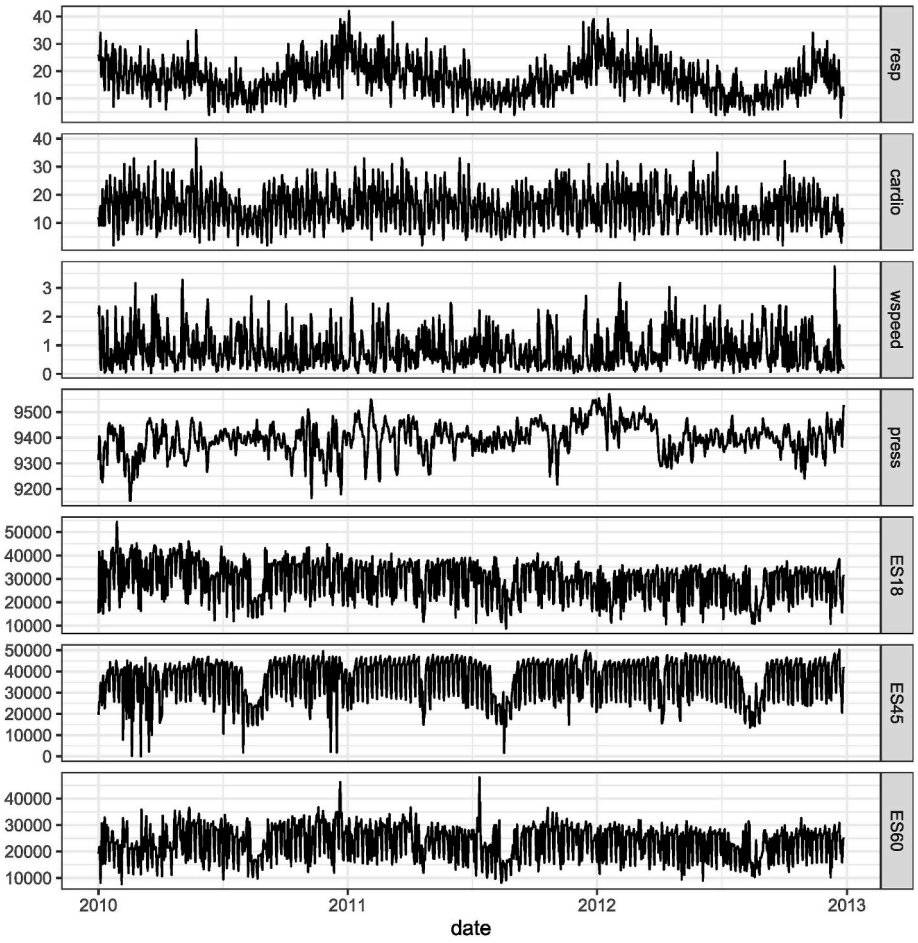


Fig. 1. Hospital admissions time series along with a sample of independent variables in this study.

diseases (Tobías et al., 2001; Linares and Díaz, 2010a, b).

2. Materials and methods

2.1. Data description

2.1.1. Target variables

Target variables consist of daily hospital admissions recorded at hospital Gregorio Marañon in Madrid due to respiratory (ICD-10: J00-J99) and cardiovascular diseases (ICD-10: I00–I99). Gregorio Marañon hospital covers the assistance of a population of 320.000 individuals which correspond to 12 primary care centers.

The study period lays between 01 and 01-2010 and 31-12-2012 (Fig. 1) and, due to data confidentiality laws, the exact origin of the patients is not provided. Consequently, independent variables source data is collected from surrounding areas assuming emergency cases reported far away are diverted to other hospitals.

2.1.2. Independent variables

Atmospheric conditions play an important role in the convection-diffusion process of pollutants (Li et al., 2017). Consequently, pressure and wind were included to represent the convection and advection processes respectively.

Wind data. Observations consist of hourly wind speed measures in m/s and wind direction in degrees. The data is provided by the Autonomous Community of Madrid at Plaza de España which is the

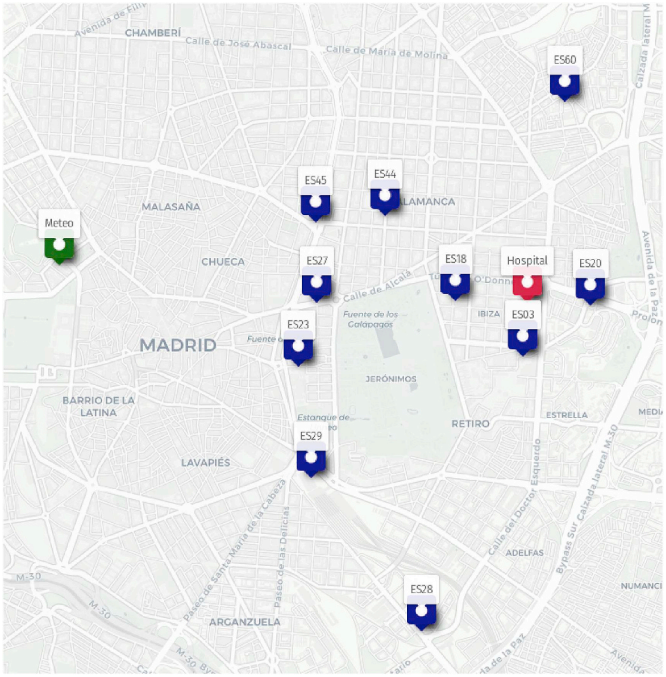


Fig. 2. Locations of the Hospital, the traffic stations and the weather station.

closest observation station to Gregorio Marañon hospital separated by 3 km.

Levels of immission are determined by pollutant dispersion processes in the atmosphere. Generally, these processes are driven by convection, which is pollutant dispersion to higher layers of the atmosphere, or advection which is the horizontal movement of the pollutants. Advection processes are mainly driven by wind while convection processes are driven by cyclonic (low pressure) or anticyclonic structures. Cyclonic structures are distinguished by the presence of upward currents which ease pollutant dispersion. Conversely, downward currents are characteristic of anticyclonic structures which hinder pollutant dispersion.

Pressure. Daily average pressure was provided by the Agencia Estatal de Meteorología (AEMET) at the observation station located in Retiro which is 700 m from Gregorio Maranõn Hospital. In order to consider a synoptic scale of meteorological changes, a variable $deltaPress = P_t - P_{t-1}$is defined being $P_t$the average daily pressure at time t. This variable serves to represent the trends which are related to more or less intense wind presence in the case of cyclonic ($deltaPress < 0$) or anticyclonic ($deltaPress > 0$) atmosphere respectively.

2.2. Methodology

The proposal consists in a first part to discriminate those variables which are not relevant, in terms of predictive capabilities, for both dependent variables estimation. With this first selection we simplify subsequent models and with their consequent gain in interpretability. As a second part, once the variables are filtered by importance, linear models are applied to analyze the independent selected variables attributable risk to the number of admissions. Firstly, a data preprocessing (Appendix A) was performed to eliminate collinearity and excessive correlations to prevent complications when applying variable selection and attribution models.

2.2.1. Variable selection

Linear models describe the relation between variables and the predictions since they have a single vector of coefficients. This interpretability eases diagnosis and it is clearly and advantage when compared to more complex computational intelligence models even though these last ones might better extract relevant information for prediction.

Nonetheless, tree-based computational intelligence models success in providing good predictive performance and interpretability. Among them, gradient boosting machines (GBM) Friedman (2001) is an ensemble technique which combines multiple weak learners to form a strong learner by additive training. At each iteration, a new weak learner (tree) is added to optimize the error function obtained by previous iteration fitted model. Even though tree based algorithms easily provide a way to extract variable importances, it is not always straightforward which assumption needs to be taken in order to obtain the importances. In tree-based models, variable importance can be interpreted as either the number of times a variable is used to split data across trees (weight) or the reduction gained in the loss function when that variable is used for splitting (gain). This decision might lead to different interpretations.

Lundberg and Lee (2017) introduce a unified approach based on Shapley values (Shapley, 1953) which describes the effect of each variable on the prediction of each data point by approximating the effect of eliminating a variable from the model. The Shapley (φ) value of a feature $x_i$is defined by

$$\varphi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!}\left(f_{S \cup \{i\}} - f_S\right), \tag{1}$$

where F denotes the set of all feature space, S a subset of F and $f_S$is the evaluation of the algorithm given a subset S of input variables (Ichiishi, 1983). Shapley values compare a prediction to a subset, which can be also composed by a single variable, instead of comparing it with the average prediction for the whole dataset allowing more contrastive explanation when compared to local surrogate models such as the local interpretable model-agnostic explanations LIME Ribeiro et al. (2016). Lundberg and Lee (2017) provides proof of the consistency and accuracy of using Shapley values in contrast to gain and weight approaches mentioned before.

2.3. Linear models

In order to estimate the impact of the total number of vehicles per day on hospital admissions due to respiratory and cardiovascular causes, both ARIMA models (Appendix B.1) and Log-linear Poisson regression models were used in this study. These two methodologies are comparable from the point of view of the quantification of the impact on health in normal distributions (Tobías et al., 2001).

2.3.1. Poisson regression models

Poisson distributions are a particularly useful theoretical model to study the contingency of epidemiological diseases. A random variable X representing the number of occurrences of an event happens in a period of time t, follows a probability Poisson distribution if complies with the following hypothesis with respect to the cumulative incidence of the disease: proportionality, stationarity and independence. Under these assumptions, the probability of an event k during a time period t for a random variable Y that follows a Poisson distribution is defined by

$$P(Y = k) = \mathrm{e}^{-\mu}.\frac{\mu^k}{k!}, \tag{2}$$

where μ represents the expected number of events during a period t (Pastor-Barriuso, 2012). One of the advantages of this type of models is that they allow to determine an estimation of the effect of certain event on health, taking into account ecological studies which use aggregated data of the population. This effect is known as relative risk (RR) which is represented in this study as the number of emergency admissions due to respiratory and cardiovascular disorders.

RR represents the difference in the risk of suffering the health event between exposed and unexposed individuals due to an increase in the unit of the corresponding independent variable. The linear regression model is constructed in which the probability of a count is determined by a Poisson distribution, where the average of the distribution is a function of the external independent variables (explanatory) as follows:

$$\ln(\hat{\mu}) = \beta_0 + \beta_1 x, \tag{3}$$

where x is the explanatory variable, $\beta_0$is the intersection and $\beta_1$the trend. Taking exponentials of both sides, it can be derived:

$$\hat{\mu} = \mathrm{e}^{\beta_0 + \beta_1 x} = \mathrm{e}^{\beta_0}.\mathrm{e}^{\beta_1 x}, \tag{4}$$

where e^{β}represents the RR for each correspondent variable.

The following covariables were included in the analysis, in order to control for the trend and seasonalities of the series, as well as the lags in the Theta mentioned before:

- Sine and Cosine functions of 365, 180, 120, 90 and 60 days to account for annual, six, four, three and two month periodicities.
- The trend of the series, using a counter (n1), which is 1 for the first day of the series, 2 for the second day, and so on, successively.
- Days of the week, using dummy variable.

The p-value was determined using the step-back procedure, in which the complete model that included all the analyzed explanatory variables, those concluded relevant out of the computational intelligence method applied in the first step of the analysis, was initially implemented, with those variables that individually showed less statistical significance gradually eliminated until concluding with a model

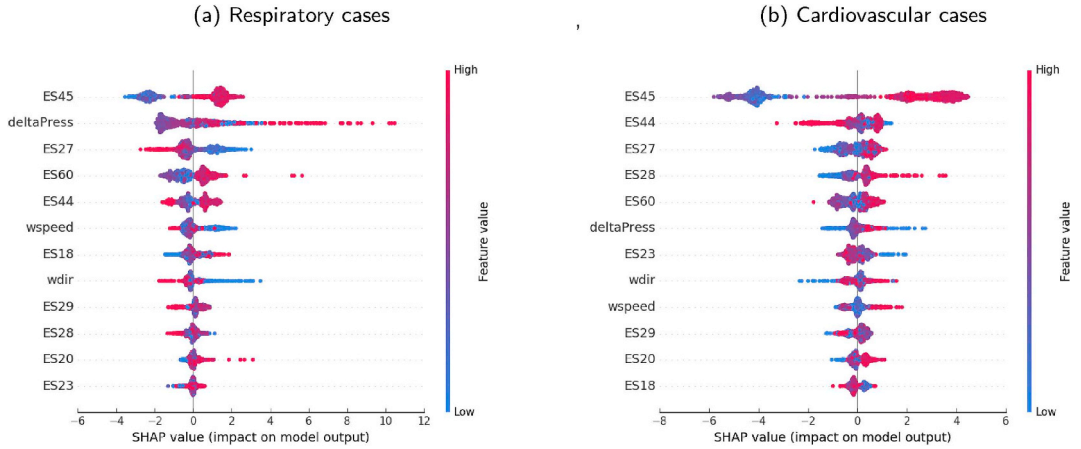


Fig. 3. GBM variable importance based on Shapley values for respiratory (a) and cardiovascular (b) cases.

that included just the statistically significant variables ($p < 0.05$). The percentage of population attributable risk (PAR) is calculated, based on RR, as follows $\%PAR = 100 \cdot [(RR - 1)/RR]$(Coste and Spira, 1991): representing the percentage of increment in emergency hospital admissions associated under the hypothesis of full population exposure. We have also to assume that all the other factors that might potentially influence should remain stable. All analyses were performed using the software IBM SPSS Statistics 22 and STATA v14.1.

3. Results

3.1. Variable selection

As we have seen in Appendix A the data obtained at ES03 was highly correlated (>70%) with the majority of the other observation points. A further collinearity analysis shows that removing ES03 drives not only the VIF with respect other observation points below the threshold of 5 as suggested by O' Brien (2007) (Table A2), but also its impact on the linear regression over the target variables is residual. Consequently, ES03 was removed from the study.

Fig. 3 (a) shows the Shapley values for respiratory cases. Among all traffic observation points, ES45 shows the higher overall influence where positive Shapley values (~ 2) correspond to high influx of patients due to this disorder. Conversely, Shpley values around −2 are related to low levels of admissions. Consequently, Table A1 shows the highest correlation for ES45. Anticyclonic atmosphere conditions (increase in pressure) show as an important factor clearly influencing high number of patient admissions due to respiratory cases.

On the other hand, atmospheric conditions seem to have less impact on the cardiovascular cases (Fig. 3 (b)), staying ES45 as the most influential among traffic observation points while being also the most correlated as low/high Shapley values correspond to low/high patient influx respectively.

Table 1
Explanatory variables obtained by the ARIMA model for the respiratory-related admissions.

		Estimations	Std. Error	*t*	Sig.
Non Seasonal	AR1	0.959	0.018	52,493	0
	MA1	0.864	0.03	28,499	0
Reg. Coefficients	Sine 365 days	2151	0.61	3528	0
	Cosine 365 days	4420	0.597	7399	0
	Sine 180 days	−1057	0.525	−2014	0.044
	Monday	4421	0.575	7684	0
	Tuesday	2841	0.592	4802	0
	Wednesday	3029	0.608	4981	0
	Thursday	1503	0.606	2481	0.013
	Friday	−1612	0.61	−2644	0.008
	Saturday	−3319	0.502	−6608	0
	LAGS (wspeed.2)	−0.428	0.228	−1875	0.061
	LAGS (deltaPress.3)	−0.007	0.004	−1854	0.064
	ES45 ([a])	0.056	0.022	2474	0.014
Constant		14,566	0.818	17,806	0

[a] In thousands.

3.2. Impact on hospital admissions

The results obtained for emergency hospital admissions due to respiratory causes through ARIMA and the calculation of risk using linear Poisson regression models are shown in Tables 1 and 2 respectively. Both models are consistent with respect to the resulting independent and control variables. Also, variables that control the annual and semiannual seasonality of the series, as well as the days of the week from Monday to Friday appear as significant. Regarding the independent variables, wind speed in two-days lag and the difference in atmospheric pressure in lagged three days are significant and with a negative coefficient, that is, at lower wind speed (less dispersion) and anticyclonic situations (greater atmospheric stability) they are related to an increase in the number of hospital admissions due to respiratory causes in the analyzed period.

Fig. 3 (a) shows that ES45 station was the most influential among all traffic observation points for respiratory cases. As can be seen in Table 1, ARIMA modeling for respiratory causes shows ES45 is significant with a positive coefficient of 0.056, meaning that for each thousand vehicles per day registered at this observation point, there is an absolute increment of 0.056 in admissions. Being the average daily number of admissions of 17.30 patients, the percentage of admissions would be 0.33% over 100 patients. Since the average daily traffic intensity at ES45 is 36.3 thousand vehicles, 12% of daily admissions due to respiratory cases are associated to the number of vehicles registered at this observation point.

Regarding the results of the Poisson modeling for the calculation of the risk, it is obtained that the ES45 station presents an increase in the relative risk, IRR = 1.004 [1.003, 1.006]; If we apply the equation to calculate the population attributable risk (PAR), we obtain that PAR = 0.40%. This result represents and increase of 0.40% patients per thousand vehicles per day. With the average of 36.3 thousand vehicles per day, it can be said that 14.5% of emergency cases due to respiratory cases are attributable to traffic intensity at ES45.

Tables 3 and 4 show the results of the ARIMA and the poisson regression models for cardiovascular cases respectively. Annual, semiannual, quarterly and bi-monthly seasonality patterns have statistical significance as well as weekdays and Saturday. Fig. 3 (b) again shows

Table 2
Poisson regression for respiratory admissions.

Poisson regression					Num. Obs.: 1088	
					LR χ^2_{13}: 1511.27	
					Prob. >χ^2: 0	
Log. Lik. −3173.861					Pseudo R2: 0.1923	
	IRR	Std. Error	z	P>|z|	[95% Conf. Interval]	
LAGS (resp, 1)	1.0088	0.00141	6.25	0	1.0060	1.0116
N1 (trend)	0.9999	0.00002	−3.72	0	0.9999	1.0000
Sine 365 days	1.1174	0.01265	9.81	0	1.0929	1.1425
Sine 180 days	0.9331	0.01015	−6.36	0	0.9134	0.9532
Cosine 365 days	1.2633	0.01581	18.68	0	1.2327	1.2947
ES45 ([a])	1.0045	0.00096	4.66	0	1.0026	1.0064
LAGS (wspeed.2)	0.9673	0.01153	−2.79	0.005	0.9449	0.9901
LAGS (deltaPress.3)	0.9996	0.00020	−2.05	0.041	0.9992	1.0000
Monday	1.2135	0.02779	8.45	0	1.1602	1.2692
Tuesday	1.0661	0.02570	2.65	0.008	1.0169	1.1177
Wednesday	1.0885	0.02605	3.54	0	1.0386	1.1408
Friday	0.8426	0.02186	−6.6	0	0.8008	0.8865
Saturday	0.7685	0.02063	−9.81	0	0.7291	0.8100
Constant	13.2985	0.53997	63.73	0	12.2812	14.4001

[a] In thousands.

Table 3
Explanatory variables obtained by the ARIMA model for the cardiovascular admissions.

		Estimations	Std. Error	t	Sig.
Non Seasonal	AR1	0.09	0.292	0.307	0.759
	MA1	−0.015	0.293	−0.052	0.958
Reg. Coefficients	N1 (trend)	−0.001	0	−2083	0.038
	Sine 365 days	0.714	0.217	3293	0.001
	Cosine 365 days	0.59	0.209	2814	0.005
	Cosine 180 days	−0.469	0.207	−2265	0.024
	Sine 120 days	0.792	0.216	3663	0
	Cosine 120 days	−0.584	0.21	−2778	0.006
	Sine 60 days	0.441	0.209	2107	0.035
	Monday	6714	0.536	12,519	0
	Tuesday	4397	0.571	7703	0
	Wednesday	3705	0.586	6322	0
	Thursday	2208	0.584	3780	0
	Friday	−1451	0.586	−2475	0.013
	Saturday	−5245	0.472	−11102	0
	ES45 ([a])	0.17	0.02	8390	0
Constante		8561	0.644	13,299	0

[a] In thousands.

ES45 as the most influential traffic intensity observation point among all considered. ES45 station is also positive associated with cardiovascular cases according to the ARIMA (Table 3), although the coefficient is higher (0.17) when compared to the number of respiratory cases. This coefficient of 0.17 corresponds to 39.3% of daily admissions due to cardiovascular disorders which are attributable to the number of vehicles registered in ES45.

The results of the estimation of attributable impact using log-linear Poisson regression (Table 4) obtain an IRR = 1.012 [1.009, 1.014] with a PAR or 1.19%. Every thousand vehicles per day registered at ES45 increases hospital admissions due to cardiovascular cases by 1.19%. Being the daily average at ES45 of 36.3 thousand vehicles a day, the percentage of cardiovascular causes at Gregorio Marañon hospital attributable to traffic intensity is 43%.

Table 4
Poisson regression for cardiovascular admissions.

Poisson regression					Num. Obs.: 1090	
					LR χ^2_{15}: 1746.19	
					Prob. >χ^2: 0	
Log. Lik.-3083.0027					Pseudo R2: 0.2207	
	IRR	Std. Error	z	P>|z|	[95% Conf. Interval]	
LAGS (cardio, 1)	1.0052	0.00173	3.02	0.003	1.0018	1.0086
N1 (trend)	0.9999	0.00003	−2.50	0.012	0.9999	1.0000
Sine 365 days	1.0420	0.01187	3.61	0.000	1.0190	1.0655
Cosine 365 days	1.0388	0.01149	3.45	0.001	1.0166	1.0616
Cosine 180 days	0.9721	0.01054	−2.61	0.009	0.9516	0.9930
Sine 120 days	1.0466	0.01198	3.98	0.000	1.0234	1.0704
Cosine 120 days	0.9677	0.01077	−2.95	0.003	0.9468	0.9890
Sine 60 days	1.0282	0.01125	2.54	0.011	1.0064	1.0505
Monday	1.4432	0.04798	11.03	0.000	1.3521	1.5403
Tuesday	1.2204	0.04843	5.02	0.000	1.1291	1.3191
Wednesday	1.1873	0.04573	4.46	0.000	1.1010	1.2804
Thursday	1.0921	0.04226	2.28	0.023	1.0123	1.1782
Friday	0.8662	0.03404	−3.66	0.000	0.8020	0.9355
Saturday	0.5787	0.02217	−14.27	0.000	0.5369	0.6239
ES45 ([a])	1.0118	0.00128	9.27	0.000	1.0093	1.0143
Constant	9.0576	0.36176	55.17	0.000	8.3757	9.7952

[a] In thousands.

4. Discussion

Even though the effects of pollution concentrations in the air and its influence on human health has been thoroughly studied, the relation between main pollution source in cities and related health disorders was not previously established. As we have seen in Section 3 road traffic relates to 13% of respiratory cases and 39% of cardiovascular disorders.

Variable selection through permutation over feature importance in tree-based models creates an interpretable output without the need to apply any transformation to the variable involved. However, it was mentioned the instability of the results since permutation adds randomness to the measurement, especially in the presence of highly

correlated variables. In order to control these situations, a previous collinearity analysis was included along with tree-based variant of Shapley values computation proposed by Lundberg et al. (2018).

Among the limitations of this proposal, there are those inherent in every longitudinal ecological study that prevent extrapolating the results at individual levels. On the other hand, averaged data from several vehicle counting points have been used, therefore, these measures do not represent an individual exposure. However, the methodology applied is common to studies in which the impact on health is analyzed through data from air pollution measurement stations (Samet et al., 2000). These biases are minimized by including in the control variable models such as trend, seasonality and autoregressive factor of the series. Finally, as in all studies that analyze the effect of pollution on health variables there is a misalignment problem (Ingebrigtsen, 2015).

Even though previously cited studies show a clear relation between primary pollutant levels immission and road traffic in large cities, atmosphere also plays an important role. In this study, the role of atmosphere was included through convection-diffusion process, via the variable deltaPress, and advection via wind speed (wspeed).

The results obtained in this study are inline with respect to the role atmospheric conditions play in pollutant diffusion (Li et al., 2017). Specifically, negative trends in pressure, as those obtained in the proposed models, represent a cyclonic tendency which is distinguished by the altitude of the mixing layer and, consequently, with a better dispersion of pollutants and lower levels of immission (Li et al., 2017). As a result, there is a decrease in hospital admissions (Díaz et al., 1999; Linares and Díaz, 2010a, b).

On the other hand, the negative sign obtained in wind influence on hospital admissions, represents lower levels of immission, therefore, lower number of hospital admissions (Díaz et al., 1999; Linares and Díaz, 2010a, b).

From a quantitative point of view, the influence of traffic is 13% per thousand vehicles in the case of respiratory disorders, and 39.3% in the case of circulatory. These results are inline when compared to the influence of chemical air pollutants and noise levels on hospital admissions where, the impact is higher on circulatory cases than in respiratory cases (Díaz et al., 1999; Díaz et al., 2001; Tobías et al., 2001; Linares and Díaz, 2010a,b; Recio et al., 2016). It should also be emphasized that the percentages obtained in this study refer to traffic intensity in contrast to pollution levels, either chemical or noise. Therefore, the results are not directly comparable to the aforementioned researches or other study focused on pollution levels. Notwithstanding, the World Health Organization (WHO Regional Office for Europe, 2013) [1] establishes that air pollution-causes deaths are 40% due to ischaemic heart disease, 40% due to stroke and 11% due to chronic obstructive pulmonary disease (COPD). Similar proportions were found in this study.

5. Conclusions

In this paper, a novel approach to quantify the effect of urban traffic on respiratory and cardiovascular diseases is presented. Although pollution-related effects on health have been extensively studied, the direct influence of what is considered a main driver of pollution (traffic) was not previously established.

Indeed, our results show that traffic is responsible for emergency hospital admissions: 13% of respiratory cases are related to traffic intensity while, in the case of cardiovascular disorders, the percentage increases to 39%.

These results represent a step forward in the understanding of how human health in contemporary cities is threatened by the way in which they are organized: concretely, it becomes clear that traffic is a major cause of respiratory and cardiovascular diseases. As a consequence, raising awareness about the risks of high traffic levels should arguably be a priority for urban institutions, which should put public health in the center of how the cities are understood and managed.

Traffic intensity. Hourly traffic intensity was provided by the Madrid Municipal Traffic Grid which consists of 4079 electromagnetic sensors placed under the pavement to detect vehicles mass that pass over the system. Instead of using the full grid, the 10 observations composing a 20 km radius which surround the hospital were selected. Records are provided as the number of cars per hour which are aggregated at each location (Fig. 2) to obtain the total number of cars per day.

Acknowledgment

The authors gratefully acknowledge grants ENS-UNED from the IMIENS.

Appendix A. Data preprocessing

Table A.1
Correlation matrix of model variables

		correlations														
	Variable	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.	13.	14.	15.
1.	ES03	1														
2.	ES18	0.85	1													
3.	ES20	0.65	0.64	1												
4.	ES23	0.38	0.18	0.27	1											
5.	ES27	0.04	−0.02	0.12	0.53	1										
6.	ES28	0.76	0.84	0.59	0.12	−0.03	1									
7.	ES29	0.71	0.72	0.6	0.44	0.25	0.68	1								
8.	ES44	0.38	0.28	0.35	0.4	0.41	0.25	0.38	1							
9.	ES45	0.52	0.48	0.42	0.33	0.31	0.42	0.49	0.72	1						
10.	ES60	0.84	0.76	0.64	0.38	0.11	0.74	0.72	0.43	0.52	1					
11.	Wspeed	−0.02	−0.02	−0.01	0.04	−0.04	0	0.01	−0.05	0	−0.02	1				
12.	Wdir	−0.02	0	−0.05	0	−0.05	−0.01	−0.03	−0.03	−0.07	−0.02	0.07	1			
13.	pressAvg	−0.05	−0.08	0.01	0.15	0.24	−0.06	0.05	0.15	0.09	0.01	−0.24	−0.26	1		
14.	Resp	0.25	0.22	0.16	0.08	−0.07	0.22	0.16	0.2	0.3	0.25	−0.05	−0.09	0.07	1	
15.	cardio	0.34	0.33	0.26	0.14	0.12	0.35	0.28	0.34	0.5	0.33	0.03	−0.03	0.03	0.42	1

[1] https://www.who.int/mediacentre/news/releases/2014/air-pollution/en/

Table A.2
Stepwise VIF values at each iteration and impact on linear regression statistics.

	VIF	
Variable	Step 1	Step 2
ES03	6.22	–
ES18	6.09	4.79
ES20	1.98	1.96
ES23	2.10	1.86
ES27	1.74	1.66
ES28	3.98	3.96
ES29	3.10	3.10
ES44	2.42	2.41
ES45	2.56	2.55
ES60	4.18	3.53
wspeed	1.10	1.09
wdir	1.09	1.08
pressAvg	1.37	1.36
deltaPress	1.12	1.12
Reg. cardiovascular		
Residual SE	5.527	5.526
R2	0.27	0.27
	Reg. respiratory	
Residual SE	6.254	6.254
R2	0.13	0.14

Wind is usually reported as two quantities, speed in m/s and direction in degrees 0–359 where 0 represents wind blowing from the North. Since data is provided at hourly level, in order to aggregate (via average) at daily granularity its vector components $\overrightarrow{u}$ and $\overrightarrow{v}$ which represent the east-west and the north-south components respectively (Glickman, 2000) and are defined by

$$\overrightarrow{u_i} = -u_i \sin\left(2\pi \frac{\theta_i}{360}\right), \quad \overrightarrow{v_i} = -u_i \cos\left(2\pi \frac{\theta_i}{360}\right), \tag{A.1}$$

where $u_i$is the wind speed at time i and $\theta_i$the angle in degrees. These components can be averaged to obtain their daily levels: $\overrightarrow{u_t} = \frac{1}{24}\sum_{i=0}^{23} \overrightarrow{u_i}$and $\overrightarrow{v_t} = \frac{1}{24}\sum_{i=0}^{23} \overrightarrow{v_i}$. Daily vector average wind speed at time t becomes $\text{wspeed}_t = (\overrightarrow{u_t}^2 + \overrightarrow{v_t}^2)^{\frac{1}{2}}$and average wind direction is defined by

$$\text{wdir}_t = \arctan\left(\frac{\overrightarrow{u_t}}{\overrightarrow{v_t}}\right) + C, \tag{A.2}$$

where $C = 180$if $\left(\frac{\overrightarrow{u_t}}{\overrightarrow{v_t}}\right) < 180$and $C = -180$otherwise.

Collinearity or excessive correlation among variables is required to be checked to prevent complications when identifying and optimal subset of explanatory variables. High correlation and severe multicollinearity among predictors might result in instability of the coefficient estimates since confidence intervals for coefficients tend to be very wide and, as a consequence, makes models difficult to interpret as they lose statistical significance. Table A1 shows the correlation among variables. It can be clearly seen some pairs of highly correlated variables such as ES03 and ES18 which are correlated at 85% or ES03 and ES60 with a correlation of 84% which, on the other hand it might be caused by the imputation method.

In order to examine for multicollinearity, a widely-used diagnostic called variance inflation factor (VIF) (Graham, 2003) is calculated for each predictor by performing a linear regression of the predictor on the remaining other to obtain the R_2, being the VIF defined by $VIF_i = \frac{1}{1-R_i^2}$, where $R_i^2$is the R^2-value obtained by regressing the i^{th}predictor on the remaining predictors. As a rule of thumb, VIF values in excess of 5 or 10 are used as an indicator of multicollinearity (Mason and Gunst, 2003) although some studies warn about using a cutoff value of 10 (O' Brien, 2007). Consequently, in this study the threshold to consider severe multicollinearity will be set at 5.

The stepwise procedure consists of calculating the VIF values iteratively, at each step the predictor variable with highest VIF is removed to subsequently recalculate the VIF until all predictors show a value lower than 5. Table A2 shows the VIF at each step along with the linear regression statistics on each target variable. It can be clearly seen a very limited impact in the residual standard error and the R_2 for each respiratory and cardiovascular regressions with a reduced set of 13 variables (Table A2) compared to the initial set of 14. this is an appendix.

Appendix B. Linear models

B.1. Arima

The acronym ARIMA stands for Auto-Regressive Integrated Moving Average. The ARIMA forecasting equation for a stationary time series is a linear (i.e., regression-type) equation in which the predictors consist of lags of the dependent variable and/or lags of the forecast errors. That is, predicted value of Y equals a constant (μ)and/or a weighted sum of one or more recent values of Y ($Y_{t-1}, \cdots, Y_{t-p}$)and a weighted sum of recent values of the errors ($e_{t-1}, \cdots, e_{t-q}$).

Lags of the stationary series in the forecasting equation are called “autoregressive” terms, lags of the forecast errors are called “moving average” terms, and a time series which needs to be differenced to be made stationary is said to be an “integrated” version of a stationary series. A nonseasonal ARIMA model is noted as an “ARIMA (p,d,q)” model, where.

- p is the number of autoregressive terms,
- d is the number of nonseasonal differences needed for stationarity, and

- q is the number of lagged forecast errors in the prediction equation.

In terms of y, the general forecasting equation is:

$$\hat{y}_t = \mu + \varphi_1 y_{t-1} + ...+\varphi_p y_{t-p} - \theta_1 e_{t-1} - ...-\theta_p e_{t-p}, \quad \text{(B.1)}$$

where $\varphi_i$is the coefficient of the autoregressive (AR) term i and $\theta_i$represents the coefficient of the moving average (MA) term i. ARIMA models with exogenous variables (Makridakis et al., 1983) include the values of these variables $(X...Z)$with their correspondent lag $(s...m)$along with the dependent variable Y, its lags (Y_{t-p}), the errors (e) and its lags (e_{t-q}) resulting in the following equation:

$$\begin{aligned}\hat{y}_t = \mu \quad & + \varphi_1 y_{t-1}+...+\varphi_p y_{t-p} \\ & - \theta_1 e_{t-1}-...-\theta_p e_{t-p} \\ & + \beta_0 X_t + \beta_1 X_{t-1}+...+\beta_s X_{t-s}+... \\ & + \gamma_0 Z_t + \gamma_1 Z_{t-1}+...+\gamma_m Z_{t-m}. \end{aligned} \quad \text{(B.2)}$$

The value of the estimators $\beta_0, ...,\beta_s$and $\gamma_0, ...,\gamma_m$of the variables that are significant at $p < 0.05$(p-value provided by SPSS v15) indicating increased Y to increment by one unit of each independent variable $(X,...,Z)$respectively. The model's goodness-of-fit was obtained by analysis of residuals (AIC, BIC, ACF, Box-Ljung).

References

Burns, J., Boogaard, H., Polus, S., Pfadenhauer, L., Rohwer, A., van Erp, A., Turley, R., Rehfuess, E., 2020. Interventions to reduce ambient air pollution and their effects on health: an abridged cochrane systematic review. Environ. Int. 135, 105400. https://doi.org/10.1016/j.envint.2019.105400. URL: http://www.sciencedirect.com/science/article/pii/S0160412019322056.

Carmona, R., Linares, C., Recio, A., Ortiz, C., Díaz, J., 2018. Emergency multiple sclerosis hospital admissions attributable to chemical and acoustic pollution: madrid (Spain), 2001–2009. Science of The Total Environment 612 111–118. https://doi.org/10.1016/j.scitotenv.2017.08.243. URL: http://www.sciencedirect.com/science/article/pii/S0048969717322519.

Coste, J., Spira, A., 1991. Le proportion de cas attributable en santé publique: definition (s), estimation(s) et interpretation. Rev. Epidemiol. Sante Publique 51, 399–411.

de Miguel-Díez, J., Hernández-Vázquez, J., López-de Andrés, A., Alvaro-Meca, A., Hernández-Barrera, V., Jiménez-García, R., 2019. Analysis of environmental risk factors for chronic obstructive pulmonary disease exacerbation: a case-crossover study (2004-2013). PloS One 14, 1–11. https://doi.org/10.1371/journal.pone.0217143. URL:10.1371/journal.pone.0217143.

Díaz, J., García, R., Ribera, P., Alberdi, J.C., Hernández, E., Pajares, M.S., 1999. Modeling of air pollution and its relationship with mortality and morbidity in madrid (Spain). Int. Arch. Occup. Environ. Health 75, 366–376.

Díaz, J., Alberdi, J.C., Pajares, M.S., López, R., López, C., Otero, A., 2001. A model for forecasting emergency hospital admissions: effect of environmental variables. J. Environ. Health 64, 9–15.

Díaz, J., López-Bueno, J., López-Ossorio, J., González, J., Sánchez, F., Linares, C., 2020. Short-term effects of traffic noise on suicides and emergency hospital admissions due to anxiety and depression in madrid (Spain). Science of The Total Environment 710 136315. https://doi.org/10.1016/j.scitotenv.2019.136315. URL: http://www.sciencedirect.com/science/article/pii/S0048969719363119.

EEA, 2020. Health Impacts of Air Pollution. https://www.eea.europa.eu/themes/air/health-impacts-of-air-pollution.

Friedman, J.H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29, 1189–1232.

Glickman, T.S., 2000. Glossary of Meteorology. American Meteorological Society (AMS), The Council of AMS.

Graham, M., 2003. Confronting multicollinearity in ecological multiple regression. Ecology 84, 2809–2815.

Ichiishi, T., 1983. Game Theory for Economic Analysis. Economic Theory, Econometrics and Mathematical Economics Series. Academic Press.

Ingebrigtsen, R., 2015. Bayesian Spatial Modelling of Non-stationary Processes and Misaligned Data Utilising Markov Properties for Computational Efficiency.

Krzyzanowski, M., Kuna-Dibbert, B., Schneider, J., 2005. Health effects of transport-related air pollution. WHO Regional Office Europe.

Li, Z., Guo, J., Ding, A., Liao, H., Liu, J., Sun, Y., Wang, T., Xue, H., Zhang, H., Zhu, B., 2017. Aerosol and boundary-layer interactions and impact on air quality. URL:. Nat. Sci. Rev. 4, 810–833. https://doi.org/10.1093/nsr/nwx117.. arXiv:10.1093/nsr/nwx117http://oup.prod.sis.lan/nsr/article-pdf/4/6/810/23827203/nwx117.pdf.

Linares, C., Díaz, J., 2010a. Short-term effect of concentrations of fine particulate matter on hospital admissions due to cardiovascular and respiratory causes among the over-75 age group in madrid, Spain. Publ. Health 124, 28–36. https://doi.org/10.1016/j.puhe.2009.11.007. URL: http://www.sciencedirect.com/science/article/pii/S0033350609003564.

Linares, C., Díaz, J., 2010b. Short-term effect of pm2.5 on daily hospital admissions in madrid (2003–2005). 129–140. URL:. Int. J. Environ. Health Res. 20, 20169485. https://doi.org/10.1080/09603120903456810. arXiv:10.1080/09603120903456810. pMID.

Lundberg, S.M., Lee, S.I., 2017. A unified approach to interpreting model predictions. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems 30. Curran Associates, Inc., pp. 4765–4774.

Lundberg, S.M., Erion, G.G., Lee, S., 2018. Consistent Individualized Feature Attribution for Tree Ensembles. CoRR abs/1802.03888. URL:. arXiv:1802.03888. http://arxiv.org/abs/1802.03888.

Makridakis, S., Wheelwright, S., Mcgee, V., 1983. Forecasting Methods and Applications. Wiley, San Francisco.

Mannucci, P.M., Harari, S., Franchini, M., 2019. Novel evidence for a greater burden of ambient air pollution on cardiovascular disease. URL:. Haematologica 104, 2349–2357. https://doi.org/10.3324/haematol.2019.225086.. arXiv:http://www.haematologica.org/content/104/12/2349.full.pdf. http://www.haematologica.org/content/104/12/2349.

Marques-Mejías, M., Tomás-Pérez, M., Hernández, I., López, I., Quirce, S., 2018. Asthma exacerbations in the pediatric emergency department at a tertiary hospital: relationship with environmental factors. J. Invest. Allergol. Clin. Immunol. 29. https://doi.org/10.18176/jiaci.0364.

Mason, R., Gunst, R., 2003. Statistical Design and Analysis of Experiments: with Applications to Engineering and Science. John Wiley & Sons.

Navares, R., Díaz, J., Linares, C., Aznarte, J., 2018. Comparing arima and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in madrid. Stoch. Environ. Res. Risk Assess. 1–11doi. https://doi.org/10.1007/s00477-018-1519-z.

O' Brien, R., 2007. A caution regarding rules of thumb for variance inflation factors. Qual. Quantity 41, 673–690. https://doi.org/10.1007/s11135-006-9018-6.

Pastor-Barriuso, R., 2012. Bioestadística. Centro Nacional de Epidemiología, Instituto de SaludCarlos III.

Quero, X., Viana, M., Moreno, T., Alastuey, A., 2012. Bases científico-técnicas para un plan nacional de mejora de la calidad del aire. Informes CSIC.

Recio, A., Linares, C., Banegas, J., Díaz, J., 2016. The short-term association of road traffic noise with cardiovascular, respiratory, and diabetes-related mortality. Environ. Res. 150, 383–390. https://doi.org/10.1016/j.envres.2016.06.014.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. why should I trust you? CoRR abs/1602.04938. URL:. Explaining the Predictions of Any Classifier. arXiv:1602.04938. http://arxiv.org/abs/1602.04938.

Samet, J., Dominici, F., Zeger, S., Schwartz, J., Dockery, D., 2000. The National Morbidity, Mortality, and Air Pollution Study. Part I: Methods and Methodologic Issues. Research report (Health Effects Institute) , 5–14; discussion 75–84URL:. . http://europepmc.org/abstract/MED/11098531.

Shapley, L.S., 1953. A value for n-person games. In: Kuhn, H.W., Tucker, A.W. (Eds.), Contributions to the Theory of Games II. Princeton University Press, Princeton, pp. 307–317.

Tobías, A., Díaz, J., Saez, M., Carlos Alberdi, J., 2001. Use of Poisson regression and box–jenkins models to evaluate the short-term effects of environmental noise levels on daily emergency admissions in madrid, Spain. Eur. J. Epidemiol. 17, 765–771.

Who Regional Office for Europe, R., 2013. Review of evidence on health aspects of air pollution – REVIHAAP Project: technical Report. (Technical Report).

Xu, Z.E., Huang, G., Weinberger, K.Q., Zheng, A.X., 2019. Gradient Boosted Feature Selection. CoRR abs/1901.04055. URL:. , 04055 arXiv:1901. http://arxiv.org/abs/1901.04055.

Chapter 12

Side project II: Geographical imputation of missing pollen data via convolutional neural networks

Type:	Published Article
Title:	*Geographical imputation of missing pollen data via convolutional neural networks*
Journal:	Atmosphere
Authors:	Ricardo Navares & José Luis Aznarte
Published:	November 2019
Impact Factor:	2.046
Quartile:	Q3
DOI	10.3390/atmos10110717

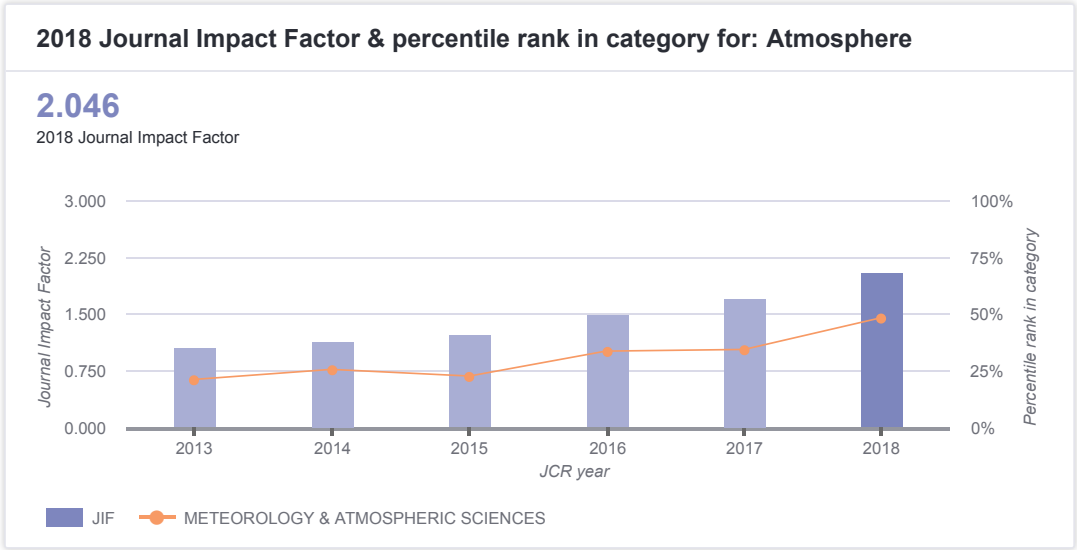


FIGURE 12.1: Impact factor Atmosphere

atmosphere

MDPI

Article

Geographical Imputation of Missing Poaceae Pollen Data via Convolutional Neural Networks

Ricardo Navares and José Luis Aznarte *

Department of Artificial Intelligence, UNED, Juan del Rosal, 16, 28040 Madrid, Spain; navares.ricardo@gmail.com

* Correspondence: jlaznarte@dia.uned.es

Received: 4 October 2019; Accepted: 14 November 2019; Published: 16 November 2019

Abstract: Airborne pollen monitoring datasets sometimes exhibit gaps, even very long, either because of maintenance or because of a lack of expert personnel. Despite the numerous imputation techniques available, not all of them effectively include the spatial relations of the data since the assumption of missing-at-random is made. However, there are several techniques in geostatistics that overcome this limitation such as the inverse distance weighting and Gaussian processes or kriging. In this paper, a new method is proposed that utilizes convolutional neural networks. This method not only shows a competitive advantage in terms of accuracy when compared to the aforementioned techniques by improving the error by 5% on average, but also reduces execution training times by 90% when compared to a Gaussian process. To show the advantages of the proposal, 10%, 20%, and 30% of the data points are removed in the time series of a Poaceae pollen observation station in the region of Madrid, and the airborne concentrations from the remaining available stations in the network are used to impute the data removed. Even though the improvements in terms of accuracy are not significantly large, even if consistent, the gain in computational time and the flexibility of the proposed convolutional neural network allow field experts to adapt and extend the solution, for instance including meteorological variables, with the potential decrease of the errors reported in this paper.

Keywords: Poaceae pollen; spatial imputation; convolutional neural networks

1. Introduction

The clinical relevance of Poaceae pollen has been increasing as the number of allergy cases continues to grow [1], which is expected to double in the next 40 years [2]. Limiting exposure to airborne pollen plays a key role in the prevention of symptoms. The prediction of future pollen concentrations is thus crucial, not only for patients, but also for clinical institutions, in order to arrange resources before the influx of pollen related allergy cases.

Observation based models employ different methods to relate records of air concentrations to one or more variables that can be measured or predicted. Examples include regression models [3,4], time series models [5], and process based phenological models [6]. In the last decade, machine learning techniques have been gaining importance due to the success of their applications [4,7–12]. However, these techniques require a significant amount of data, and when dealing with pollen time series, where high concentrations are especially harmful when they are over 25 grains/m^3 [1], the data are incomplete during the full year (Figure 1). Even though there have been advances in automatic pollen monitoring [13], the European volumetric spore trap network is mostly operated manually. Furthermore, defects and maintenance imply that pollen observation networks are highly sensitive to missing data points.

Despite the numerous techniques available for missing data imputation, most of the literature focuses on the conditions under which they lead to unbiased estimates, conditions that do not often hold. There is no consensus about the exact proportion of missing data for which it is considered unacceptable to use such techniques. For instance, Schafer [14] asserted that less than 5% is inconsequential, while Bennett [15] claimed that statistical analysis is likely to be biased over 10%. In this proposal, we use 10%, 20%, and 30% of missing data, which are greater amounts than asserted by previous literature. Moreover, many techniques do not take into consideration the spatial relations of the data. Geographical imputation overcomes this problem by estimating missing data points with approximate locations derived from associated data. Among the techniques available, inverse distance weighting [16] and kriging or Gaussian process regression [17] are two of the most popular among field experts.

However, in the last few decades, artificial intelligence methods have been gaining attention due to their competitive advantage in solving real-world problems [18]. In particular, convolutional neural networks (CNNs) [19] have been proven very effective in areas such as computer vision [20] and natural language processing [21]. The main difference from traditional neural networks lies in using the convolution operation [22] applied to filters, which allows exploiting the strong, spatial correlation present in the data.

Even though there is extensive literature about computational intelligence techniques applied to pollen time series, such as random forests [7,12,23,24], artificial neural networks [9,10], and deep neural architectures [25], very few works have applied convolutional neural networks to time series. Nonetheless, CNNs have been extensively used in identifying and classifying pollen grains [26,27].

The objective of this paper is to extend the application of CNNs and increase the awareness of their advantage and potential. In order to do so, we propose a network architecture that will be compared to the aforementioned traditional spatial imputation techniques. By artificially producing missing data points in the time series of one of the observation stations in the region of Madrid, the study estimates such points from the available observations from the surrounding stations.

2. Materials and Methods

2.1. Data Description

Poaceae pollen observations were provided daily in grains per cubic meter registered at eight locations in or around the city of Madrid: Alcalá de Henares, Alcobendas, Aranjuez, Complutense University of Madrid (Pharmacy Faculty), Coslada, Getafe, Leganés, and Villalba. Series for these locations are shown in Figure 1. Pollen counts followed the standard methodology of the Spanish Aerobiological Network [28] and were provided by Red Palinológica de la Comunidad de Madrid. Observations were available for 14 years starting from 1 January 2000 to 31 December 2013.

The region of Madrid has particular geographical characteristics (Figure 1). The observation station in Aranjuez is at the lowest elevation (495 m above sea level) and has a yearly average temperature above 14 °C with a yearly average rainfall below 400 mm. On the other hand, Villalba is located 903 m above sea level with a yearly average temperature of 10–11 °C and a yearly average precipitation of 1250–1500 mm. The remaining locations are in metropolitan areas between 594 and 668 m above sea level with yearly average temperatures above 15.2 °C and precipitation around 440 mm.

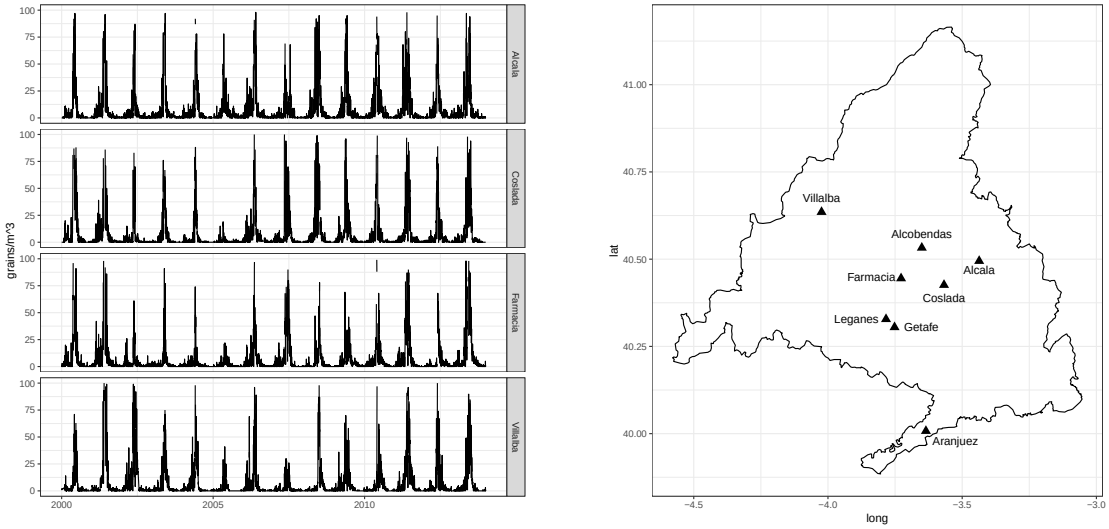


Figure 1. Selected stations with Poaceae pollen concentrations capped at 100 grains/m^3 (**left**). Distribution of the locations in the region of Madrid (**right**).

2.2. Methodology

Inverse distance weighted (IDW) interpolation is based on the principle that nearer observations are more related than distant ones [29]. Consequently, pollen counts measured closer to the location of the station for which we want to estimate the pollen counts will have more influence than those that are distant. This influence is represented by the distance, and the estimation $\hat{y}_j$ is calculated as the weighted sum of the measured pollen counts at the observation stations x_i:

$$\hat{y}_j = \frac{\sum_{i=1}^{n} \left(\frac{x_i}{d_{ij}^p} \right)}{\sum_{i=1}^{n} \left(\frac{1}{d_{ij}^p} \right)}, \quad (1)$$

where n is the number of observation stations, d_{ij} is the distance between the observation station i and the observation station j where we want to estimate the pollen count, and p is the power function, which is set to 2 as the default value.

While in IDW, the power function defines how fast the influence (weight) of an observed pollen count measure decreases based on distance, a Gaussian process or kriging [17] creates a model of spatial correlation that provides the proper weights by relying on the covariance matrix to control the values that are close together in the input space to generate values that are similar. A Gaussian process (GP) assumes that the probability $p(f(x_1), \ldots, f(x_n))$ is jointly Gaussian, x_i being the set of observed points, with mean μ and covariance given by $\sum_{ij} = k(x_i, x_j)$, where k is the kernel function [30]. The underlying idea is that having the joint probability of the variables, it is possible to get the conditional probability of one of the variables given the others [31].

Based on their success in other fields, experts have started to use convolutional neural networks for time series analysis [32]. CNNs differ from feedforward neural networks mainly by the existence of convolutional layers, which are hidden layers that utilize the power of the mathematical operation of convolution to transform the inputs. Convolution allows for the encoding of the local properties of the input in such a way that the information propagates in a more efficient manner. CNN filters, obtained by the convolution of inputs and weights, are local in input space and are thus able to exploit the strong, spatial correlation present in the time series. That means that they work well in identifying simple patterns within local regions of the data (subset of features), which then will be used by subsequent layers to form more complex patterns.

In order to compare the results with similar research studies, the common scoring rule of the root mean squared error (RMSE) will be used to measure the average magnitude of the error.

2.3. Experimental Design

The aim was to compare the aforementioned methods in order to see their ability to impute data properly for the observation station in Farmacia (most central location available), by inferring airborne pollen counts at one station based on the levels measured in its surroundings. In order to do so, series with 10%, 20%, and 30% of missing data points were generated and then used as a test set to check the estimations. However, as we can see in Figure 1, airborne pollen time series are a particular kind of series where during most of the year, pollen counts are either nonexistent or very low. For this reason, a stratified random sample was drawn based on the criteria of an observation belonging to a peak or off-peak pollen season.

There is no general consensus about the definition of the pollen season, and hence, season dates might differ according to their definition [11]. This notwithstanding, in Spain, the first symptoms are observed over 25 grains/m^3 [33]. Accordingly, this level is selected to define the boundary dates of the main pollen season as the first and the last day that 25 grains/m^3 are observed, corresponding to the start and the end of the peak season, respectively. Thus, every random sample drawn would include X% of observations from the peak season and X% from the off-peaks season, $X \in \{10, 20, 30\}$.

For each percentage of missing data points, a 10-fold cross-validation was run to cover the full dataset. Subsequently, the GP and CNN algorithms were trained and tested against the corresponding test set. Since IDW is an unsupervised method, meaning that there is no need to train the algorithm in order to extract the relations between pollen counts at different locations and the control station of Farmacia, it was used as a benchmark to evaluate the Gaussian process and the neural network. For the Gaussian process regression, a dot product covariance function $k(x_i, x_j) = \sigma_0^2 + x_i \cdot x_j$ was used along with a noise level estimation $\sigma^2 \delta(x_i, x_j)$ [31] where $\delta(x_i, x_j)$ is the Kronecker delta function.

In order to parse the pollen counts through the CNN as the input, at each time t, the seven locations surrounding location i and the pollen observations p_i were transformed into a 3×3 matrix, as seen in Equation (2), which in turn was transformed into a 5×5 matrix (Equation (2)) by adding zeros in order to capture the contributions of the individual locations and enrich the information flow through the network. Thus, by using a 2×2 filter with a stride equal to one, we ensured the parsing of an individual location, as seen in Figure 2 with element X_{11} of the input matrix:

$$[p_1, p_2, p_3, p_4, p_5, p_6, p_7]_t \rightarrow \begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & p_1 & p_2 & p_3 & 0 \\ 0 & p_4 & p_5 & p_6 & 0 \\ 0 & p_7 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}_t \tag{2}$$

Given the aforementioned setup, with 14 filters, it would be sufficient to cover the full feature map without taking into account the last two filters, which would result in 0 (Figure 2). However, the generalization ability of CNNs is not based on limiting the number of parameters [34], and with an incremental experiment, 32 filters are adequate to solve this problem. As a final step, the filters were fully connected to a 5 neuron layer, which at the same time was connected to the output. CNNs usually suffer from an abrupt increment of the number of parameters as their complexity, in terms of topology, increases. Thus, it is common to include a pooling layer in order to reduce the number of parameters. Since the architecture proposed was fairly simple, adding this kind of layer was avoided, as the number of parameters was already small.

Figure 1 shows the particular nature of airborne pollen time series, with a high presence of observations equal or close to 0, especially outside the main pollen season. This may lead to what it is know as dead neurons, which results in the weights being equal to 0, since the amount of information

from the inputs is limited. To avoid this situation, the network was trained using a "LeakyReLu" activation function ($\alpha = 0.1$) along with the Adam optimization algorithm [35] instead of the traditional stochastic gradient descent. A learning rate $\alpha = 0.001$, an exponential decay of $\beta_1 = 0.9$, and $\beta_2 = 0.999$ were used to train the network over 60 epochs.

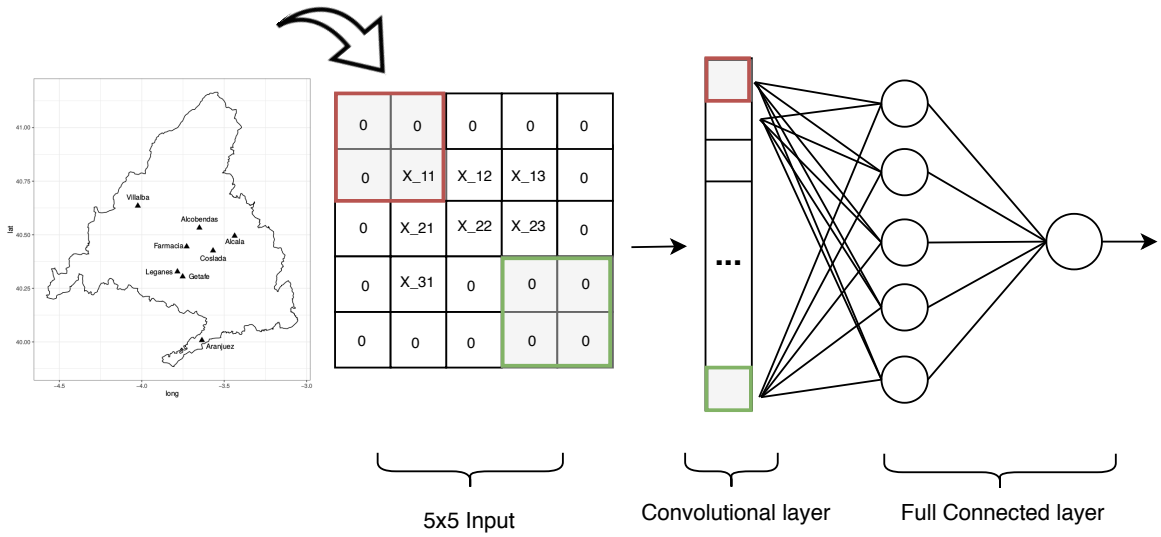


Figure 2. Convolutional neural network.

3. Results

The study was based on removing data points from the pollen observation station located at the Faculty of Pharmacy (Farmacia). Table 1 shows the average 10-fold RMSE and standard deviation for each percentage of observations subtracted from the time series. As mentioned in Section 2.3, the proportions of missing data were equally selected within and outside the main pollen season.

When removing 10% of the data points, the CNN provided a more accurate estimation, both during the peak and off-peak season, with an RMSE equal to 39.89 and 4.53 grains per square meter, respectively. These quantities were compared to an RMSE of 42.44 using Gaussian process regression and 43.97 obtained by IDW. The differences were closer during the off-peak season with 5.07 and 5.36 for GP and IDW, respectively. This situation was expected since, on the one hand, there were more observations outside the main pollen season and, on the other, airborne pollen concentrations were close to zero. Additionally, the stability of the estimations seemed higher for the CNN given the lowest standard deviation of the results.

Table 1. Average and standard deviation (in parenthesis) of the RMSE per percentage of missing data and methodology.

% of Missing	Peak Season			Off-Peak Season			All		
	IDW	GP	CNN	IDW	GP	CNN	IDW	GP	CNN
10%	43.97 (5.84)	42.44 (8.56)	39.89 (5.25)	5.36 (0.91)	5.07 (1.09)	4.53 (0.98)	18.02 (2.05)	17.41 (3.02)	16.46 (2.02)
20%	41.55 (3.95)	39.69 (4.72)	37.35 (4.21)	5.76 (1.50)	5.24 (1.55)	4.80 (1.20)	17.26 (1.57)	16.43 (1.80)	15.42 (1.48)
30%	42.79 (3.77)	41.50 (4.20)	40.14 (4.05)	6.45 (0.68)	5.92 (0.79)	5.40 (1.15)	17.87 (1.52)	17.24 (1.66)	16.60 (1.60)

CNN also improved the accuracy of other methods, both during the peak and off-peak season, when 20% of the observations were removed. With respect to the peak season, the CNN performed

10% better than IDW. This accuracy went to 5% better when compared to GP. In this case, as happened when removing 30% of the observations, the standard deviation was higher than that obtained by IDW. This behavior was expected since in computational intelligence models, the principle of the more data, the better applies in general. Still, the differences were residual.

Figure 3 shows the estimation of all three methods for the peak season during sample years 2008 to 2011. It can be clearly seen that the CNN (red circle) tended to adjust better to sudden peaks in concentrations. Furthermore, it managed to mitigate the influence of extreme observations from other locations, as it did not overestimate as much as the other two methodologies. This situation was demonstrated both in 2009 and 2011, where all models estimated an airborne concentration over 100 grains/m^3, while the true observation was closer to 50 grains/m^3.

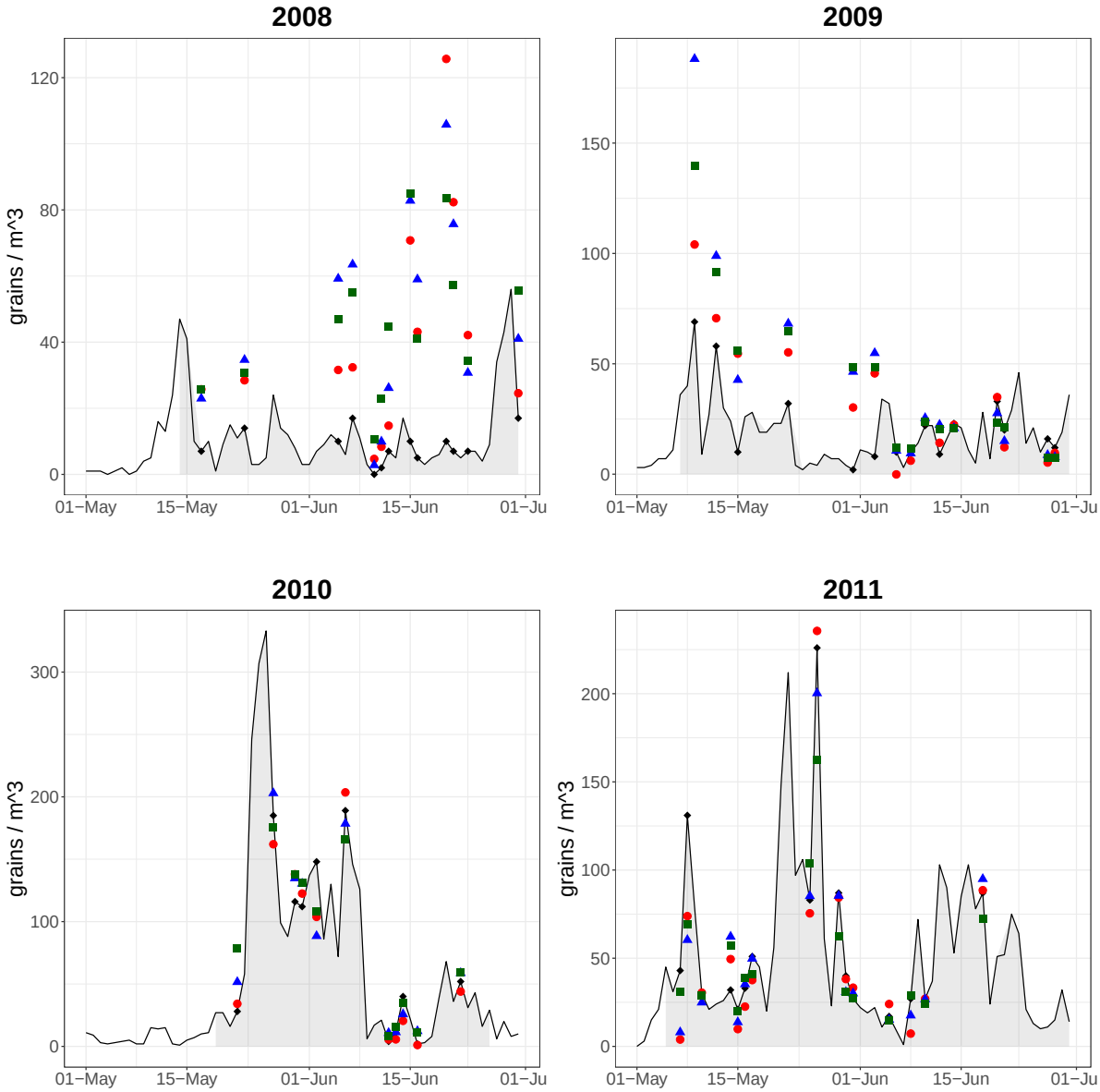


Figure 3. Sample estimation points during peak seasons (grey area) in 2008 to 2011 representing CNN (red circle), GPR(blue triangle), and IDW (green square). The missing observed data points are represented by a black diamond.

One of the well known disadvantages of neural networks is the training execution times. For this reason, the training times were tracked (Table 2) to have a fair comparison in every aspect. For this analysis, IDW was discarded since it is a deterministic method, which was not comparable, in terms of time performance, to the other two. Intuitively, the more data removed, the shorter the execution times, as the training set decreased in size. We can see a certain stability in CNN time performance when compared to the Gaussian process consuming on average about 10% of the time of the GP in the training process.

Table 2. Average and standard deviation (in parenthesis) of the 10-fold execution time in seconds per percentage of missing data.

	GP	CNN
10% missing	243.38 (10.56)	17.47 (0.66)
20% missing	171.22 (18.00)	15.91 (0.64)
30% missing	148.12 (24.19)	15.55 (1.65)

4. Discussion

As we saw in Section 3, the competitive advantage of using CNNs to impute airborne pollen concentrations was clear. In terms of accuracy, both the Gaussian process and the CNN performed better than the inverse distance weighting method (Table 1). At the same time, both models could be extended by including a temporal dimension.

In order to provide insightful results, these were reported based on whether the observations belonged to the main pollen season or not. The main pollen season, or peak season, was defined using a threshold approach [36] of 25 grains/m^3. This threshold differed from the literature based on the study region and pollen genus. However, the work in [37] concluded that Poaceae is ranked highest in terms of allergic significance, and the work in [10] established a threshold of 25 grains/m^3 for Plantago pollen. Furthermore, high pollen concentration thresholds might lead to very short peak seasons and consequently few test points.

During the main peak pollen season, all models suffered from the influence of extreme values in other locations (Figure 3), resulting in an overestimation of the concentrations at the target location. However, this influence was mitigated by the CNN due to the increase in the number of filters, which increased the model's generalization.

During off-peak periods, the differences between the techniques proposed were marginal, but regarding their practical application, these observations were not as important, since they did not imply a high risk for the allergic population. There is no consensus about how much missing data is allowed in order to have unbiased statistical analyses when using inference models [14,15], the cutoff values being around 10% depending on the dataset. This was the reason why 10%, 20%, and 30% of missing data were selected. As a consequence, an increase of the number of filters used in the CNN topology was necessary to provide the generalization of the estimations as proven by the results. However, the larger the amount of missing data, the smaller the number of training observations, which negatively influences the learning process of the CNN. This explains why a decrease in the accuracy was obtained as a result.

Even though only pollen observations were used in this study, mainly to compare the proposed solution with the benchmark IDW, the CNN provided the flexibility to include meteorological measures or predictions as input variables. There is evidence that including such variables [12,24] improves the estimations of airborne pollen concentrations. Moreover, these variables serve as a differential factor to mitigate under- and over-estimation of sudden high peaks during the main pollen season. On the other hand, the simplicity of the topology of the proposed solution was lost. As a consequence, execution

training periods increased as the number of hyper-parameters increased. This is a well known drawback of machine learning models; however, the applied method performed better compared to the others tested, mainly by computation time (Table 2), and is expected to outperform them significantly when additional co-factors are used, such as meteorological variables.

5. Conclusions

In this study, we tackled the problem of the spatial imputation of missing values for pollen time series in Madrid. We proposed the use of convolutional neural networks and conducted a comparison with two traditional geoimputation techniques, inverse distance weighting and Gaussian process regression. The CNN's competitive advantage was shown both in terms of accuracy and execution times.

The results show that it is possible to apply this technique to fields outside computer vision and linguistics. Field experts can take advantages of the potential of CNNs and their application to spatial imputation. Even though the results were promising, they could be improved by including meteorological measures or predictions in the model, yet increasing the computational cost and complexity. This notwithstanding, it was also intended to increase the awareness of the advantages and disadvantages of such a technique.

Author Contributions: R.N.: conceptualization, methodology, software, formal analysis, investigation, writing-m –original draft, and visualization. J.L.A.: writing, review editing, and supervision.

Funding: This research received no external funding.

Acknowledgments: Pollen data were kindly provided by Patricia Cervigón (Palinocam network, Comunidad de Madrid) and Montserrat Gutiérrez Bustillo (Department of Botany, Complutense University of Madrid).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. De Weger, L.A.; Bergmann, K.C.; Rantio-Lehtimaki, A.; Dahl, A.; Buters, J.; Déchamp, C.; Belmonte, J.; Thibaudon, M.; Cecchi, L.; Besancenot, J.P.; et al. Impact of Pollen. In *Allergenic Pollen*; Sofiev, M., Bergmann, K.C., Eds.; Springer: Dordrecht, The Netherlands, 2013; pp. 161–215._6. [CrossRef]
2. Lake, I.; Jones, N.; Agnew, M.; Goodess, C.; Giorgi, F.; Lynda, H.L.; Semenov, M.; Solmon, F.; Storkey, J.; Vautard, R.; et al. Erratum: "Climate Change and Future Pollen Allergy in Europe". *Environ. Health Perspect.* **2018**, *126*. [CrossRef] [PubMed]
3. Sabariego, S.; Cuesta, P.; Fernández-González, F.; Pérez-Badia, R. Models for forecasting airborne Cupressaceae pollen levels in central Spain. *Int. J. Biometeorol.* **2012**, *56*, 253–258. [CrossRef] [PubMed]
4. Smith, M.; Emberlin, J. A 30-day-ahead forecast model for grass pollen in north London, UK. *Int. J. Biometeorol.* **2006**, *50*, 233–242. [CrossRef] [PubMed]
5. Silva-Palacios, I.; Fernández-Rodríguez, S.; Durán-Barroso, P.; Tormo-Molina, R.; Maya-Manzano, J.; Gonzalo-Garijo, A. Temporal modelling and forecasting of the airborne pollen of Cupressaceae on the southwestern Iberian peninsula. *Int. J. Biometeorol.* **2016**, *60*, 1509–1517. [CrossRef]
6. Schaber, J.; Badeck, F.W. Physiology-based phenology models for forest tree species in Germany. *Int. J. Biometeorol.* **2003**, *47*, 193–201. [CrossRef]
7. Navares, R.; Aznarte, J. *Forecasting the Start and End of Pollen Season in Madrid*; Springer International Publishing: Berlin/Heidelberg, Germany, 2017; Chapter 26, pp. 387–399.
8. Puc, M. Artificial neural network model of the relationship between Betula pollen and meteorological factors in Szczecin (Poland). *Int. J. Biometeorol.* **2011**, *56*, 395–401. [CrossRef]
9. Castellano-Méndez, M.; Aira, M.J.; Iglesias, I.; Jato, V.; González-Manteiga, W. Artificial neural networks as a useful tool to predict the risk level of Betula pollen in the air. *Int. J. Biometeorol.* **2005**, *49*, 310–316. [CrossRef]
10. Iglesias-Otero, M.A.; Fernández-González, M.; Rodríguez-Caride, D.; Astray, G.; Mejuto, J.C.; Rodríguez-Rajo, F.J. A model to forecast the risk periods of Plantago pollen allergy by using ANN methodology. *Aerobiologia* **2015**, *31*, 201–211. [CrossRef]

11. Navares, R.; Aznarte, J. Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features. *Int. J. Biometeorol.* **2016**. [CrossRef]
12. Navares, R.; Aznarte, J. What are the most important variables for Poaceae airborne pollen forecasting? *Sci. Total Environ.* **2016**, *579*, 1161–1169. [CrossRef]
13. Oteros, J.; Sofiev, M.; Smith, M.; Clot, B.; Damialis, A.; Prank, M.; Werchan, M.; Wachter, R.; Weber, A.; Kutzora, S.; et al. Building an automatic pollen monitoring network (ePIN): Selection of optimal sites by clustering pollen stations. *Sci. Total Environ.* **2019**, *688*, 1263–1274. [CrossRef]
14. Schafer, J.L. Multiple imputation: A primer. *Stat. Methods Med. Res.* **1999**, *8*, 3–15. [CrossRef] [PubMed]
15. Bennett, D. How can I deal with missing data in my study? *Aust. N. Z. J. Public Health* **2001**, *25*, 464–469. [CrossRef] [PubMed]
16. Shepard, D. A Two-dimensional Interpolation Function for Irregularly-spaced Data. In Proceedings of the 23rd ACM National Conference, Las Vegas, NV, USA, 27–29 August 1968; ACM: New York, NY, USA, 1968; pp. 517–524. [CrossRef]
17. Matheron, G. Principles of geostatistics. *Econ. Geol.* **1963**, *58*, 1246–1266. [CrossRef]
18. Kordon, A.K. Competitive Advantages of Computational Intelligence. In *Applying Computational Intelligence: How to Create Value*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 233–256._9. [CrossRef]
19. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [CrossRef]
20. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*; Pereira, F., Burges, C.J.C., Bottou, L., Weinberger, K.Q., Eds.; Curran Associates, Inc.: Makati, Philippines, 2012; pp. 1097–1105.
21. Gehring, J.; Auli, M.; Grangier, D.; Yarats, D.; Dauphin, Y.N. Convolutional Sequence to Sequence Learning. In *Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia, 6–11 August 2017*; Precup, D., Teh, Y.W., Eds.; PMLR: Sydney, Australia, 2017; Volume 70, pp. 1243–1252.
22. Smith, S.W. *The Scientist and Engineer's Guide to Digital Signal Processing*; California Technical Publishing: San Diego, CA, USA, 1997.
23. Nowosad, J. Spatiotemporal models for predicting high pollen concentration level of Corylus, Alnus, and Betula. *Int. J. Biometeorol.* **2016**, *60*, 843–855. [CrossRef]
24. Navares, R.; Aznarte, J.L. Forecasting Plantago pollen: improving feature selection through random forests, clustering, and Friedman tests. *Theor. Appl. Climatol.* **2019**. [CrossRef]
25. Zewdie, G.K.; Lary, D.J.; Levetin, E.; Garuma, G.F. Applying Deep Neural Networks and Ensemble Machine Learning Methods to Forecast Airborne Ambrosia Pollen. *Int. J. Environ. Res. Public Health* **2019**, *16*, 1992. [CrossRef]
26. Sevillano, V.; Aznarte, J.L. Improving classification of pollen grain images of the POLEN23E dataset through three different applications of deep learning convolutional neural networks. *PLoS ONE* **2018**, *13*, e0201807. [CrossRef]
27. Khanzhina, N.; Putin, E.; Filchenkov, A.; Zamyatina, E. Pollen grain recognition using convolutional neural network. In Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges, Belgium, 25–27 April 2018.
28. Galán Soldevilla, C.; Cariñanos González, P.; Alcázar Teno, P.; Domínguez Vílches, E. *Manual de Calidad y Gestión de la Red Española de Aerobiología*; Universidad de Córdoba: Córdoba, Spain, 2007.
29. Tobler, W.R. A Computer Movie Simulating Urban Growth in the Detroit Region. *Econ. Geogr.* **1970**, *46*, 234–240. [CrossRef]
30. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*, 1st ed.; MIT Press: Cambridge, MA, USA, 2012.
31. Edward Rasmussen, C.; Bousquet, O.; von Luxburg, U.; Rätsch, G. Gaussian Processes in Machine Learning. In *Advanced Lectures on Machine Learning: ML Summer*; Springer: Berlin/Heidelberg, Germany, 2004; Volume 3176._4. [CrossRef]
32. Gamboa, J.C.B. Deep Learning for Time-Series Analysis. *arXiv* **2017**, arXiv:1701.01887.
33. Rodríguez-Rajo, F.; Frenguelli, G.; Jato, M. Effect of air temperature on forecasting the start of the Betula pollen season at two contrasting sites in the south of Europe (1995–2001). *Int. J. Biometeorol.* **1983**, *47*, 117–125.
34. Zhang, C.; Bengio, S.; Hardt, M.; Recht, B.; Vinyals, O. Understanding deep learning requires rethinking generalization. *arXiv* **2016**, arXiv:1611.03530.
35. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.

36. Jato, V.; Rodríguez-Rajo, F.J.; Alcázar, P.; Nuntiis, P.D.; Galán, C.; Mandrioli, P. May the definition of pollen season influence aerobiological results? *Aerobiologia* **2006**, *22*, 13–25. [CrossRef]
37. Peternel, R.; Srnec, L.; Culig, J.; Hrga, I.; Hercog, P. Poaceae pollen in the atmosphere of Zagreb (Croatia), 2002–2005. *Grana* **2005**, *45*, 130–136. [CrossRef]

Chapter 13

Conclusions

This doctoral thesis is focused on the research of computational intelligence techniques and its applications in the context of air quality forecasting. The exposure to poor air quality is an outstandingly important topic for governmental and independent agencies due to its health and economic implications. Air quality forecasting is a pillar for taking effective control measures both at a large scale, for instance setting global air quality guidelines, and regional and private scale such as city traffic control measures or emergency admissions operational planning for clinical institutions.

Air quality, together with meteorology, is bound to produce large volumes of data. Moreover, given the distribution of the observation stations, the dimensionality exponentially increases both temporal and spatially. All this data is interrelated and shows influence among each other. Due to this reason along with the different nature of the sources of the pollutants, current approaches to the problem require specific expert knowledge. The main purpose of this work, as defined in Section 1.3, was to come out with a methodology able to automatically filter and extract relevant information, from an engineering point of view, which allows accurate air quality predictions independent from the number and nature of its inputs. In order to achieve this objective, the research plan was split in incremental phases, outlined in Section 1.4, towards the final system which correspond to each specific objective. In light of the results presented in Chapters 3-10, we can draw the following principal conclusions:

1. By surveying several machine learning we have demonstrated that tree-based models were appropriate to both forecast concentrations during the main pollination season, which is the period with higher risk for allergic patients, and extract the influential variables. These models managed to outperform other proposals found in the literature aimed at easing the prevention of exposures to risk concentration levels.

2. Through automatic feature selection it was shown the strong links between phenological and meteorological studies performed and the automatic way taken with the pure data point of view. Results were inline with the aforementioned studies and the methods enable a new family of models saving time-consuming research of individual problems which, at the same time, requires resources in research centers and medical institutions that are not always available.

3. Even though previous results supported the approach, the process required to acquire specific knowledge for each pollutant and, the generation of input features that are selected in an automatic manner, plays against one of the major hypotheses of this thesis. We have shown how long short-term memory units are able

to properly extract the information required to predict chemical pollutants (influenced by current environmental conditions), and pollen concentrations (driven also by meteorology conditions during plant formation) in a unified approach.

4. We also demonstrated that the topology of the network is determinant in order to isolate the potential relations among the same type of pollutant as a first step and to avoid noisy interactions. Results showed that network topologies play an essential role in model performance. Research outcome proved that grouping pollutant types first in order to obtain temporal relations and then combine the results to represent spatial relations, eases the network to focus on the relevant information and provides a more stable results across locations. It was demonstrated that including in the configuration fully connected layers, either as an output or hidden layer, the networks are able to better identify the relations among pollutants with no data preprocessing. However, we have seen that there is still room for improvement as the LSTMs struggle to identify the presence of sudden high peaks as past information weights on the predictions. This situation can be mitigated by capping pollutant observation levels to thresholds over which it implies risk for human health.

5. All previous findings were wrapped in the final contribution with a direct application of the forecasts provided by the deep learning architecture. Additionally, it was found the benefits of stacking methodologies as it eases forecasting generalization. Among the stacking techniques, the application of convolutional neural networks for this task was extended from their traditional computer vision and speech recognition fields.

One of the main challenges during the thesis was to present the proposals to journals (Appendix A) and reviewers belonging to non-engineering fields. Communicating our ideas required a special effort to adapt the concepts to domain-specific language even though these proposals were backed up by results. The experience over these years has led not only to acknowledge and welcome the workload required to contribute to science, but also to understand the full research process and the importance of contributions to scientific communication. The outcome of this research has opened new perspectives to the scientific community based on the number of reads and citations the different contributions have generated so far and the collaboration proposals received. This has encouraged to continue researching on the new perspectives this work opened to improve the understanding about the quality of the air and its impacts on society.

Appendix A

Publications list

Chapter 3 R. Navares and JL. Aznarte. "Forecasting the Start and End of Pollen Season in Madrid". In: *Advances in Time Series Analysis and Forecasting*. Ed. by Ignacio Rojas, Héctor Pomares, and Olga Valenzuela. Cham: Springer International Publishing, 2017, pp. 387–399

Chapter 4 R. Navares and J.L. Aznarte. "Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features". In: *Int. J. Biometeorol* (2016). DOI: 10.1007/s00484-016-1242-8

Chapter 5 R. Navares and J.L. Aznarte. "What are the most important variables for Poaceae airborne pollen forecasting?" In: *Science of the Total Environment* 579 (2016), pp. 1161–1169

Chapter 6 R. Navares and JL. Aznarte. "Forecasting Plantago pollen: improving feature selection through random forests, clustering, and Friedman tests". In: *Theoretical and Applied Climatology* (2019). ISSN: 1434-4483. DOI: 10.1007/s00704-019-02954-1. URL: https://doi.org/10.1007/s00704-019-02954-1

Chapter 7 D. Valput, R. Navares, and JL. Aznarte. "Forecasting hourly NO2 concentrations by ensembling neural networks and mesoscale models". In: *Neural Computing and Applications* (2019). ISSN: 1433-3058. DOI: 10.1007/s00521-019-04442-z. URL: https://doi.org/10.1007/s00521-019-04442-z

Chapter 8 R. Navares and JL. Aznarte. "Predicting air quality with deep learning LSTM: Towards comprehensive models". In: *Ecological Informatics* 55 (2020), p. 101019. ISSN: 1574-9541. DOI: https://doi.org/10.1016/j.ecoinf.2019.101019. URL: http://www.sciencedirect.com/science/article/pii/S1574954119303309

Chapter 9 R. Navares et al. "Comparing ARIMA and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in Madrid." In: *Stoch Environ Res Risk Assess* (2018), pp. 1–11. DOI: 10.1007/s00477-018-1519-z

Chapter 10 R. Navares and JL Aznarte. "Deep learning architecture to predict daily hospital admissions". In: *Neural Computing and Applications* (2020). DOI: https://doi.org/10.1007/s00521-020-04840-8

Chapter 11 R. Navares et al. "Direct assessment of health impacts on hospital admission from traffic intensity in Madrid". In: *Environmental Research* 184 (2020), p. 109254. ISSN: 0013-9351. DOI: https://doi.org/10.1016/j.envres.2020.109254. URL: http://www.sciencedirect.com/science/article/pii/S0013935120301468

Chapter 12 R. Navares and JL. Aznarte. "Geographical Imputation of Missing Poaceae Pollen Data via Convolutional Neural Networks". In: *Atmosphere* 10 (2019), pp. 1–10. DOI: https://doi.org/10.3390/atmos10110717. URL: https://www.mdpi.com/2073-4433/10/11/717

Bibliography

[1] G. Abraham, G.B. Byrnes, and C.A. Bain. "Short-term forecasting of emergency inpatient flow." In: *Inf Technol Biomed* 13 (2009), pp. 380–388.

[2] F. Aguilera et al. "Phenological models to predict the main flowering phases of olive (Olea europaea L.) along a latitudinal and longitudinal gradient across the Mediterranean region." In: *Int. J. Bioeteorology* 59 (2014), pp. 629–641.

[3] J. C. Alberdi et al. "Daily mortality in Madrid Community (Spain) 1986-1991: Relationship with atmospheric variables." In: *European Journal of Epidemiology.* 14 (1998), pp. 571–578.

[4] T. B. Andersen. "A model to predict the beginning of the pollen season". In: *Grana* 30 (1991), pp. 269–275.

[5] I. Antépara et al. "Pollen allergy in the Bilbao area (European Atlantic seaboard climate): pollination forecasting methods". In: *Clinical & Experimental Allergy* 25.2 (1995), pp. 133–140.

[6] M.Y. Anwar et al. "Time series analysis of malaria in Afghanistan: using ARIMA models to predict future trends in incidence." In: *Malar J* 15 (2016), p. 566.

[7] G. Astray et al. "Airborne castanea pollen forecasting model for ecological and allergological implementation." In: *Science of the Total Environment* 548–549 (2016), pp. 110–121.

[8] J. L. Aznarte et al. "Forecasting airborne pollen concentration time series with neural and neuro-fuzzy models". In: *Expert Systems with Applications* 32.4 (2007), pp. 1218–1225.

[9] C. Bergmeir, R.J. Hyndman, and B. Koo. "A note on the validity of cross-validation for evaluating autoregressive time series prediction." In: *Computational Statistics and Data Analysis* 120 (2018), pp. 70–83. DOI: 10.1016/j.csda.2017.11.003.

[10] C. M. Bishop. "Neural networks for pattern recognition". en. In: Oxford University Press, 1995. ISBN: 0198538642.

[11] A. Blum and P. Langley. "Selection of relevant features and examples in machine learning." In: *Artificial Intelligence* 97 (1997), pp. 245–271.

[12] G. E. P. Box and G. M. Jenkins. "Time series analysis: forecasting and control." In: *Holden-Day* (1976).

[13] L. Breiman. "Bagging predictiors." In: *Machine Learning.* 25 (1996), pp. 123–140.

[14] L. Breiman. "Manual on seeting up, using and understanding random forest". In: *Stat. Dept. University of California Berkley.* v3.1 (2002).

[15] L. Breiman. "Random Forest". In: *Machine Learning* 45 (2001), pp. 5–32.

[16] M. A. Brighetti et al. "Multivariate statistical forecasting modeling to predict Poaceae pollen critical concentrations by meteoclimatic data." In: *Aerobiologia* 30 (2013), pp. 25–33.

[17] M.G.R. Cannell and R.I. Smith. "Thermal time, chill days and prediction of budburst in Picea sitchensis." In: *Journal of Applied Ecology* 20 (1983), pp. 269–275.

[18] R. Carmona et al. "Geographical variation in relative risks associated with cold waves in Spain: The need for a cold wave prevention plan." In: *Environment International. 2016* 88 (2016), pp. 103–111.

[19] M. Castellano-Méndez et al. "Artificial neural networks as a useful tool to predict the risk level of Betula pollen in the air." In: *Int. J. Biometeorology* 49 (2005), pp. 310–316.

[20] M. Catalano et al. "Improving the prediction of air pollution peak episodes generated by urban transport networks." In: *Environmental Science & Policy* 60 (2016), pp. 69–83.

[21] Giulia Cesaroni et al. "Health benefits of traffic-related air pollution reduction in different socioeconomic groups: the effect of low-emission zoning in Rome". In: *Occupational and Environmental Medicine* 69.2 (2012), pp. 133–139. ISSN: 1351-0711. DOI: 10.1136/oem.2010.063750.

[22] S. le Cessie and J.C. van Howelingen. "Ridge estimators in logistic regression." In: *Applied Statistics* 41 (1992), pp. 191–201.

[23] A. Chaloulakou, M. Saisana, and N. Spyrellis. "Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens." In: *Sci Total Environ* 313 (1998), pp. 1–13.

[24] A.B. Chelani et al. "Prediction of sulphur dioxide concentration using artificial neural networks." In: *Environmental Modelling & Software* 17 (2002), pp. 161–168.

[25] François Chollet. "Xception: Deep Learning with Depthwise Separable Convolutions". In: *CoRR* abs/1610.02357 (2016). arXiv: 1610.02357. URL: http://arxiv.org/abs/1610.02357.

[26] W. J. Conover. "Nonparametric methods". en. In: *Practical nonparametric statistics*. Ed. by Brad Wiley and Mary O'Sullivan. DOI: 10.1002/bimj.19730150311. John Wiley and Sons, 1999, pp. 233–305. ISBN: 978-0-471-16068-7.

[27] C. Cortes and V.N. Vapnik. "Support-vector networks." In: *Machine Learning* 20 (1995), pp. 273–276.

[28] T.R. Cotos-Yáñez, F.J. Rodríguez-Rajo, and M.V. Jato. "Short-term prediction of Betula airborne pollen concentration in Vigo (NW Spain) using logistic additive models and partially linear models". In: *Int J Biometeorol* 48 (2004), pp. 179–185.

[29] Z. Csépe et al. "Predicting daily ragweed pollen concentrations using Computational Intelligence techniques over two heavily polluted areas in Europe." In: *Science of the Total Environment* 476–477 (2014), pp. 542–552.

[30] G. D'Amato et al. "Climate Change, Migration, and Allergic Respiratory Diseases: An Update for the Allergist." In: *World Allergy Organ J* 4 (2011), pp. 120–125.

[31] A.J. Deák et al. "Climate sensitivity of allergenic taxa in Central Europe associated with new climate change related forces". In: *Science of the Total Environment* 442 (2013), pp. 36–47.

[32] J. Díaz, C. Linares, and A. Tobías. "Short term effects of pollen species on hospital admissions in the city of Madrid in terms of specific causes and age." In: *Aerobiologia* 23 (2007), pp. 231–238.

[33] J. Díaz et al. "A model for forecasting emergency hospital admissions: effect of environmental variables." In: *Journal of Environmental Health* 64 (2001), pp. 9–15.

[34] J. Díaz et al. "Geographical variation in relative risks associated with heat: update of Spain's Heat Wave Prevention Plan." In: *Environment International. 2015* 85 (2015), pp. 273–283.

[35] J. Díaz et al. "Heat waves in Madrid, 1986-1997: effects on the health of the elderly." In: *International Archives Occupational and Environmental Health* 75 (2002), pp. 163–170.

[36] J. Díaz et al. "Modeling of Air Pollution and its Relationship with mortality and morbidity in Madrid (Spain)." In: *Int Arch Occup Environ Health* 75 (1999), pp. 366–376.

[37] J. Díaz et al. "Mortality impact of extreme winter temperatures." In: *International Journal of Biometeorology* 49 (2005), pp. 179–183.

[38] M. Dominak, L. Swiecicki, and J. Rybakowski. "Psychiatric hospitalizations for affective disorders in Warsaw, Poland: Effect of season and intensity of sunlight." In: *Phychiatry Res* 229 (2015), pp. 289–294.

[39] A. Earnest et al. "Using autoregressive integrated moving average (ARIMA) models to predict and monitor the number of beds occupied during a SARS outbreak in a tertiary hospital in Singapore." In: *BMC Health Serv Res.* 5 (2005), p. 36.

[40] EEA. "Health impacts of air pollution". In: (2020). URL: https://www.eea.europa.eu/themes/air/health-impacts-of-air-pollution.

[41] Report WHO Regional Office for Europe. *Review of evidence on health aspects of air pollution – REVIHAAP Project: Technical Report*. Tech. rep. 2013.

[42] Publications Office of the European Union. *Spatial representativeness of air quality monitoring sites: Outcomes of the FAIRMODE-AQUILA intercomparison exercise*. Tech. rep. 2017. URL: https://ec.europa.eu/jrc/en/publication/spatial-representativeness-air-quality-monitoring-sites-outcomes-fairmodeaquila-intercomparison.

[43] M. Fawcett. "ROC Graphs: Notes and Practical Considerations for Data Mining Researchers." In: *Tech. rep., HP Laboratories* (2003).

[44] Z. Feher and M. Jarai-Komlodi. "An examination of the main characteristics of the pollen seasons in Budapest, Hungary (1991-1996)". In: *Grana* 36 (1997), pp. 169–174.

[45] Anna Font et al. "A tale of two cities: is air pollution improving in Paris and London?" In: *Environmental Pollution* 249 (2019), pp. 1 –12. ISSN: 0269-7491. DOI: https://doi.org/10.1016/j.envpol.2019.01.040. URL: http://www.sciencedirect.com/science/article/pii/S0269749118321687.

[46] J. H. Friedman. "Greedy Function Approximation: A Gradient Boosting Machine". In: *The Annals of Statistics* 29 (2001), pp. 1189–1232.

[47] M. Friedman. "The use of ranks to avoid the assumption of normality implicit in the analysis of variance." In: *J. of American Statistical Association* 32 (1937), pp. 674–701.

[48] C. Galán et al. "A comparative analysis of daily variations in the Gramineae pollen counts at Cordoba, Spain and London, UK". In: *Grana* 34 (1995), pp. 189–198.

[49] John Cristian Borges Gamboa. "Deep Learning for Time-Series Analysis". In: *CoRR* abs/1701.01887 (2017). arXiv: 1701.01887. URL: http://arxiv.org/abs/1701.01887.

[50] S. García et al. "Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power." In: *Information Science* 180 (2010), pp. 2044–2064.

[51] M.W. Gardner and S.R. Dorling. "Artificial neural networks (the multilayer perceptron) a review of applications in the atmospheric sciences." In: *Atmos Environ* 32 (1998), pp. 2627–2636.

[52] J. Gehring et al. "Convolutional Sequence to Sequence Learning". In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. International Convention Centre, Sydney, Australia: PMLR, 2017, pp. 1243–1252. URL: http://proceedings.mlr.press/v70/gehring17a.html.

[53] F.A. Gers, J. Schmidhuber, and F. Cummins. "Learning to forget: Continual prediction with LSTM." In: *Neural Computation* 12 (2000), pp. 2451–2471.

[54] S. González et al. "Relationship between atmospheric pressure and mortality in the Madrid Autonumus Region : A time series study." In: *Int J Biometeorol* 45 (2001), pp. 34–40.

[55] B. James Green et al. "Atmospheric Poaceae pollen frequencies and associations with meteorological parameters in Brisbane, Australia: a 5 year record, 1994–1999." In: *Int. J. Biometeorology* 40 (2004), pp. 172–178.

[56] G. Grivas and A. Chaloulakou. "Artificial neural network models for prediction of PM10 hourly concentrations, in the greater area of Athens, Greece". In: *Atmospheric Environment* 40 (2006), pp. 1216–1229.

[57] M. A. Hall. "Correlation-based feature selection for machine learining." In: *PhD. Thesis. University of Waikato* (1999).

[58] S. Haykin. "Neural networks and learning machines". en. In: ed. by Marcia J. Horton and Alice Dworkin. Pearson Prentice Hall, 1999. ISBN: 978-0-13-147139-9.

[59] Jianjun He et al. "Numerical Model-Based Artificial Neural Network Model and Its Application for Quantifying Impact Factors of Urban Air Quality". In: *Water, Air, & Soil Pollution* 227 (2016), pp. 227 –235. DOI: https://doi.org/10.1007/s11270-016-2930-z.

[60] S. Hochreiter and J. Schmidhuber. "Long short-term memory." In: *Neural Computation* 9 (1997), pp. 1735–1780.

[61] S. Hochreiter et al. "Gradient flow in recurrent nets: the difficulty of learning long-term dependencies". In: *A Field Guide to Dynamical Recurrent Neural Networks*. Ed. by S. C. Kremer and J. F. Kolen. IEEE Press, 2001.

[62] S. Holm. "A simple sequentially rejective multiple test procedure." In: *Scandinavian Journal of Statistics* 6 (1979), pp. 65–70.

[63] Oystein Hov et al. "Comparison of numerical techniques for use in air pollution models with non-linear chemical reactions". In: *Atmospheric Environment (1967)* 23.5 (1989), pp. 967 –983. ISSN: 0004-6981. DOI: https://doi.org/10.1016/0004-6981(89)90301-6. URL: http://www.sciencedirect.com/science/article/pii/0004698189903016.

[64] X. Hu, D. Xu, and Q. Wan. "Short-Term Trend Forecast of Different Traffic Pollutants in Minnesota Based on Spot Velocity Conversion." In: *Int J Environ Res Public Health* 9 (2018). DOI: 10.3390/ijerph15091925.

[65] V. Jato et al. "May the definition of pollen season influence aerobiological results?" In: *Aerobiologia* 22 (2006), pp. 13–25.

[66] E. Jiménez et al. "Role of Saharan dust in the relationship between particulate matter and short-term daily mortality among the elderly in Madrid (Spain)." In: *Science of Total Environment* 408 (2010), pp. 5729–5736.

[67] A.M. Jones and R.M. Harrison. "The effects of meteorological factors on atmospheric bioaerosol concentrations—a review". In: *Science of the Total Environment* 326 (2004), pp. 151–181.

[68] F.J. Kelly and J.C. Fussell. "Air pollution and public health: emerging hazards and improved understanding of risk." In: *Environ Geochem Health* 37 (2015), pp. 631–649. DOI: 10.1007/s10653-015-9720-1.

[69] DP. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". In: *CoRR* abs/1412.6980 (2014). URL: http://arxiv.org/abs/1412.6980.

[70] M. Kmenta et al. "The grass pollen season 2014 in Vienna: A pilot study combining phenology, aerobiology and symptom data." In: *Science of the Total Environment* 566-567 (2016), pp. 1614–1620.

[71] R. Kohavi and G.H. John. "Wrappers for feature subset selection." In: *Artificial Intelligence* 97 (1997), pp. 273–324.

[72] A. Krizhevsky, I. Sutskever, and GE. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira et al. Curran Associates, Inc., 2012, pp. 1097–1105. URL: http://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[73] A. Kumar and P. Goyal. "Forecasting of daily air quality index in Delhi." In: *Sci Total Environ* 409 (2011), pp. 5517–23.

[74] V. Kumar et al. "Forecasting malaria cases using climatic factors in Delhi, India: a time series analysis." In: *Malar Res Treat* 2014 (2014), p. 482851.

[75] Y. Lecun et al. "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. ISSN: 0018-9219. DOI: 10.1109/5.726791.

[76] E. Levetin. "Daily Ragweed Pollen Forecasting." In: *Journal of Allergy and clinical Immunology* 133 (2014), AB17.

[77] C. Linares and J. Díaz. "Impact of high temperatures on hospital admissions: comparative analysis with previous studies about mortality (Madrid)." In: *European Journal of Public Health* 18 (2008), pp. 318–322.

[78] C. Linares et al. "Time trend in natural-cause, circulatory-cause and respiratory-cause mortality associated with cold waves in Spain, 1975-2008." In: *Stochastic Research and Risk Assessment* 30 (2016), pp. 1565–1574.

[79] M.A. Luque et al. "Influence of temperature and rainfall on the evolution of cholera epidemics in Lusaka, Zambia 2003-2006: Analysis of a time series." In: *Transactions of the Royal Society of Tropical Medicine and Hygiene* 103 (2009), pp. 137–143.

[80] S. Makridakis, S.C. Wheelwright, and V.E. Mcgee. "Forecasting methods and applications." In: *Wiley, San Francisco* (1983).

[81] T. Masuko. "Computational cost reduction of long short-term memory based on simultaneous compression of input and hidden state". In: *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. 2017, pp. 126–133. DOI: 10.1109/ASRU.2017.8268926.

[82] I. Matyasovszky et al. "Plants remember past weather: a study for atmospheric pollen concentrations of Ambrosia, Poaceae and Populus". In: *Theoretical and Applied Climatology* 122 (2015), pp. 181–193.

[83] S. McWilliams, A. Kinsella, and E. O'Callaghan. "Daily weather variables and affective disorder admissions to psychiatric hospitals." In: *Int J Biometeorol* 58 (2014), pp. 2045–57.

[84] J.C. Montero et al. "Relationship between mortality and heat waves in Castile-La Mancha (1975-2003): influence of local factors." In: *Science of Total Environment* 414 (2012), pp. 73–78.

[85] D. Myszkowska. "Predicting tree pollen season start dates using thermal conditions". In: *Aerobiologia* 30 (2014), pp. 307–321.

[86] R. Navares and JL Aznarte. "Deep learning architecture to predict daily hospital admissions". In: *Neural Computing and Applications* (2020). DOI: https://doi.org/10.1007/s00521-020-04840-8.

[87] R. Navares and JL. Aznarte. "Forecasting Plantago pollen: improving feature selection through random forests, clustering, and Friedman tests". In: *Theoretical and Applied Climatology* (2019). ISSN: 1434-4483. DOI: 10.1007/s00704-019-02954-1. URL: https://doi.org/10.1007/s00704-019-02954-1.

[88] R. Navares and JL. Aznarte. "Forecasting the Start and End of Pollen Season in Madrid". In: *Advances in Time Series Analysis and Forecasting*. Ed. by Ignacio Rojas, Héctor Pomares, and Olga Valenzuela. Cham: Springer International Publishing, 2017, pp. 387–399.

[89] R. Navares and JL. Aznarte. "Geographical Imputation of Missing Poaceae Pollen Data via Convolutional Neural Networks". In: *Atmosphere* 10 (2019), pp. 1–10. DOI: https://doi.org/10.3390/atmos10110717. URL: https://www.mdpi.com/2073-4433/10/11/717.

[90] R. Navares and JL. Aznarte. "Predicting air quality with deep learning LSTM: Towards comprehensive models". In: *Ecological Informatics* 55 (2020), p. 101019. ISSN: 1574-9541. DOI: https://doi.org/10.1016/j.ecoinf.2019.101019. URL: http://www.sciencedirect.com/science/article/pii/S1574954119303309.

[91] R. Navares and J.L. Aznarte. "Predicting the Poaceae pollen season: six month-ahead forecasting and identification of relevant features". In: *Int. J. Biometeorol* (2016). DOI: 10.1007/s00484-016-1242-8.

[92] R. Navares and J.L. Aznarte. "What are the most important variables for Poaceae airborne pollen forecasting?" In: *Science of the Total Environment* 579 (2016), pp. 1161–1169.

[93] R. Navares et al. "Comparing ARIMA and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in Madrid." In: *Stoch Environ Res Risk Assess* (2018), pp. 1–11. DOI: 10.1007/s00477-018-1519-z.

[94] R. Navares et al. "Direct assessment of health impacts on hospital admission from traffic intensity in Madrid". In: *Environmental Research* 184 (2020), p. 109254. ISSN: 0013-9351. DOI: https://doi.org/10.1016/j.envres.2020.109254. URL: http://www.sciencedirect.com/science/article/pii/S0013935120301468.

[95] S. Nilsson and S. Persson. "Tree pollen spectra in the Stockholm region (Sweden), 1973-–1980". In: *Grana* 20 (1981), pp. 179–182.

[96] J. Nowosad. "Spatiotemporal models for predicting high pollen concentration level of Corylus, Alnus and Betula." In: *Int. J. Biometeorology* 60 (2016), pp. 843–855.

[97] E.O. Nsoesie et al. "Modeling to predict cases of hantavirus pulmonary syndrome in Chile." In: *PLoS Negl Trop Dis* 8 (2014), e2779.

[98] Z. Obermeyer and E.J. Emanuel. "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine." In: *N Engl J Med* 375 (2016), pp. 1216–1219. DOI: 10.1056/NEJMp1606181..

[99] Andrea Obersteiner et al. "Pollen-Associated Microbiome Correlates with Pollution Parameters and the Allergenicity of Pollen". In: *PLOS ONE* 11.2 (Feb. 2016), pp. 1–16. DOI: 10.1371/journal.pone.0149545. URL: https://doi.org/10.1371/journal.pone.0149545.

[100] OECD. *The Economic Consequences of Outdoor Air Pollution*. Tech. rep. 2016. URL: https://www.oecd-ilibrary.org/environment/the-economic-consequences-of-outdoor-air-pollution_9789264257474-en.

[101] J.R. Olsen et al. "Effects of new urban motorway infrastructure on road traffic accidents in the local area: a retrospective longitudinal study in Scotland." In: *J Epidemiol Community Health* 70 (2016), pp. 1088–1095.

[102] World Health Organization. *Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide*. Tech. rep. 2005. URL: https://www.who.int/airpollution/publications/aqg2005/en/.

[103] M. A. Otero et al. "A model to forecast the risk periods of Plantago pollen allergy by using ANN methodology." In: *Aerobiologia* 31 (2015), pp. 201–211.

[104] J. Oteros et al. "Biometeorological and autoregressive indices for predicting olive pollen intensity." In: *Int. J. Biometeorology* 57 (2013), pp. 307–316.

[105] H. Ozkaynak et al. "Summary and findings of the EPA and CDC symposium on air pollution exposure and health." In: *J Expo Sci Environ Epidemiol* 19 (2009), pp. 19–29.

[106] I. Silva Palacios, R. Tormo Molina, and A. F. Muñoz Rodríguez. "Influence of wind direction on pollen concentration in the atmosphere." In: *Int. J. Biometeorology* 44 (2000), pp. 128–133.

[107] A. Pauling, R. Gehrig, and B. Clot. "Toward optimized temperature sum parametrizations for forecasting the start of the pollen season". In: *Aerobiologia* 30 (2014), pp. 45–57.

[108] R. Peternel et al. "Poaceae pollen in the atmosphere of Zagreb (Croatia), 2002–2005." In: *Grana* 45 (2005), pp. 130–136.

[109] X. Quero et al. "Bases Científico-técnicas para un Plan Nacional de mejora de la calidad del aire." In: *Informes CSIC.* (2012).

[110] A. Rakotomamonjy. "Variable Selection Using SVM-based Criteria." In: *Journal of Machine Learning* 3 (2003), pp. 1357–1370.

[111] A. Rantio-Lehtimäki et al. "Significance of sampling height of airborne particles for aerobiological information." In: *Allergy* 46 (1991), pp. 68–76.

[112] H. Ribeiro, M. Cunha, and I. Abreu. "Definition of main pollen season using logistic model." In: *Ann Agric Environ Med* 14 (2007), pp. 259–264.

[113] F.J. Rodríguez-Rajo, A. Dopazo, and V. Jato. "Environmental factors affecting the start of pollen season and concentrations of airborne Alnus pollen in two localities of Galicia (NW Spain)". In: *Ann Agric Environ Med* 11 (2004), pp. 35–44.

[114] F.J. Rodríguez-Rajo, G. Frenguelli, and M.V. Jato. "Effect of air temperature on forecasting the start of the Betula pollen season at two contrasting sites in the south of Europe (1995-2001)". In: *Int J. of Biometeorology* 47 (1983), pp. 117–125.

[115] F.J. Rodríguez-Rajo et al. "Prediction of airborne Alnus pollen concentration by using ARIMA models." In: *Ann Agric Environ Med* 13 (2006), pp. 25–32.

[116] Paul G. Rogers. "The Clean Air Act of 1970". In: *EPA Journal* (1990).

[117] E. Roldán et al. "The effect of climate-change-related heat waves on mortality in Spain: Uncertainties in health on a local scale." In: *Stochastic Research and Risk Assessment* 30 (2016), pp. 831–839.

[118] S. Ruder. "An overview of gradient descent optimization algorithms". In: *CoRR* abs/1609.04747 (2016). URL: http://arxiv.org/abs/1609.04747.

[119] D. E. Rumelhart, G. E. Hinton, and R. J. Ronald. "Learning representations by back-propagating errors". In: *Nature* 323 (1986), pp. 533–536.

[120] William F. Ryan. "The air quality forecast rote: Recent changes and future challenges". In: *Journal of the Air & Waste Management Association* 66.6 (2016), pp. 576–596. DOI: 10.1080/10962247.2016.1151469.

[121] S. Sabariego et al. "Models for forecasting airborne Cupressaceae pollen levels in central Spain." In: *Int J Biometeorol* 56 (2012), pp. 253–258.

[122] S. Sabour, N. Frosst, and GE. Hinton. "Dynamic Routing Between Capsules". In: *NIPS*. 2017.

[123] J.A. Sánchez-Mesa et al. "Characteristics of grass pollen seasons in areas of southern Spain and the United Kingdom". In: *Aerobiologia* 19 (2003), pp. 243–250.

[124] J. Schaber and F-W. Badeck. "Physiology-based phenology models for forest tree species in Germany." In: *Int J Biometeorol* 47 (2003), pp. 193–201.

[125] J. Shaffer. "Modified sequentially rejective multiple test procedures." In: *J. of AmericanStatistical Association* 81 (1986), pp. 826–831.

[126] Zhigen Shang et al. "A novel model for hourly PM2.5 concentration prediction based on CART and EELM". In: *Science of The Total Environment* 651 (2019), pp. 3043 –3052. ISSN: 0048-9697. DOI: https://doi.org/10.1016/j.scitotenv.2018.10.193. URL: http://www.sciencedirect.com/science/article/pii/S0048969718340841.

[127] S. Sharma et al. "Statistical behavior of ozone in urban environment." In: *Sustainable Evironment Research* (2016), pp. 142–148. DOI: 10.1016/j.serj.2016.04.006.

[128] I. Silva-Palacios et al. "Temporal modelling and forecasting of the airborne pollen of Cupressaceae on the southwestern Iberian peninsula." In: *Int J Biometeorol* 60 (2016), pp. 1509–1517.

[129] M. Smith and J. Emberlin. "A 30-day-ahead forecast model for grass pollen in north London, UK." In: *Int J. Biometeorology* 50 (2006), pp. 233–242.

[130] M. Sofiev and K.C. Bergmann. "Allergenic Pollen: A review of the production, release, distribution and health impacts." In: Springer Science and Business Media, 2012. Chap. Impact of pollen, pp. 161–215.

[131] M. Sofiev et al. "A dispersion modelling system SILAM and its evaluation against ETEX data". In: *Atmospheric Environment* 40 (2006), pp. 674–685.

[132] M. Sofiev et al. "A numerical model of birch pollen emission and dispersion in the atmosphere. Description of the emission module". In: *Int J Biometeorol* 57 (2013), pp. 45–58.

[133] J. Subiza et al. "Allergenic pollen pollinosis in Madrid." In: *J Allergy and Clinical Immunology* 96 (1995), pp. 15–23.

[134] F. Tassan-Mazzocco, A. Felluga, and P. Verardo. "Prediction of wind-carried Gramineae and Urticaceae pollen occurrence in the Friuli Venezia Giulia region (Italy)". In: *Aerobiologia* 31 (2015), pp. 559–574.

[135] James G. Titus. "Greenhouse effect, sea level rise, and barrier Islands: Case study of long beach Island, New Jersey". In: *Coastal Management* 18.1 (1990), pp. 65–90. DOI: 10.1080/08920759009362101.

[136] UNFCCC. *Kyoto Protocol to the United Nations Framework Convention on Climate Change adopted at COP3 in Kyoto, Japan*. Tech. rep. 1997.

[137] D. Valput, R. Navares, and JL. Aznarte. "Forecasting hourly NO2 concentrations by ensembling neural networks and mesoscale models". In: *Neural Computing and Applications* (2019). ISSN: 1433-3058. DOI: 10.1007/s00521-019-04442-z. URL: https://doi.org/10.1007/s00521-019-04442-z.

[138] H. Vogel, A. Pauling, and B. Vogel. "Numerical simulation of birch pollen dispersion with an operational weather forecast system". In: *Int J Biometeorol* 52 (2008), pp. 805–814.

[139] A. Wagner et al. "Evaluation of the MACC operational forecast system potential and challenges of global near-real-time modelling with respect to reactive gases in the troposphere." In: *Atmos. Chem. Phys.* 15 (2015), pp. 14005–14030.

[140] P. K. Van de Water et al. "An assessment of predictive forecasting of Juniperus ashei pollen movement in the Southern Great Plains, USA". In: *Int J Biometeorol* 48 (2003), pp. 74–82.

[141] Letty A. de Weger et al. "Impact of Pollen". en. In: *Allergenic Pollen*. Ed. by Mikhail Sofiev and Karl-Christian Bergmann. DOI: 10.1007/978-94-007-4881-1_6. Springer Netherlands, 2013, pp. 161–215. ISBN: 978-94-007-4880-4 978-94-007-4881-1. (Visited on 08/22/2016).

[142] David H. Wolpert. "Stacked Generalization". In: *Neural Networks* 5 (1992), pp. 241–259.

[143] M. Yousefi et al. "Chaotic genetic algorithm and Adaboost ensemble metamodeling approach for optimum resource planning in emergency departments". In: *Artificial Intelligence in Medicine* 84 (2018), pp. 23–33. DOI: https://doi.org/10.1016/j.artmed.2017.10.002.

[144] Yang Zhang et al. "Real-time air quality forecasting, part I: History, techniques, and current status". In: *Atmospheric Environment* 60 (2012), pp. 632 –655. ISSN: 1352-2310. DOI: https://doi.org/10.1016/j.atmosenv.2012.06.031.

[145] T. Zhu et al. "Time Series Approaches for Forecasting the Number of Hospital Daily Discharged Inpatients." In: *IEEE J Biomed Health Inform* (2015). DOI: 10.1109/JBHI.2015.2511820.

[146] Chiara Ziello et al. "Changes to Airborne Pollen Counts across Europe". In: *PLOS ONE* 7.4 (Apr. 2012). DOI: 10.1371/journal.pone.0034076. URL: https://doi.org/10.1371/journal.pone.0034076.