Environmental Pollution 229 (2017) 321–328

Contents lists available at ScienceDirect

Environmental Pollution

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

Probabilistic forecasting for extreme NO_2 pollution episodes☆

José L. Aznarte [1]

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

ARTICLE INFO

Article history:
Received 8 November 2016
Received in revised form
23 May 2017
Accepted 28 May 2017
Available online 10 June 2017

Keywords:
Probabilistic forecasting
Air quality
Quantile regression
Nitrogen dioxide
Madrid

ABSTRACT

In this study, we investigate the convenience of quantile regression to predict extreme concentrations of NO_2. Contrarily to the usual point-forecasting, where a single value is forecast for each horizon, probabilistic forecasting through quantile regression allows for the prediction of the full probability distribution, which in turn allows to build models specifically fit for the tails of this distribution.

Using data from the city of Madrid, including NO_2 concentrations as well as meteorological measures, we build models that predict extreme NO_2 concentrations, outperforming point-forecasting alternatives, and we prove that the predictions are accurate, reliable and sharp. Besides, we study the relative importance of the independent variables involved, and show how the important variables for the median quantile are different than those important for the upper quantiles. Furthermore, we present a method to compute the probability of exceedance of thresholds, which is a simple and comprehensible manner to present probabilistic forecasts maximizing their usefulness.

1. Introduction

With increasing problematic pollution levels in cities around the world, and given the scientific consensus about their adverse effects on health (Kim et al., 2015; Sellier et al., 2014), traffic restrictions emerge as a temporary remedy when high pollution episodes occur. To comply with European regulations (European Commission, 2008) the city of Madrid has enforced a new air quality protocol which includes restrictions to the use of polluting vehicles when NO_2 concentrations reach certain thresholds. Anticipating the activation of such restrictions is critical both for the decision makers (which need to announce them in advance) and for the vehicle owners (which need to plan their transport alternatives).

Notwithstanding, research on forecasting extreme pollution events is meagre in general and, in particular, to our knowledge, probabilistic forecasting (the prediction of the full future distribution of a magnitude, as opposed to point forecasting) has never been put in practice to deal with NO_2 concentrations. Forecasting the central tendencies of the data distribution, i.e. the conditional mean, is not the best approach if the main interest is to predict possible exceedances of thresholds that lie on the tails of the distribution.

Air quality forecasting is an active field of study which gathers contributions from meteorology, physics, chemistry, statistics and computational intelligence (Zhang et al., 2012). There are several approaches to air quality forecasting, roughly divided in two classes: those which rely upon analyzing the atmosphere status and evolution from a fluid mechanics and chemical point of view and those which study pollution measures from a statistical time series analysis perspective. The latter can be subsequently divided in approaches that assume that the data are generated by a given stochastic data model and those that use algorithmic models and treat the data mechanisms as unknown (Breiman, 2001a). These algorithmic models can be included in what is called computational intelligence (CI) and are applicable to many different forecasting problems, including air quality.

However, previous research (excluding extreme value theory (Thompson et al., 2001)) has focused mostly on the central tendencies of the data distribution. This is also the case for CI-based forecasting of air quality (Gardner and Dorling, 1998; Wang et al., 2003) and its applications in different parts of the world: London (Gardner and Dorling, 1999), Santiago de Chile (Perez and Trier, 2001), Helsinki (Kukkonen et al., 2003), Bilbao (Agirre-Basurko et al., 2006), Palermo (Brunelli et al., 2007) and Athens (Vlachogianni et al., 2011).

Although point forecasts are widely used, they have some

☆ This paper has been recommended for acceptance by Eddy Y. Zeng.

E-mail address: jlaznarte@dia.uned.es.

[1] This work has been partially funded by Ministerio de Economía y Competitividad, Gobierno de España, through a Ramón y Cajal grant (reference: RYC-2012–11984).

http://dx.doi.org/10.1016/j.envpol.2017.05.079
0269-7491/© 2017 Elsevier Ltd. All rights reserved.

obvious disadvantages. For example, they do not readily inform about the inherent uncertainty of the predictions, and are generally unsuitable for the cases of heavily skewed data or if there is a need to examine certain important strata of the series (Koenker, 2005).

Hence, considering the complex nature of the interactions between meteorological and human factors which affect air quality, it can be problematic to assume that the relationship between those factors and the concentrations of airborne pollutants is the same for unusually low concentrations as for unusually high (peak) concentrations. And it can be even more problematic to assume that both relationships are of the same form as for the central part of the conditional distribution. Furthermore, there is no need for the explanatory variables used in forecasting the concentrations of airborne pollutants on the tails of a conditional distribution to be the same as the explanatory variables used in forecasting the expected concentrations or point-forecasts.

This is especially true when forecasting air quality in the framework of anti-pollution regulations, which impose certain actions that must be taken when pollutants exceed thresholds set by the authorities. Modelling the upper quantiles of the conditional distribution becomes a necessity in this case. In addition, modelling the full conditional distribution allows to obtain estimations of the probability of exceedance of the thresholds, which is a more useful estimate in terms of communicative power to the general public, as shown by its extensive use in meteorological forecasting (Raftery,). Two alternative applications of these ideas are (Balashov et al., 2017; Garner and Thompson, 2013).

2. Probabilistic forecasting with quantile regression

The prediction from most regression models is an estimate of the conditional mean of a dependent variable, or response, given a set of independent variables or predictors. However, the conditional mean measures only the center of the conditional distribution of the response, and if we need a more complete summary of this distribution, for example in order to estimate the associated uncertainty, quantiles are in order. The 0.5 quantile (i.e., the median) can serve as a measure of the center, and the 0.9 quantile marks the value of the response below which reside the 90% of the predicted points. Recent advances in computing have inducted the development of regression models for predicting given quantiles of a conditional distribution, using a technique called quantile regression (Roger Koenker, 1978).

Quantile regression (QR) has gained an increasing attention from diverse scientific disciplines (Yu et al., 2003), including financial and economic applications (Fitzenberger et al., 2002), medical applications (Soyiri et al., 2012), wind power forecasting (Zhang et al., 2014), electric load forecasting (Taieb et al., 2016; Gibbons and Faruqui, 2014), environmental modelling (Cade and Noon, 2003) and meteorological modelling (Bjornar Bremnes, 2004). To our knowledge, despite its success in other areas, quantile regression has not been applied in the framework of NO_2 forecasting.

As an illustration of the concept (profusely discussed in (Koenker, 2005)), given a set of vectors (x_i, y_i), in point forecasting we are usually interested in what prediction $\hat{y}(x) = \alpha_0 + \alpha_1 x$ minimizes the mean squared error,

$$E = \frac{1}{n}\sum_i^n \varepsilon_i = \frac{1}{n}\sum_i^n [y_i - (\alpha_0 + \alpha_1 x)]^2. \tag{1}$$

This prediction is the conditional sample mean of y given x, or the location of the conditional distribution. But we could be interested in estimating the conditional median (i.e., the 0.5 quantile) instead of the mean, in which case we should find the prediction $\hat{y}(x)$ which minimizes the mean absolute error,

$$E = \frac{1}{n}\sum_i^n \varepsilon_i = \frac{1}{n}\sum_i^n \left| y_i - (\alpha_0 + \alpha_1 x) \right|. \tag{2}$$

The fact is that, apart from the 0.5 quantile, it is possible to estimate any other given quantile τ. In that case, instead of (2), we could minimize

$$E = \frac{1}{n}\sum_i^n f(y_i - (\alpha_0 + \alpha_1 x)) \tag{3}$$

where

$$f(y-q) = \left\{ \begin{array}{cc} \tau(y-q) & \text{if } y \geq q \\ (1-\tau)(q-y) & \text{if } y < q \end{array} \right\}, \tag{4}$$

with $\tau \in (0, 1)$. Equation (3) represents the median when $\tau = 0.5$ and the τ-th quantile in any other case.

Thus, if we can estimate an arbitrary quantile and forecast its values, we can also estimate the full conditional distribution. Among the array of methods that allow to estimate and forecast data-driven conditional quantiles, in this study we have chosen quantile regression forests for its ease of use (few parameters have to be chosen) and for its availability in the free software mathematical environment R (Core Team, 2015). For a detailed discussion on quantile regression forests, see (Meinshausen, 2006).

3. Data description and experimental design

3.1. Protocol for high NO_2 concentration episodes

Complying with European regulations (European Commission, 2008), the city of Madrid has an atmospheric pollution monitoring system including 24 stations around the city. The data gathered by this system are public and are made available on an hourly basis (Madrida). In 2016, the local government imposed new anti-pollution measures in a protocol (Madridb) which include traffic restrictions when NO_2 concentrations reach the thresholds set by the EU. Concretely, this protocol establishes three action levels that are raised according to hourly average NO_2 concentrations: a *pre-warning* for breaches of a threshold of 180 $\mu g/m^3$, a *warning* when concentrations are over 200 $\mu g/m^3$ and an *alert* when values over 400 $\mu g/m^3$ are registered. The city is divided into 5 zones and, in order to activate the different action levels, these limits must be violated in at least two stations of the same zone during at least two consecutive hours.

In this paper, we will deal with predicting the probability of a single station breaching the limits, leaving for future works the computation of aggregated probabilities of pairs of stations in order to predict the activation of the restrictions imposed by the protocol. The forecasting model currently used by the city is a point-forecasting one, and hence cannot predict the probability of the breaching of these thresholds.

In this paper, we will center our attention on NO_2 concentrations over the pre-warning threshold. However, the results are applicable to other thresholds for NO_2 and other pollutant species like ozone or particulate matter and their respective regulatory limits.

3.2. Nitrogen dioxide data

For this study, from the 24 stations of Madrid's monitoring system we chose the Plaza de España station. This station is known to register a high proportion of peak values exceeding the NO_2

safety limits, thus frequently contributing to the activation of the protocol. The time series for Plaza de España consists of hourly measured values of the concentrations of NO_2 from 01/01/2000 to 30/11/2015. As can be seen in Fig. 1, these values exhibit a clear intraday pattern, in which the higher values are located in two peaks around the morning and evening (with highest average value at 19 h) while the night hours (from 00 h to 05 h) have lower average concentrations. In the same figure we can see that the distribution of NO_2 presents a positive skew, with values over the thresholds being rare. In fact, the 99.8% of the data are equal or below the pre-warning threshold, while a total amount of 1379 points exceed it.

3.3. *Weather data*

Apart from air quality data, some of the monitoring stations of the atmospheric monitoring system also register meteorological variables as hourly average temperature and wind speed, hourly accumulated rainfall and barometric pressure. These variables encode an approximate picture of the local atmospheric situation, and hence they are related to the variability of the NO_2 concentrations (Arain et al., 2009). The Plaza de España station registers these variables except for barometric pressure. Because typically the variations in pressure are small across the city, we use the pressure data from another nearby meteorological station: Casa de Campo (approximately 3 km away from Plaza de España). In order to account for changes in the atmosphere status, the hourly differences of pressure were used instead of the original series.

3.4. *Experimental design*

Fig. 2 shows a diagram of the process. Once all the data were gathered, the series were aligned and merged using their respective time dimension, and UTC times were used for all of them. Then, each of them was lagged in order to create new variables to take into account the possible autocorrelative features of the series, in the following manner: for NO_2, a set of embedding dimensions or lags was selected with the objective to take into account the recent past hours, the analogous hours of the day before and the analogous hours of the same day of the week before. That is, $\mathbf{d}_{NO_2} = (1, 2, 3, 4, 24, 25, 26, 27, 168, 169, 170, 171)$. For the meteorological variables we selected only the past hours: $\mathbf{d}_{met} = (1, 2, 3, 4)$. Then, for each hour t in the 15 years covered by the data, a vector was constructed as $\mathbf{z}_t = (y_{t-\mathbf{d}_{NO_2}}, x^{Temp}_{t-\mathbf{d}_{met}}, x^{Wind}_{t-\mathbf{d}_{met}}, x^{Rain}_{t-\mathbf{d}_{met}}, x^{Press}_{t-\mathbf{d}_{met}}; y_{t+h})$.

To account for periodic effects, the time of day, the weekday and the day of the year were also included as dummy variables in the set of predictors. As a result, a matrix with 139,513 rows and 33 columns (32 predictor variables and one response variable y_{t+h}) was constructed. Throughout this paper, we take $h = 1$, although h can take any positive integer value: in order to forecast 12 h in advance, it would suffice to set $h = 12$.

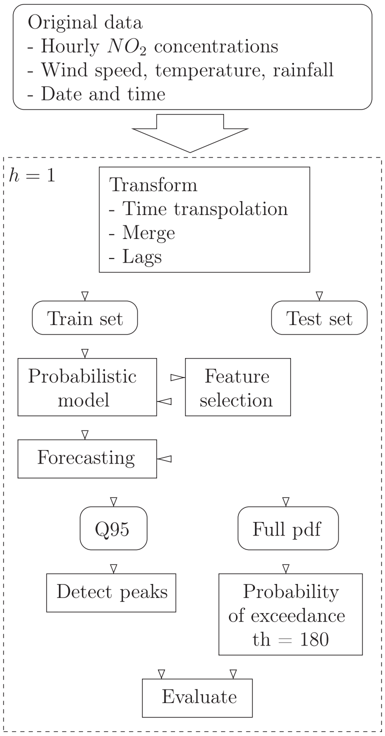

Fig. 2. Diagram of the experimental design.

Finally, in order to avoid overfitting, and assuming that NO_2 concentrations remained stable during the whole period, a standard validation scheme was followed. We divided our data in two blocks: a block from 01/01/2000 to 31/12/2009 will be used to train the models, and the rest of the data, from 01/01/2010 to 31/12/2015 will be used to test their properties.

3.5. *Evaluation of probabilistic forecasts*

The evaluation of probabilistic forecasts is an open issue in the literature (McSharry et al., 2009). Given the fact that we can derive expected values from probabilistic forecasts, the simpler way is to use the standard metrics for evaluating point forecasts, as root mean squared error (RMSE), mean average error (MAE), correlation or bias. These measures can be compared against some reference models, to establish how the probabilistic forecasts perform when forecasting expected values in the central part of the distribution. In this work, we used three point forecasts models as benchmark: persistence, a linear regression and random forests.

Persistence, also known as naive predictor, is the simplest way of producing a forecast. The persistence forecasts are obtained by

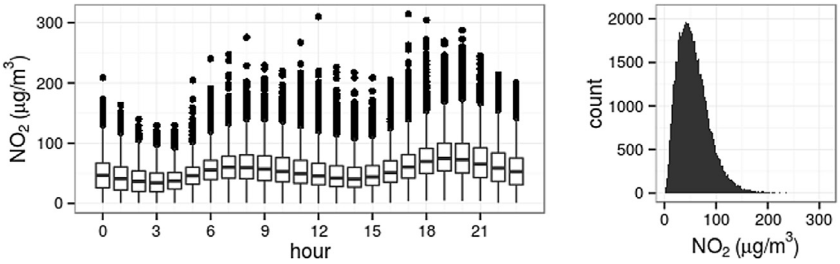

Fig. 1. Intraday distribution (left) and histogram (right) of the concentration of NO_2.

issuing the last observation as the forecast for all future horizons: $\widehat{y}_{t+h|t} = y_t$. This method works well when the patterns of the series change very slowly with time. Another simple method of producing forecasts is through a least-squares linear regression which combines the predictors $\mathbf{z}_t$ (as defined in section 3.4) in a linear manner to compute estimations for future values of the series. Finally, to obtain a more state-of-the-art benchmark, a nonlinear model from the machine learning field was selected and trained to point-forecast future values of the series. Among the panoply of nonlinear machine learning models available, we selected Random Forests (Breiman, 2001b), which have been used successfully to forecast air quality before (Yu et al.,; Yang et al.,2016). We used the R implementation (Liaw and Wiener, 2002) with its default parameters.

However, on the other hand, if non-central regions of the forecast distributions are to be evaluated, more elaborated metrics must be used. Amongst the main attributes of probabilistic forecasts used for evaluation, reliability and sharpness are two of the most common (McSharry et al., 2009), and will be used in Section 4.3.

Reliability deals with how close the actual distribution of the data is to the predicted one and is related to the unconditional coverage of a prediction interval. That is, if a predicted 50% interval covers 50% of the observed load values, then it is considered reliable.

However, reliability is not sufficient to characterize the quality of a probabilistic forecast since a forecast based on climatology is perfectly reliable and yet has no skill. A model is said to have no skill when it provides the same forecast distribution for all situations. A skillful model will provide sharper distributions for more certain situations and wider distribution when the uncertainty on the outcome is higher. Hence sharpness deals with how tightly the predicted distribution covers the actual one. A 95% interval is said to be sharp if the maximum and minimum values of the observed values are very close to the upper and lower bounds of the predicted interval.

Reliability is related to sharpness in the same way bias is related to variance in deterministic forecast evaluation: there is a sharpness-reliability performance trade-off in the same way that there is a bias-variance trade-off for point forecast models.

3.6. Evaluation of alert forecasting

The evaluation of binary forecasts (forecasts of events which might occur or not) is usually performed through summary statistics of the contingency table, from which different ways of measuring the goodness of an alert forecast have been proposed (McSharry et al., 2009). A contingency table arranges the four different outcomes which are expected from an alert forecast: true positives (the alert is forecast and it actually happens, noted as *TP*), false positives (the alert is forecast but it does not happen, *FP*), true negatives (no alert is forecast and none happens, *TN*) and false negatives (no alert is forecast but it actually happens, *FN*).

Two available performance measures are specificity and sensitivity. Specificity, also known as true negative rate, measures the proportion of negatives that are correctly forecast as such: $\text{spec} = TN/(TN + FP)$. Sensitivity, also called true positive rate, is the fraction of positives that were correctly forecast: $\text{sens} = TP/(TP + FN)$.

However, predicting extreme values is a special situation in which the imbalance of the different types of events plays a crucial role (Ferro and Stephenson, 2011). Both specificity and sensitivity behave poorly in the framework of problems which have few true positives with respect to the total, which is our case. Hence, some other measures have been proposed to account for imbalanced contingency tables:

Balanced accuracy represents the average accuracy (number of correct predictions divided by number of predictions) in each class: $\text{BAcc} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)$. Ranges from 0 to 1, with 0 indicating no accuracy.

True skill statistic measures how well did the forecast separate the positive from the negative events: $\text{TSS} = \frac{TP}{TP+FN} - \frac{FP}{TN+FP}$. Ranges from −1 to 1, and 0 indicates no skill.

Extreme dependency score compares the fraction of the observed events with the fraction of the correctly forecast events. $\text{EDS} = 2\left(\ln\left(\frac{TP+FP}{n}\right)\right) \Big/ \left(\ln\left(\frac{TP}{n}\right)\right) - 1,$ where $n = TP + FN + TN + FN$. Ranges from −1 to 1, perfect score is 1.

Finally, another available tool to evaluate alert forecasts is the receiver operating characteristic or ROC curve (Powers, 2011), which is a graphical plot that shows the performance of a binary classifier as its discrimination threshold is varied. The curve is created by plotting the true positive rate (TPR or sensitivity) against the false positive rate (FPR or 1− specificity) at various threshold settings.

4. Results and discussion

This section summarizes the four experiments performed. First we compare the expected values obtained through probabilistic forecasts with reference point forecasts models. Secondly, we study the relative importance associated with the variables used in the first experiment, with the aim to discover any distinctive pattern in how each variable is used for the different regions of the forecast distributions. In the third experiment, we center our attention in the prediction of the extreme values which exceed the regulation's thresholds, while in the fourth we show how the probability of alerts can be predicted and presented.

4.1. Reference models

In the first experiment, we used quantile regression to compute forecasts of the expected value (median) for NO_2 concentrations. Table 1 shows error indicators for the aforementioned reference models and the median forecast by the probabilistic model. As we can see, the median-based model Q50 behaves well in general compared to the other models, being especially good in terms of MAE and bias. This might be related to the median being more robust than the mean in the presence of outliers. However, in this framework, we are interested in those outliers, as they precisely are the values which trigger the activation of the air quality protocol.

4.2. Variable importance

Quantile regression forests offer a straightforward method for feature selection: the method directly computes measures of importance for each variable. We will use it in this second experiment to investigate which variables are more important for the

Table 1
Point forecast error measures for reference models persistence, linear regression, random forests and median of the probabilistic model (QRF).

	RMSE	MAE	Bias	Corr
Persistence	13.47	9.23	0.04	0.88
LR	11.51	8.16	−1.62	0.91
RF	11.27	7.89	−2.14	0.92
Q50	11.30	7.63	−0.27	0.91

problem under consideration.

The variable importance concept (Breiman, 2001b) is based in a direct measure of the impact of each feature on the accuracy of the model. This is done through a random alteration of the values of each feature. Once this alteration is done, it is possible to measure how much it decreases the accuracy of the model. For unimportant variables, this should have little or no effect on the model accuracy, while altering important variables should significantly decrease it.

In Fig. 3 we can see the importance of each of the 31 variables considered by the model according to their mean decrease in accuracy, for quantiles 0.05, 0.5 and 0.95. The variables are sorted according to their importance in the latter quantile, as this is the region of the distribution that we are most interested in.

In accordance to the hypothesis of autocorrelation stated in Section 3.4, the NO_2 value measured the hour before (NO2_1) is the most important variable in the three quantiles. The value of NO_2 measured at the same hour of the day before (NO2_24), is the second variable in importance for all the three considered quantiles, which accounts for the intra-daily effect shown in Fig. 1.

If we center our attention in the 0.95 quantile, it is remarkable that the difference of barometric pressure (difPres) does not seem to be as important as expected, although it is clearly more important in the upper quantile than in the other two. Concerning wind, we can see how the wind speed registered at time $t-1$ (WindSpeed_1) is of high importance while other, more old wind speed measures are not. This hints at the immediate washing effect of high wind speeds, which carry NO_2 outside the city (its effect being even more important for quantile 0.05).

Finally, it is interesting to highlight the differences amongst the three considered quantiles, which prove the need and usefulness of quantile forecasting of NO_2. The distribution of variable importances for the median quantile, in particular, shows a separate pattern from the upper and lower quantiles, being the case that the difference between the value of NO_2 at time $t-1$ and the rest of the variables is much more acute for the median quantile than for the others.

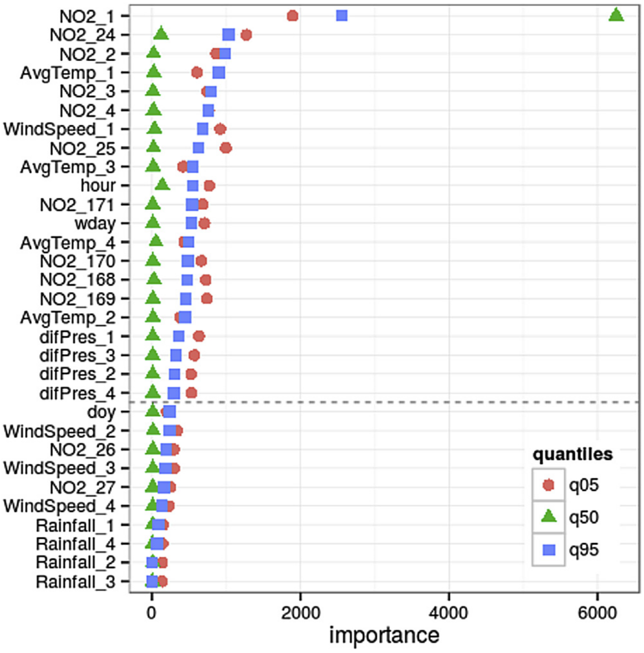

Fig. 3. Importance of the variables in quantiles 0.05, 0.50 and 0.95, in descending order according to the importance in the higher quantile.

4.3. Probabilistic forecasting of extreme values

The third experiment consisted in modelling the full distribution of NO_2 concentrations using quantile regression forests in order to investigate its usefulness to forecast extreme values.

A reliability diagram (Hartmann et al., 2002) is shown in the left part of Fig. 4. We see that the reliability of the forecasts is satisfactory for all considered probabilities. The model predictions are centered around the ideal values, although the model slightly over-predicts for low probabilities and under-predicts for higher probabilities. This is not necessarily negative in the frame of NO_2 forecasting: it means that the model is slightly conservative and this can be considered as a good property since the cost of false alarms is high.

In the right part of Fig. 4, the sharpness of the forecasts, for five different coverage intervals (90%, 70%, 50%, 30% and 10%) is presented. As expected, the inter-quantile range decreases with the width of the interval, while for all coverage rates it increases with forecast horizon due to the increasing forecast uncertainty. Also, the forecast can be said to be skillful since the minimum and especially the maximum observed inter-quantile distances are significantly different from the median values.

In order to show the performance of the model when forecasting extreme values, Fig. 5 shows the 0.95 quantile forecasts for the last six years of the NO_2 series. As we can see, the forecasts for this model (represented by the grey line) acts as an upper envelope for the series, which results in peaks over the 180 $\mu g/m^3$ warning threshold being properly forecast in many cases. This is especially true in the event of high concentrations episodes, as those occurred in late 2014 and around the start of 2015, for example.

Table 2 shows the contingency table and the performance measures described in Section 3.6 for the various models considered. The first two columns correspond to the persistence and the linear regression models. The third and fourth column correspond to the 0.50 and 0.95 quantiles, respectively. The fifth column will be analyzed below.

The first thing we verify in the table is that, for all the models, the amount of true negatives (values below the threshold) largely exceeds the amount of true positives. This causes the specificity to be trivially almost perfect in all the cases, an evidence of why this measure is not very useful in this context. However, in terms of balanced accuracy (*BAcc*), true skill score (*TSS*) and extreme dependency score (*EDS*), there is a clear pattern which differentiates the different mean or median forecasts from the 0.95 quantile forecast. As expected, the latter manages to predict a higher number of true positives, whereas incurring in far less false negatives. This is reflected by the three performance measures, the 0.95 quantile forecast consistently outperforming the other models.

As stated above, quantile regression allows for the prediction of any arbitrary quantile. Hence, in order to predict the protocol alerts, corresponding to the 180 $\mu g/m^3$ threshold, we might as well compute which quantile does this threshold define in the full distribution of the data. In fact, it corresponds to the 0.995 quantile, and although conventional large sample theory for quantile regression does not apply sufficiently far in the tails (for such extreme quantiles there usually are not enough data to guarantee that the model is correct), we have used QRF to predict it: it corresponds to the fifth column of Table 2. It is shown in this column that the model manages to predict extreme values outperforming the 0.95 quantile in all the measures except for specificity. Out of the 68 alerts registered, this model manages to properly predict 64, although it also produces around three times more false positives than the 0.95 quantile. This is a clear limitation of the proposal which should be investigated further by adding, for example, numerical weather or pollution predictions.

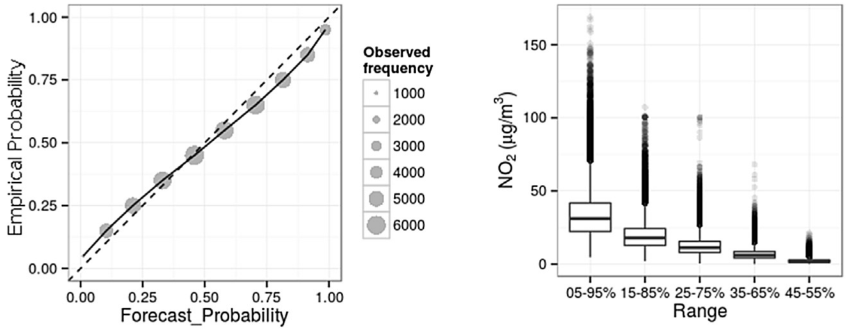

Fig. 4. Reliability (left) and sharpness (right).

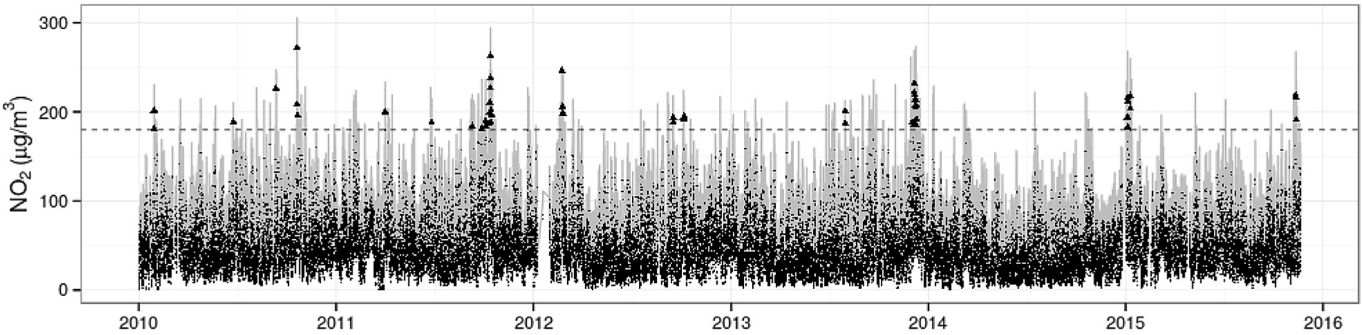

Fig. 5. Forecasts for quantile 0.95. The horizontal line marks the threshold of 180, while the grey line represents the forecast values for quantile 0.95, small dots are observed values and triangles represent correctly forecast peak values.

Table 2
Performance measures (as defined in Section 3.6) for the considered models over the test set.

	Persist.	LR	Q50	Q95	th = 180
TP	24	25	15	59	64
FP	45	27	12	309	1006
TN	47.884	47.902	47.917	47.620	46.923
FN	44	43	53	9	4
spec	1.00	1.00	1.00	0.99	0.98
sens	0.35	0.37	0.22	0.87	0.94
BAcc	0.68	0.68	0.61	0.93	0.96
TSS	0.35	0.37	0.22	0.86	0.92
EDS	0.73	0.74	0.63	0.96	0.98

Finally, in Fig. 6, the standard ROC curves for the 0.50 and 0.95 quantile forecasts are shown. As we can see, the point of the curve corresponding to the threshold of 180 $\mu g/m^3$ (marked by the dashed lines) shows much better properties in the case of the 0.95 quantile than for the 0.50 quantile.

4.4. Forecasting the probability of alerts

The fourth experiment is an example of an elaboration over the probabilistic results of the models that could help decision-making authorities to better understand the forecasts and adapt their policies to them. Fig. 7 shows, in two different fashions, the predictions for a period in November 2015, in which a high NO_2 concentrations episode was registered.

The upper part of the figure is what is known in the meteorology field as an “EPSgram”. EPSgrams (which take its name from the Ensemble Prediction System of the ECMWF) display the time evolution of the distribution of a magnitude. In the graph, we can see how the forecast NO_2 distribution varied during the episode. The forecast distribution at each forecast time is represented by a box-and whiskers plot showing the median (short horizontal line), the 25th and 75th percentiles (vertical box) and 5th and 95th percentiles (vertical lines).

In order to provide a more easily interpretable and understandable representation, the lower part of Fig. 7 shows, for each forecast, the probability of exceedance for the 180 $\mu g/m^3$ threshold which triggers the first alert in the NO_2 protocol.

5. Conclusions

In this paper, a first application of probabilistic forecasting to the problem of predicting extreme NO_2 pollution episodes has been presented. Data from the city of Madrid have been used to develop quantile regression models tailored to predict NO_2 concentrations in an urban location. Through four different experiments, it has been shown that probabilistic forecasting using quantile regression has advantages over traditional point forecasting, allowing for a better prediction of extreme concentrations and also a more insightful representation of the predictions. The proposed approach allows for the prediction of the whole distribution of the future values of NO_2 concentrations. This is especially useful when, as is the case, the tails of the distribution are of interest, allowing for a more precise prediction of extreme values.

We have shown how, when used to predict the median, the proposed model compares favourably to point-forecasting approaches including simple models as persistence or linear regression as well as complex and highly nonlinear models as random forests. On the other hand, we have shown the accuracy and usefulness of the probabilistic predictions, which have good sharpness and reliability. We have used the forecast upper quantiles to predict high NO_2 concentrations and to produce alerts in the air quality protocol of the municipality of Madrid. The results indicate that our

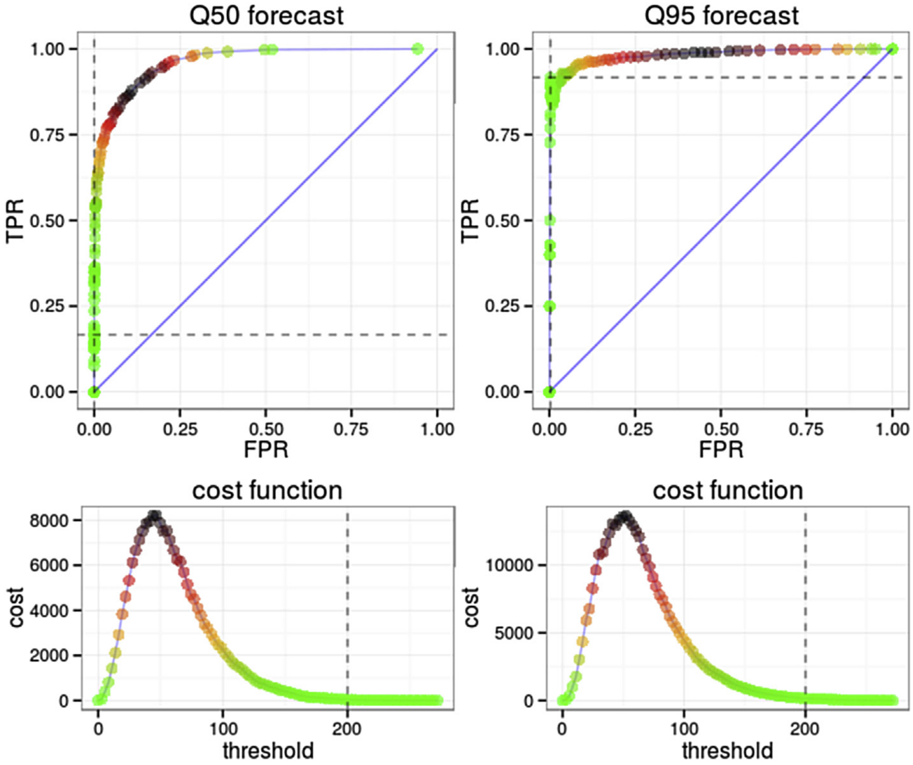

Fig. 6. ROC curve for the two considered quantile forecasts. The dashed lines indicate the location of the (TPR, FPR) point corresponding to a threshold of 180 $\mu g/m^3$. TPR is true positive rate and FPR false positive rate.

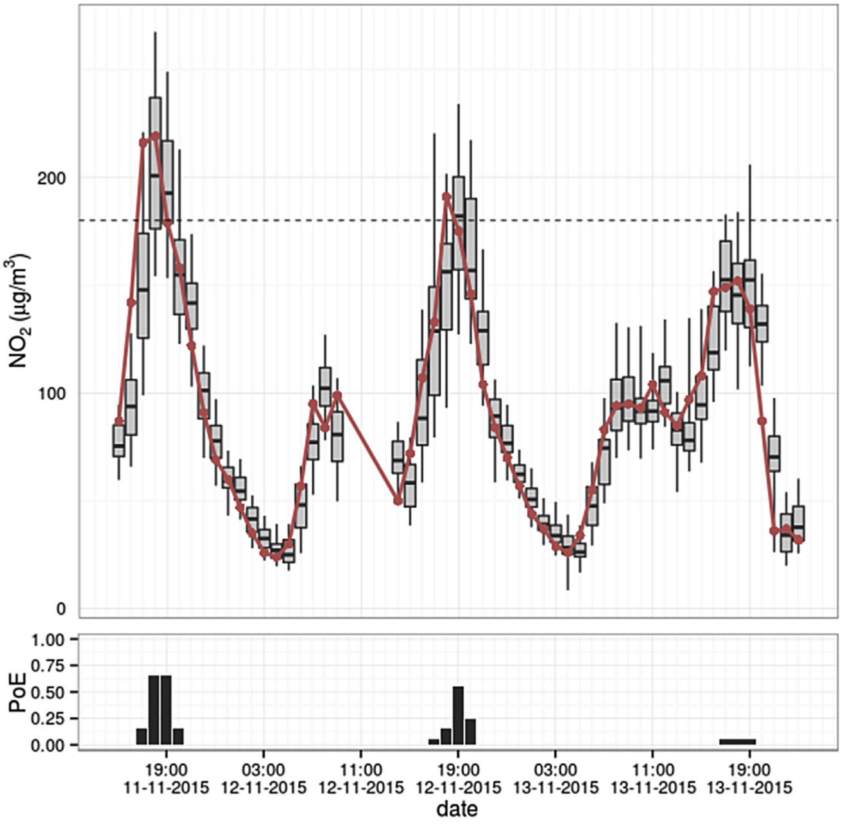

Fig. 7. For a period in November 2015, the upper graph represents observed values (red line), alert threshold (dashed horizontal line) and distribution of the forecasts (boxplots). The lower graph shows probability of exceedance of the threshold. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

proposal is able to predict these episodes in a much more dependable way.

As a by-product of the chosen model, we performed a study about the relative importance of the (autoregressive and meteorologic) independent variables available, confirming, in a purely data-driven approach, some of the assumptions made by other authors about the atmospheric interactions affecting the NO_2 concentrations.

Finally, we have shown how probabilistic forecasts can be represented in a comprehensible and intelligible way, thus allowing authorities and decision-makers, as well as the general public, to make a more beneficial use of the predictions.

Extensions to this work, which has been considered by the Municipality of Madrid to renew its predictive operational models, are under development, including spatio-temporal considerations, longer forecasting horizons and the inclusion of other covariates including numerical predictions.

References

Agirre-Basurko, E., Ibarra-Berastegi, G., Madariaga, I., 2006. Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area. Environ. Model. Softw. 21 (4), 430–446. http://dx.doi.org/10.1016/j.envsoft.2004.07.008.

Arain, M.A., Blair, R., Finkelstein, N., Brook, J., Jerrett, M., 2009. Meteorological influences on the spatial and temporal variability of NO2 in Toronto and Hamilton. Can. Geogr./Le Ogr. Can. 53 (2), 165–190. http://dx.doi.org/10.1111/j.1541-0064.2009.00252.x.

Balashov, N.V., Thompson, A.M., Young, G.S., 2017. Probabilistic forecasting of surface ozone with a novel statistical approach. J. Appl. Meteorol. Climatol. 56 (2), 297–316. http://dx.doi.org/10.1175/JAMC-D-16-0110.1.

Bjornar Bremnes, J., 2004. Probabilistic forecasts of precipitation in terms of quantiles using NWP model output. Mon. Weather Rev. 132 (1), 338–347. http://dx.doi.org/10.1175/1520-0493(2004)132<0338:PFOPIT>2.0.CO;2.

Breiman, L., 2001. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Stat. Sci. 16 (3), 199–231. http://dx.doi.org/10.1214/ss/1009213726.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32. http://dx.doi.org/10.1023/A:1010933404324, 10.1023/A:1010933404324.

Brunelli, U., Piazza, V., Pignato, L., Sorbello, F., Vitabile, S., 2007. Two-days ahead prediction of daily maximum concentrations of SO2, O3, PM10, NO2, CO in the urban area of Palermo, Italy. Atmos. Environ. 41 (14), 2967–2995. http://dx.doi.org/10.1016/j.atmosenv.2006.12.013.

Cade, B.S., Noon, B.R., 2003. A gentle introduction to quantile regression for ecologists. Front. Ecol. Environ. 1 (8), 412–420. http://dx.doi.org/10.1890/1540-9295(2003)001[0412:AGITQR]2.0.CO;2.

Core Team, R., 2015. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing. Vienna, Austria. URL https://www.R-project.org/.

European Commission, 2008. Air Quality Standards According to Directive 2008/50/EC.

Ferro, C.A.T., Stephenson, D.B., 2011. Deterministic forecasts of extreme events and warnings. In: Jolliffe, I.T., Stephenson, D.B. (Eds.), Forecast Verification. John Wiley & Sons, Ltd, pp. 185–201. http://dx.doi.org/10.1002/9781119960003.ch10/summary.

Fitzenberger, B., Koenker, R., Machado, J.A.F. (Eds.), 2002. Economic Applications of Quantile Regression. Physica-Verlag HD, Heidelberg.

Gardner, M., Dorling, S., 1998. Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. Atmos. Environ. 32 (14–15), 2627–2636. http://dx.doi.org/10.1016/S1352-2310(97)00447-0.

Gardner, M.W., Dorling, S.R., 1999. Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London. Atmos. Environ. 33 (5), 709–719. http://dx.doi.org/10.1016/S1352-2310(98)00230-1.

Garner, G.G., Thompson, A.M., 2013. Ensemble statistical post-processing of the national air quality forecast capability: enhancing ozone forecasts in Baltimore, Maryland. Atmos. Environ. 81, 517–522. http://dx.doi.org/10.1016/j.atmosenv.2013.09.020.

Gibbons, C., Faruqui, A., Jul. 2014. Quantile Regression for Peak Demand Forecasting, SSRN Scholarly Paper ID 2485657. Social Science Research Network, Rochester, NY.

Hartmann, H.C., Pagano, T.C., Sorooshian, S., Bales, R., 2002. Confidence builders: evaluating seasonal climate forecasts from user perspectives. Bull. Am. Meteorol. Soc. 83 (5), 683–698. http://dx.doi.org/10.1175/1520-0477(2002)083<0683:CBESCF>2.3.CO;2.

Kim, K.-H., Kabir, E., Kabir, S., 2015. A review on the human health impact of airborne particulate matter. Environ. Int. 74, 136–143. http://dx.doi.org/10.1016/j.envint.2014.10.005.

Koenker, R., 2005. Quantile Regression, No. 38 in Econometric Society Monographs. Cambridge University Press.

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. http://dx.doi.org/10.1016/S1352-2310(03)00583-1.

Liaw, A., Wiener, M., 2002. Classification and regression by randomforest. R News 2 (3), 18–22. URL http://CRAN.R-project.org/doc/Rnews/.

Ayuntamiento de Madrid. Sistema de Vigilancia de la Calidad del Aire. URL http://www.mambiente.munimadrid.es/sica/scripts/index.php.

Ayuntamiento de Madrid, Protocolo de medidas a adoptar durante episodios de alta contaminación por dióxido de Nitrógeno. URL http://www.madrid.es/UnidadesDescentralizadas/Sostenibilidad/CalidadAire/Ficheros/ProtocoloSuperaNO2consol.pdf.

McSharry, P.E., Pinson, P., Girard, R., Oct. 2009. Methodology for the Evaluation of Probabilistic Forecasts, Tech. Rep. Deliverable Dp-6.2. European Commission.

Meinshausen, N., 2006. Quantile regression forests. J. Mach. Learn. Res. 7, 983–999.

Perez, P., Trier, A., 2001. Prediction of NO and NO2 concentrations near a street with heavy traffic in Santiago, Chile. Atmos. Environ. 35 (10), 1783–1789. http://dx.doi.org/10.1016/S1352-2310(00)00288-0.

Powers, D.M., 2011. Evaluation: from precision, Recall and F-measure to ROC, informedness, markedness and correlation. J. Mach. Learn. Technol. 2 (1), 37–63.

Raftery A. E.. Use and communication of probabilistic forecasts, arXiv:1408.4812 [stat]ArXiv: 1408.4812. URL http://arxiv.org/abs/1408.4812.

Roger Koenker, G.B., 1978. Regression quantiles. Econometrica 46 (1), 33–50.

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. http://dx.doi.org/10.1016/j.envint.2014.02.001.

Soyiri, I.N., Reidpath, D.D., Sarran, C., 2012. Forecasting peak asthma admissions in London: an application of quantile regression models. Int. J. Biometeorol. 57 (4), 569–578. http://dx.doi.org/10.1007/s00484-012-0584-0.

Taieb, S.B., Huser, R., Hyndman, R.J., Genton, M.G., 2016. Forecasting uncertainty in electricity smart meter data by boosting additive quantile regression. IEEE Trans. Smart Grid PP 99, 1–8. http://dx.doi.org/10.1109/TSG.2016.2527820.

Thompson, M.L., Reynolds, J., Cox, L.H., Guttorp, P., Sampson, P.D., 2001. A review of statistical methods for the meteorological adjustment of tropospheric ozone. Atmos. Environ. 35 (3), 617–630. http://dx.doi.org/10.1016/S1352-2310(00)00261-2.

Yu, R., Yang, Y., Yang, L., Han, G., Move, O.A.. RAQ-A random forest approach for predicting air quality in urban sensing systems. Sensors (Basel, Switzerland) 16(1). http://dx.doi.org/10.3390/s16010086.

Vlachogianni, A., Kassomenos, P., Karppinen, A., Karakitsios, S., Kukkonen, J., 2011. Evaluation of a multiple regression model for the forecasting of the concentrations of NOx and PM10 in Athens and Helsinki. Sci. Total Environ. 409 (8), 1559–1571. http://dx.doi.org/10.1016/j.scitotenv.2010.12.040.

Wang, W., Lu, W., Wang, X., Leung, A.Y., 2003. Prediction of maximum daily ozone level using combined neural network and statistical characteristics. Environ. Int. 29 (5), 555–562. http://dx.doi.org/10.1016/S0160-4120(03)00013-8.

Yang, R., Zhao, N., Yan, F., 2016. A novel approach based on an improved random forest to forecasting the air quality of second-hand housing. In: 2016 9th International Symposium on Computational Intelligence and Design (ISCID), Vol. 1, pp. 274–277. http://dx.doi.org/10.1109/ISCID.2016.1069.

Yu, K., Lu, Z., Stander, J., 2003. Quantile regression: applications and current research areas. J. R. Stat. Soc. Ser. D (The Statistician) 52 (3), 331–350. http://dx.doi.org/10.1111/1467-9884.00363.

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, 632–655. http://dx.doi.org/10.1016/j.atmosenv.2012.06.031.

Zhang, Y., Wang, J., Wang, X., 2014. Review on probabilistic forecasting of wind power generation. Renewable and Sustainable Energy Reviews 32, 255–270. http://dx.doi.org/10.1016/j.rser.2014.01.033.