A comparison of probabilistic forecasting methods for extreme NO_2 pollution episodes

Sebastián Pérez Vasseur

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

José L. Aznarte[1]

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

Abstract

Keywords: probabilistic forecasting, air quality, quantile regression, nitrogen dioxide, Madrid

1. Introduction

2. Probabilistic forecasting with quantile regression

As mentioned above, the prediction from most regression models is a point 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 the conditional distribution. The technique is called quantile regression (QR) and was first proposed by Koenker in 1978 [?] based on the intuitions of the astronomer and polymath Rudjer Boscovich in the 18th century. Elaborating from the same concept of estimating conditional quantiles from different perspectives, several statistical and CI models that implement this technique have been developed: from the original linear proposal to multiple or additive regression, neural networks, support vector machines, random forests etc.

Quantile regression has gained an increasing attention from very different scientific disciplines [?], including financial and economic applications [?], medical applications [?], wind power forecasting [?], electric load forecasting [? ?], environmental modelling [?] and meteorological modelling [?] (these references are just examples and the list is not exhaustive). To our knowledge, despite its success in other areas, quantile regression has not been applied in the framework of air quality.

Thus, as we can estimate an arbitrary quantile and forecast its values, we can also estimate the full conditional distribution, which will entail us to the results presented in Section 4.

Also, probabilistic forecasting is an advantage as we need to predict when the target will be above a certain threshold (180). So instead of having a Yes/No Answer, we are calculating the probability of the target being the above the threshold.

Among the array of methods that allow to estimate and forecast data-driven conditional quantiles, in this study we have chosen k-neighbors quantile regression, quantile regression forests and quantile XGBoost. We will compare the different algorithms through the CRPS metric for the distribution and the RMSE, MAE, Correlation and Bias for the quantile 50.

3. Data description and experimental design

3.1. Training Data

In 20 pollution stations in Madrid, we have an almost complete hourly record of the last 5 years of:

- the levels of NO2 and O3
- calendar variables that represent the status of the day: bank holiday, laboral, type of holiday ...

Email address: `jlaznarte@dia.uned.es` (José L. Aznarte)

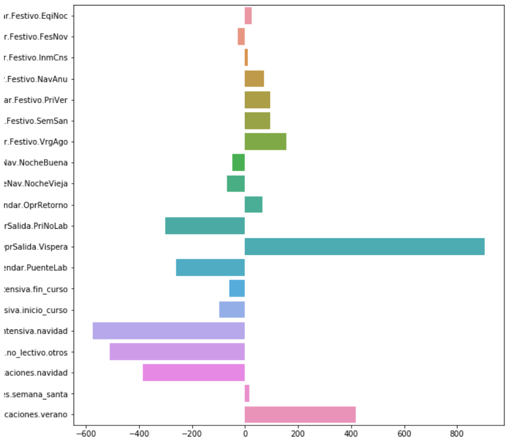
[1] This work has been partially funded by Ministerio de Economía y Competitividad, Gobierno de España, through a *Ramón y Cajal* grant (reference: RYC-2012-11984).

- weather data: Precipitation, Temperature, Humidity, Wind
- ECMWF numerical pollution prediction

3.2. Feature Preprocessing

3.2.1. Calendar Fields Summary

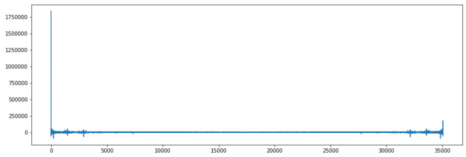
We will create 2 variables: one which indicates the day has a positive effect on pollution and another one for negative effect. We use a linear regression to predict the NO2 value only from the calendar variables. As we can see in the chart, some variables have a positive effect and others a negative effect, based on the sign of the coefficient applied to that feature:



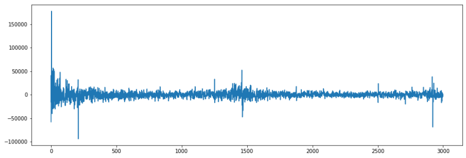
Then we create the 2 new variables as the sum of the positive features and the sum of the negative features respectively.

3.2.2. Seasonality Extraction

We extract the 6 main factors from the Fourier Transform of the signal. The chart displays the absolute value of the Fourier components of the time series:



If we zoom on the first 3000 components:



As we can see there are some dominant frequencies that show the time series has a strong seasonality. The 5 main frequencies are:

- Every 12 hour Seasonality
- Yearly Seasonality
- Daily Seasonality
- Every 4 year seasonality
- Weekly Seasonality

Therefore, we will create as input the output of periodic functions (cos) whose frequency is equal to the ones found above. This will enable the machine learning model learn the seasonality of our time series.

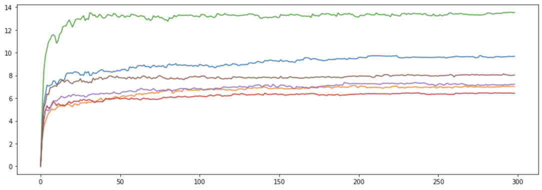
3.2.3. Previous Values

As with any forecast technique based on machine learning, we add previous values to improve the accuracy of the analysis. Based on the seasonal analysis, we see it's interesting to add a week of values. We will not add more to keep a reasonable number of features as input.

3.3. k-Neighbors

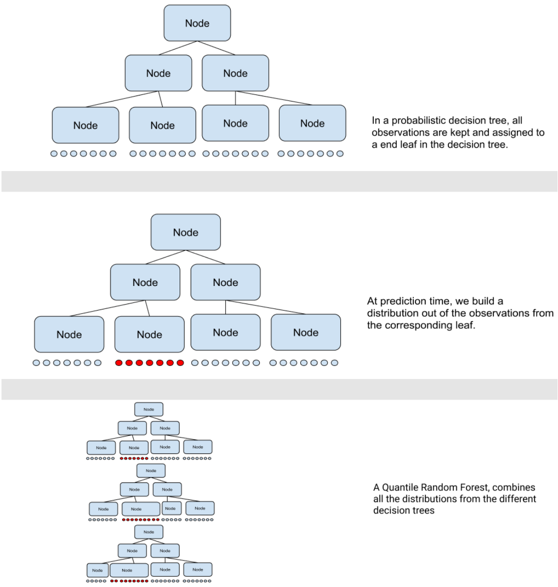
The probabilistic k-Neighbors is based on the competition entry from [K-nearest neighbors for GEFCom2014 probabilistic wind power forecasting]. This algorithm is based on the standard k-Neighbor, where instead of aggregating the targets of the k nearest points to the input (by taking the mean or median for example), it builds a distribution from those neighbors.

We need to find an optimal number k for our model. We will use the CRPS of the predicted distribution to get the best k. As you can see in the chart, 50 seems to be the optimal number of neighbor.



3.4. Quantile Random Forest

We will use the quantile random forest as described in [1]. Quantiles are built from the observations in the training set. We take the observation points that belong to the same leaf than the input and we build a distribution from those points. See figure for explanation:



In a probabilistic decision tree, all observations are kept and assigned to a end leaf in the decision tree.

At prediction time, we build a distribution out of the observations from the corresponding leaf.

A Quantile Random Forest, combines all the distributions from the different decision trees

For a detailed discussion on quantile regression forests, see [**?**].

3.5. Gradient Boosted Tree

Gradient Boosted Trees is a technique that consist on growing trees based on the compromise of a cost function and a regularization functions. This cost function is usually used to forecast the mean of the signal. But we are modifying the cost function in order to forecast the quantile of the function.

s an illustration of the concept (for a detailed discussion of quantile regression, refer to [**?**]), 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,

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

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,

$$E = \frac{1}{n}\sum_{i}^{n} \epsilon_i = \frac{1}{n}\sum_{i}^{n} |y_i - (\alpha_0 + \alpha_1 x)|. \tag{2}$$

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

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

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

By forecasting different quantiles, we can forecast the CDF of the time series. The main drawback is that we need to build a model per quantile. And since quantiles are calculated separately, we can have quantile crossing, i.e. the non monotonicity of the predicted CDF.

In order to solve that, we will apply the technique from . This chart shows the predicted 10 and 90 percentiles of the first 100 values.

3.6. Protocol for high NO_2 concentration episodes

3.7. Nitrogen dioxide data

3.8. Weather data

3.9. ECMWF numerical pollution prediction

3.10. Experimental design

We will train the models with data prior to 2017 and we will test our models with 2017 data. We will always test with predictions done at 10:00, as this is the time the forecast will be done and the alert will be decided or not.

Also, we will create different models for each horizon (how many hours in advance we do the forecast)

3.11. Evaluation of probabilistic forecasts

3.11.1. Variance of the Time series per hour

The time series have very different variance for each hour of the day. This is important when we evaluate our model, since we re always forecasting at 10:00.

3.11.2. Metrics

We will evaluate the 50 percentile through the RMSE, MAE, Bias and Corr and the whole forecasted CDF through the CRPS.

3.11.3. CRPS

The CRPS measures the accuracy of a probabilistic forecast.

3.12. Evaluation of alert forecasting

The alert forecasting is a classifier problem where we evaluate if the

3.12.1. Alert Protocol

The Madrid region is divided into zones and each zone contains pollution stations that track the levels of pollutants.

There are 3 levels of alert based on the levels of NO2:

- Prewarning (More than 180 $\frac{\mu gr}{m^3}$): 2 consecutive hours on 2 stations in the same zone OR 3 consecutive hours in 3 stations on any zone.

- Warning (More than 200 $\frac{\mu gr}{m^3}$): 2 consecutive hours on 2 stations in the same zone OR 3 consecutive hours in 3 stations on any zone.

- Alert (More than 400 $\frac{\mu gr}{m^3}$): 3 consecutive hours on 3 stations in the same zone OR 2 consecutive hours in zone 4.

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

3.12.2. Training data

We have few alerts in the last 5 years, so it is difficult to evaluate meaningfully the alert prediction.

4. Results and discussion

4.1. Reference models

In the first experiment, we used quantile regression to compute point-forecasts of the expected value (median) for one-day ahead predictions of NO_2 concentrations.

Table 1 shows the values of the root mean squared error (RMSE), mean average error (MAE), bias and correlation 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, as a matter of fact, interested in those outliers, as they precisely are the values which trigger the activation of the air quality protocol.

4.2. Probabilistic forecasting of extreme values

4.3. Forecasting the probability of alerts

5. Conclusions

References

[1] D. of Random Forest, quantregforest.
URL https://stat.ethz.ch/~nicolai/quantregforests.pdf