A comparison of probabilistic forecasting methods for extreme NO_2 pollution episodes

Sebastián Pérez

(afiliación y dirección)

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.

As 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.

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).

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.

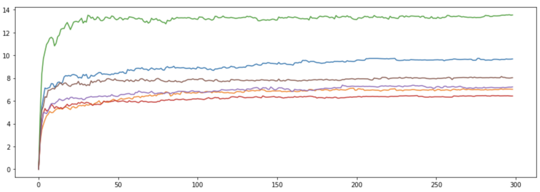
Among the array of methods that allow to estimate and forecast data-driven conditional quantiles, in this study we have chosen quantile regression forests for its ease of use (few parameters have to be chosen) and for its availability in the free software mathematical environment `R` [?]. For a detailed discussion on quantile regression forests, see [?].

3. Data description and experimental design

3.1. k-Neighbors

The probabilistic k-Neighbors is based on the competition entry from [K-nearest neighbors for GEFCom2014 probabilistic wind power forecasting]. We are using a k-neighbor algorithm, where instead of aggregating the targets for those k-neighbors (by taking the mean or the median), we are calculating the quantiles of those targets.

We then use the CRPS of our distribution estimation to get the best number of k-neighbor. As you can see in the chart, we chose 50 as the optimal number of neighbor.



3.2. Gradient Boosted Tree

We modify the cost function to calculate the . This means that we need to train a model for each of the percentiles we want to forecast. The other problem is that the percentiles can cross. We will use the method in order to correct that.

3.3. Quantile Random Forest

The quantile random forest is based on the usual random forest. The diferent trees create partitions out of the data and they select a partition in order to make the prediction. However, this time, the targets on the partition are not aggregated, but are used to create a distribution function.

3.4. Protocol for high NO_2 concentration episodes

3.5. Nitrogen dioxide data

3.6. Weather data

3.7. ECMWF numerical pollution prediction

3.8. Experimental design

3.9. Evaluation of probabilistic forecasts

3.10. Evaluation of alert forecasting

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: 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

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