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/

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 $\widehat{y}(x) = \alpha_0 + \alpha_1 x$ minimizes the mean squared error,

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

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 $\widehat{y}(x)$ which minimizes the mean absolute error,

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

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

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

where

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

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