Agricultural and Forest Meteorology 226–227 (2016) 96–107

Contents lists available at ScienceDirect

Agricultural and Forest Meteorology

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

Statistical modelling of non-stationary processes of atmospheric pollution from natural sources: example of birch pollen

Olga Ritenberga[a,*], Mikhail Sofiev[b], Victoria Kirillova[c], Laimdota Kalnina[a], Eugene Genikhovich[c]

[a] University of Latvia Faculty of Geography and Earth Sciences, Rainis bvld 19, Riga LV-1586, Latvia
[b] Finnish Meteorological Institute, Erik Palmenin aukio 1, 00560, Helsinki, Finland
[c] Voeikov Main Geophysical Observatory, Karbysheva street 7, St.Petersburg, 194021, Russia

ARTICLE INFO

Article history:
Received 16 November 2015
Received in revised form 19 May 2016
Accepted 22 May 2016

Keywords:
Statistical models
Non-stationary processes
Pollen forecasting
Betula spp

ABSTRACT

A statistical model for predicting daily mean pollen concentrations during the flowering season is constructed and its parameterization and application to birch pollen in Riga (Latvia) are discussed. The model involves several steps of transformations of both meteorological data and pollen observations, aiming at a normally distributed homogeneous stationary dataset with linearized dependencies between the transformed meteorological predictors and pollen concentrations. The data transformation includes normalization of daily mean birch pollen concentrations, a switch of the independent axis from time to heat sum, a projection of governing parameters to pollen concentrations, and a reduction of non-stationarity via removal of the mean pollen season curve. These transformations resulted in a substantial improvement of statistical features of the data and, consequently, a higher efficiency of statistical procedures and better scores of the model. The transformed datasets are used for the model construction via multi-linear regression. For the application in Riga, the model coefficients were calculated using 9 years of birch pollen observations. The model was evaluated using years withheld from the training dataset. The evaluation showed robust model performance with the overall Model Accuracy exceeding 80% and Odds Ratio = 30.

1. Introduction

Prevalence of sensitization to aeroallergens in Europe has continuously risen in recent decades (Ring et al., 2012) and presently exceeds 20% (Bauchau and Durham, 2004; Newson et al., 2014). The main aeroallergens in northern Europe are birch and grasses (D'Amato et al., 2007; Huynen et al., 2003) but hazel, alder, and mugwort, are also important (Akdis et al., 2014; D'Amato et al., 2007; Gadermaier et al., 2008). Adverse health effects of allergens can be significantly reduced by pre-emptive medication and behavioural adaptation. However, their planning requires reliable forecasts of expected pollen concentrations a few days ahead (Huynen et al., 2003).

Monitoring of atmospheric concentrations of pollen usually provides, with one-to-two weeks' delay, information about pollen in the air, which is used for pollen information and forecasting services in many European countries. Integrated pollen samplers allow collection of near-real-time data but they are expensive and so far not commonly used (Scheifinger et al., 2013).

Arguably, the most important parameter of the pollen season, from a practical point of view, is its starting date. It is followed by the season end date and the season-long sum of daily mean pollen concentrations – the seasonal pollen index (SPI). Determination of the start/end dates is not straightforward and several criteria were formulated (Jato et al., 2006). For instance, the season start (end) can be defined as a date when cumulative daily mean pollen concentration reaches 5% (95%) of the SPI (Taylor and Andersen, 2009). Specific numbers vary between the studies (Andersen, 1991; Emberlin et al., 2007; Pathirane, 1975; Smith et al., 2009; Stach et al., 2008b). However, this approach does not allow determination of the season start date until it already ends and the SPI is known, which makes it unsuitable for forecasting and real-time assessment purposes.

Two types of forecasting models are the most popular: regional-to-continental dispersion models and local-scale statistical models. Dispersion models (Helbig et al., 2004; Prank et al., 2013; Sofiev et al., 2015, 2012, 2006; Zink et al., 2013, 2012) are capable of predicting the pollen distribution over large areas but their accuracy strongly varies in space and depends on available information on plant distribution (Siljamo et al., 2012; Sofiev et al., 2015).

* Corresponding author.
E-mail addresses: olga.ritenberga@lu.lv (O. Ritenberga), mikhail.sofiev@fmi.fi (M. Sofiev), ego@main.mgo.rssi.ru (E. Genikhovich).

http://dx.doi.org/10.1016/j.agrformet.2016.05.016
0168-1923/

The local-scale statistical models exploit empirically established relations between the predicted quantity (predictant, such as pollen concentration) and independent predictors (meteorological factors and historical pollen concentrations) (Rodriguez-Rajo 2000, as referred by Castellano-Méndez et al., 2005). The ways of establishing these relations vary widely and include: (i) *artificial neural networks* (Castellano-Méndez et al., 2005; Puc, 2012); (ii) *discriminant linear analysis* (Sánchez Mesa et al., 2005); (iii) *multiple regression analysis* (Inatsu et al., 2014); (iv) *autoregressive integrated moving average* (Rodríguez-Rajo et al., 2006; García-Mozo et al., 2014); (v) *Gamma, Gaussian*, or *logistic distribution models* (Kasprzyk and Walanus, 2014).

Such statistical models commonly predict the start of season (Emberlin et al., 2002; Frei and Gassner, 2008; Laaidi, 2001; Laaidi et al., 2003; Siniscalco et al., 2014), its peak and duration (Ribeiro et al., 2007), 10-day mean concentrations (Makra et al., 2011), etc. The "classical" task of forecasting the daily/hourly pollen concentrations a few days ahead is less common (Chapter 7 of Sofiev and Bergman, 2013).

A common methodological difficulty of statistical models is stringent requirements to features of the analysed data: (i) the least-square-error and correlation quality criteria are justified only for normally distributed stochastic processes; (ii) averaging- and correlation- based methods require the processes to be stationary and ergodic, so that the averaging over a statistical ensemble of realizations can be substituted with averaging over time; (iii) (multi-) linear regressions imply near-linear dependencies of the predicted quantity (predictant) and independent predictors. None of these assumptions is fulfilled in case of pollen modelling. Indeed, the mere existence of the start and the end of the season makes the process both non-stationary and non-ergodic. The distribution function of daily mean pollen concentrations is closer to a log-normal than to a normal distribution (Limpert et al., 2008; Toro et al., 1998). And relations between the main controlling parameters (temperature, humidity, precipitation, etc) and daily mean pollen concentrations are by no means linear. Finally, models are usually developed with observed meteorological data as an input, whereas their forecasting applications have to use weather model predictions. Consequences of such substitution are rarely analysed despite known limited accuracy of the meteorological models.

The most common method for data transformation applied in the literature is the log-transform of pollen concentrations (Masaka, 2001; Méndez et al., 2005) as a precaution against log-normally distributed data. A similar effect, albeit with thinner theoretical ground, is sometimes obtained via square root function (Toro et al., 1998) or employing the SPI as shown by Moseholm et al. (1987). Other difficulties have not been considered in the studies we are aware of.

The current study aims to construct a local statistical forecasting model that takes the above peculiarities of the pollen time series into account. The objective is *to predict the daily mean pollen concentrations using only basic meteorological parameters*. For this purpose, we shall modify the methodology developed at Voeikov Main Geophysical Observatory for urban air pollution (Berlyand, 1991; Genikhovich et al., 2004). The model will be applicable to any monitoring site location and any taxa, whose flowering is controlled by accumulated heat. For illustrative purposes, we shall use birch pollen observations obtained in Riga (Latvia).

2. Materials and methods

2.1. Study area

Pollen monitoring was conducted in the central part of Riga (N56°57′02″, E24°06′57″, Fig. 1). Due to its location next to the Gulf of Riga (Baltic Sea), the city has temperate (humid continental) climate with frequent rain. Mean annual temperature of air in Riga is 6.9 °C and annual precipitation is 708 mm. The monitoring site is located in the centre of the city and surrounded by parks. The coastline is 12 km NW of the site.

2.2. Pollen sampling

Birch pollen monitoring was performed with a *Burkard* 7-day pollen spore trap of the Hirst design (Hirst, 1954) from March to September during the period 2003–2014. The sampler was situated at a height of 23 m agl. Pollen was collected with an airflow rate of 10 l min^{-1}, airflow rate controlled with an external flow meter *G 1.6. BK Premagas*. Pollen counting was done by using Primo Star light microscope with a magnification of × 400 over 12 full vertical traverses. The method with 12 vertical traverses produces comparable results to other commonly used counting methods, such as 4 horizontal traverses (Carinanos et al., 2000), and can produce both daily average and bi-hourly values. It also examines the whole traverses of the slide rather than the central parts where most of the pollen is deposited, thereby avoiding overestimation (Carinanos et al., 2000; Käpylä and Penttinen, 1981). However, it can, in theory, miss short peaks in pollen concentrations if they fall between the counted transects (Carinanos et al., 2000; Comtois et al., 1999)

Birch (*Betula spp*) pollen in Latvia comes from more than 31 species, however widely distributed are only four of them – *B. nana L., B. pendula Roth, B.humilis Schrank and B.pubescens Ehrh*. Birch tree contribution within the forests near Riga is about 14.6%, it is the second largest taxon after Scots pine tree (73.3%). The major forests are located about 40 km NE of Riga (Fig. 1).

Due to similar characteristics, birch pollen from different species are not distinguishable by light microscopy. Therefore, all pollen grains were counted jointly as one general birch group *Betula spp*.

From the 12-year-long data set, we have randomly picked 9 years for the model construction and withheld 2009 (typical-to-low pollen season), 2012, and 2014 (both high pollen seasons) for its evaluation (Figs. 2 and 3 left-hand panel)

2.3. Meteorological information

Meteorological data for the years 2006–2014 included daily mean and maximum values of air temperature, relative humidity, wind speed and direction, atmospheric pressure a.s.l., cloud fraction, visibility, and daily sum of precipitation. They were extracted from the meteorological station "Riga- LU", as provided by Latvian Environment, Geology and Meteorology Centre. Meteorological observations were divided into two parts:

(i) data on wind speed and direction, visibility, precipitation and cloud fraction came from the same place where the aerobiological monitoring was performed;
(ii) data on air temperature, relative humidity of air, atmospheric pressure came from the second part of the monitoring station located at about 1 km distance but also in downtown Riga.

The meteorological parameters for the early years 2003–2005 came from another meteorological station in Riga, located about 9 km from the aerobiological monitoring site.

For modelling purposes, wind speed and direction were recalculated to longitudinal wind component U and latitudinal one V.

2.4. Sensitivity study and comparison with other approaches

To mimic the application of the developed model in the forecasting regime, a sensitivity study has been performed: the meteorological observations were replaced with the forecasted