Aerobiologia
DOI 10.1007/s10453-012-9271-1

ORIGINAL PAPER

Analysis and forecasting of airborne pollen–induced symptoms with the aid of computational intelligence methods

D. Voukantsis · K. Karatzas · S. Jaeger · U. Berger · M. Smith

Received: 10 February 2012 / Accepted: 10 July 2012
© Springer Science+Business Media B.V. 2012

Abstract Allergies due to airborne pollen affect a considerable percentage of Europeans; thus, the provision of health-related information services concerning pollen-induced symptoms can improve the overall quality of life. In this paper, we demonstrate the development of personalized, health-related, quality-of-life information services by adopting a data-driven approach. The data we use consist of allergic symptoms reported by people as well as detailed pollen count information of the most allergenic taxa. We apply computational intelligence methods in order to analyze symptoms, identify possible interrelationships with several pollen taxa and develop models that associate pollen count levels with allergic symptoms on a personal level. The results for the case of Austria show that this approach can lead to accurate personalized symptom forecasting models; we report an average correlation coefficient of $r = 0.70$ for a sample of 102 users of the Patients Hayfever Diary. We conclude that some of these models could serve as the basis for personalized health information services.

Keywords Allergy · Computational intelligence · Symptoms forecasts · Personalized health services · Patients Hayfever Diary

1 Introduction

Allergic rhinitis (hay fever) and asthma are closely related allergic diseases that are associated with exposure to aeroallergens such as pollen (Demoly and Bousquet 2006). It is estimated that approximately 25 % of Europeans suffer from respiratory allergy diseases (Annesi-Maesano 1999; Beasley et al. 1998), while there are reports that 15–25 % of the same population are suffering from pollen-related allergies (Huynen et al. 2003; de Monchy et al. 2004; Bousquet et al. 2004), although pollen-related symptoms may vary, depending on local environmental factors (Bousquet et al. 2007). During recent years, there has been an increasing trend in pollen-induced allergic symptoms (WHO 2006), while the World Allergy Organization (www.worldallergy.org) reports an

D. Voukantsis (✉) · K. Karatzas
Informatics Systems and Applications Group, Department of Mechanical Engineering, Aristotle University, P.O. Box 483, 54124 Thessaloniki, Greece
e-mail: voukas@isag.meng.auth.gr

K. Karatzas
e-mail: kkara@eng.auth.gr

S. Jaeger · U. Berger · M. Smith
Medizinische Universität Wien, Universitätsklinik für Hals, Nasen und Ohrenkrankheiten, Währinger Gürtel 18-20, 1090 Vienna, Austria
e-mail: siegfried.jaeger@meduniwien.ac.at

U. Berger
e-mail: uwe.berger@meduniwien.ac.at

M. Smith
e-mail: matthew.smith@meduniwien.ac.at

Published online: 28 July 2012

increase in severity of the symptoms. In this context, it is important to assist sensitized people (i.e., those who have developed and demonstrated allergic reactions) in making decisions concerning their daily habits.

Allergy significantly reduces quality of life and has a nonnegligible economic impact on society, either directly through treatments or indirectly through diminishing performance at work and loss of productive hours (Malone et al. 1997; Szeinbach et al. 2007). Therefore, it would be beneficial for sensitized people to receive advice that would help them manage their symptoms. Avoiding exposure to aeroallergens is extremely important for dealing with IgE-associated allergic diseases (Peden and Reed 2010), and knowledge of airborne concentrations of aeroallergens helps allergy sufferers and health care professionals to plan effective treatment (Emberlin et al. 1999). This issue has already been partly addressed by existing information systems (www.polleninfo.org/) that provide information of pollen concentration levels (an equivalent of pollen counts) for allergenic taxa. The latter piece of information is usually disseminated at a simplified version of a nominal scale (low, medium, high) that can be easier interpreted by people. However, the severity of allergic symptoms is strongly dependent on the person under consideration (living conditions, mobility patterns, etc.), as well as the pollen taxon, since the allergenicity depends on the pollen concentration levels (Traidl-Hoffmann et al. 2009). Therefore, a more personalized approach that addresses symptoms rather than pollen concentration levels seems appropriate. The latter approach is supported by the absence of a regulatory or management framework, based on common limit values and threshold levels to serve as a reference for any health-related decision.

The design and development of personalized information services, addressing airborne pollen–related allergy symptoms, is a challenging task. This is mainly due to the lack of a generic (universal) mechanism responsible for the triggering of allergic symptoms. In order to avoid such difficulties, we adopt a human-in-the-loop approach, and utilize the database of pollen-induced symptoms reported by people themselves through the PHD (Patients Hayfever Diary) (Berger et al. 2011). The PHD system has been developed in cooperation with the Austrian Pollen Information Service, the Stiftung Deutscher Polleninformationsdienst, the Réseau National de Surveillance Aérobiologique in France, the ORL Department at the Medical University of Vienna and the Allergy-Centre-Charité in Berlin (www.pollendiary.com/Phd/ and www.pollenstiftung.de), in order to allow for the structured and continuous reporting of pollen-induced symptoms and relevant medication use by allergy sufferers throughout Europe. PHD users are reporting pollen-induced symptoms and medication use on a daily basis, through the web interface of the PHD system and with the aid of a relevant questionnaire. It should be noted that the PHD provides the opportunity to record personal symptoms, view pollen levels from areas of interest and develop reports on personalized symptoms and pollen with the aid of graphs, while other diaries can only be linked with pollen data *ex post* and thereby ignore the patient's actual whereabouts. This was the main reason for not using an existing symptoms reporting questionnaire, but developing a new one in the frame of PHD, which is based on existing and validated questionnaires and follows EMEA guidelines (www.emea.europa.eu). Additionally, we use detailed pollen concentration levels from EAN (European Aeroallergen Network, https://ean.polleninfo.eu/Ean/), monitored in several sampling sites, for the area of interest. We demonstrate the development of data-driven models for the case of Austria, using data collected during the year 2010. We argue that the resulting personalized models could serve as the main modeling tool of existing or future information systems, providing user-specific warnings concerning the occurrence and severity of airborne pollen–induced symptoms.

2 Materials and methods

2.1 Data presentation

The computational approach adopted in this study has been based on two distinct databases:

2.1.1 Pollen concentration data

Daily average pollen data for 2010 were collected at more than 20 sites in Austria using volumetric spore traps of the Hirst design (Hirst 1952) situated at roof level and stored in the EAN database held in Vienna. Air flows into the trap at a rate of 10 l/min through a 2 × 14-mm orifice, over a rotating drum (or microscope slide) that moves past the inlet at 2 mm/h. The

drum is covered with an adhesive coated, transparent plastic tape. Particles in the air impact on the tape to give a time-related sample (Emberlin 2000). Following its removal from the trap, the tape is divided into segments corresponding to 24-h periods (48 mm in length) (Piotrowska and Weryszko-Chmielewska 2006). Each segment is mounted between a glass slide and cover slip using a gelvatol- and glycerol-based mixture, dyed with basic fuchsin. The samples are then examined by light microscopy, and pollen grains are generally identified at $\times 400$ magnification. Slides are examined along five longitudinal transects following the method described in the literature (Mandrioli 1995; Mandrioli et al. 1998; Jäger 2000; Galán et al. 2007). Finally, raw data are converted to airborne pollen concentrations per cubic meter of air.

The data used in our study were pre-processed excluding taxa with more than 10 % missing values, finally resulting in 20 taxa (Table 1). These variables were considered as possible predictors for the models, and each one was evaluated on a user-specific base, applying the criteria presented in Sect. 2.3.2.

2.1.2 Pollen-induced symptoms

The PHD database includes a total of 716 users, contributing on an average of approximately 57 records during the year 2010, thus summing to a total of 40,700 distinct data records. The total number of points related to the symptoms that a person can report may vary from zero (no symptoms) up to twenty-six. The symptoms were reported by patients, and the total number of points was calculated as follows: Three body areas of interest are identified: lungs, nose and eyes. For each body area, a discomfort score ranging from zero to three points is used, leading to a total of $3 \times 3 = 9$ points (at maximum). Moreover, four specific symptoms are identified (such as eye itching or watering), for each one of the three body areas, with each one receiving one point, thus leading to a maximum score of four points per body area and an overall maximum score of $3 \times 4 = 12$ points. In this study, up to five additional points could be added as a result of the medication being used (one point per medication type: eye drops, nose drops, tablets, homeopathic, other), leading to an overall maximum number of points that a patient can report equal to twenty-six. It should be noted, however, that this is a first step in symptom score calculations, and different approaches have been considered, such as different weighting given to the different medications used, or a combined (and different) weighting of medications and symptoms. The variables associated with pollen-induced symptoms are presented in Table 1. The names of the variables consist of two parts: the first one specifying the body area of interest and the second specifying the symptom.

Finally, we normalized the data using the variance scaling. The latter step was necessary in order to apply certain computational intelligence (CI) algorithms during the modeling process, such as the artificial neural networks. The normalized variables are characterized by an average of $\mu = 0$ and a standard deviation of $\sigma = 1$. Figure 1 provides an aggregated visualization of the databases involved in this study.

Table 1 Structure of the dataset used to develop the models[a]

Pollen concentration (in grains/m^3)	Symptoms (according to PHD system)
Acer	Lungs—asthma
Alnus	Lungs—cough
Ambrosia	Lungs—shortness of breath
Betula	Lungs—wheezing
Carpinus	
Corylus	Nose—blocked
Cupressaceae	Nose—running
Fraxinus	Nose—sneezing
Juglans	Nose—itching
Pinus	
Plantago	Eye—watering
Platanus	Eye—redness
Poaceae	Eye—foreign body
Populus	Eye—itching
Quercus	
Rumex	Overall symptoms (total points)
Salix	
Sambucus	
Tilia	
Urtica	

[a] The left column summarizes the pollen concentration data provided by the European Aeroallergen Network (only taxa with <10 % missing values were considered). The right column summarizes variables related to symptoms experienced by the users of the PHD system. Symptoms as well as symptom details are reported for lungs, nose and eyes

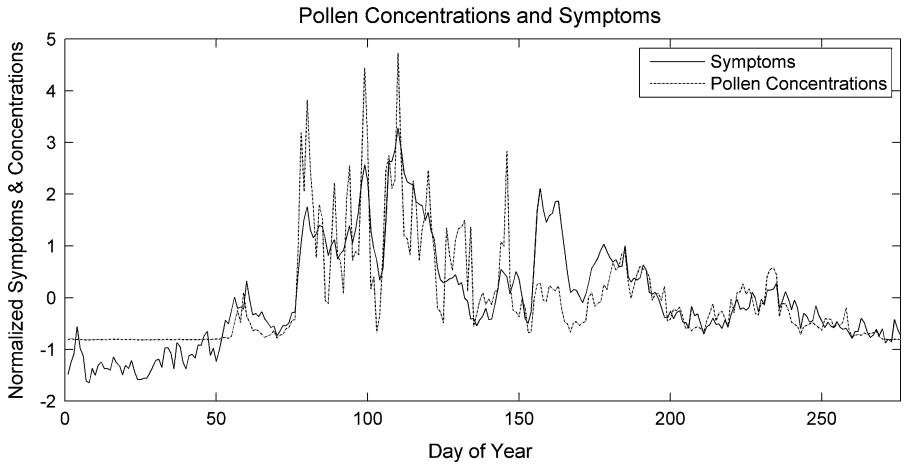


Fig. 1 Overall symptoms (average of all users, *solid line*) and pollen concentration (average of all taxa, *dash-dot line*) during the year 2010 in Austria. Both variables have been normalized, using the variance scaling ($\mu = 0$, $\sigma = 1$)

2.2 Clustering

Clustering is an unsupervised learning technique capable of identifying structures in large datasets and representing the results in the form of data groups of similar characteristics. Similarity is measured in terms of certain metrics, such as the Euclidean distance in *N*-dimensional space (where *N* is the number of variables in the dataset). Therefore, the data groups (clusters) express specific associations among the variables of a dataset—in this case, pollen concentrations and allergic symptoms—and could represent interesting patterns that require further analysis and modeling (hypothesis generation).

In this case, clustering was the first step of our study, with the aim to reveal and investigate relationships between symptoms and pollen taxa. We have applied a two-step clustering process, combining the self-organizing map (SOM) with the *K*-means clustering.

The *self-organizing map* is an unsupervised neural learning algorithm (Kohonen 1997). The goal of the SOM algorithm is to find prototype vectors that represent the structure of the input dataset and at the same time realize a continuous mapping from the input space to a low-dimensional lattice. SOM is an algorithm characterized by robustness and computational efficiency, thus being an appropriate tool to apply to the large amount of data available in this application. Furthermore, it is often combined with other clustering algorithms, serving as an intermediate step of the clustering process. The advantage of this two-step approach has been already outlined in Vesanto and Alhoniemi (2000). In this case, we have combined SOM with *K*-means clustering.

K-means is a clustering algorithm (MacQueen 1967) that has been applied in many cases in analyzing environmental data (Kolehmainen et al. 2000; Voukantsis et al. 2009). The *K*-means algorithm consists of two steps: (a) assigning data records to clusters on the basis of a similarity criterion, for example, Euclidean distance, and (b) repositioning the cluster centers (centroids) by calculating the average of the new cluster. The application of *K*-means requires that the number of clusters is known in advance. In this case, we have evaluated several choices and identified the optimum number of cluster by using the Davies–Boulding Index (Davies and Bouldin 1979).

2.3 Estimating symptoms

In this study, we identified the parameter "overall symptoms" as the main parameter of interest concerning the development of personalized symptoms models. For this purpose, we develop computational models that can estimate the overall symptoms score based on pollen count information. These models can be applied on an operational basis, providing personalized information concerning pollen-induced symptoms.

2.3.1 Computational intelligence methods

In order to model the interrelations between pollen concentrations and user-specific symptoms, we have applied a series of CI methods. From the modelers' point of view, the latter task can be approached as a

regression problem, where the target variable ("overall symptoms") is estimated as a function of exposure to various pollen taxa and their concentrations. The following CI methods were evaluated for the task at hand:

Multi-layer perceptron (MLP) (Haykin 1994) models are artificial neural network models capable of modeling complex and highly nonlinear processes. In this case, we have evaluated network architectures consisting of one hidden layer, setting the number of neurons equal to $N = (inputs + outputs)/2$. Where *inputs* is the number of input variables of the model and *outputs* the number of output variables; in this case, outputs = 1. In addition, we have used the Levenberg–Marquardt backpropagation algorithm during the training of the models, with a maximum number of epochs at 500. The MLP models were developed with the aid of the ANN Toolbox for Matlab.

Support vector regression (SVR) is a modification of the original algorithm introduced by Vapnik (1995) capable of tackling regression problems. In this case, we have chosen the SVR kernel to be the radial basis function (RBF). We have also optimized the regularization parameter (C) and the ε-intensive zone of the model, as well as the γ parameter of the kernel. The SVR models have been developed using the Spider Toolbox for Matlab.

Least squares support vector regression (LS-SVR) is a modification of support vector machines (Suykens et al. 2002). In this case, we have used the RBF kernel and optimized the regularization (C) and kernel parameter (γ). The LS-SVR models were developed using the LS-SVMlab toolbox for Matlab.

k-nearest neighbors (kNN) is nonparametric CI algorithm that can be used to tackle regression problems. In this case, we have set the number of nearest neighbors to be $k = 5$, and we applied the $1/r$ (r is the distance to the neighbor) weighting scheme.

Multiple linear regression (MLR) is a well-known statistical method that was used as the reference for comparison of the performance of models based the aforementioned CI methods.

2.3.2 Feature selection

The selection of the appropriate features as input parameters of the models is an important modeling step as irrelevant or noisy features may result in poor model performance. Mutual information (MI) as well as other information-related criteria has been used in the past in order to identify relevant input variables. In this case, we have used MI following the approach presented in Peng et al. (2005).

2.3.3 Training and evaluation

Due to the limited number of data records per user, the training of the models was based on the tenfold cross-validation, in order to utilize all training examples. The root-mean-square error (RMSE), the correlation coefficient (r) and the index of agreement (d) have been used to validate the models. The latter index is defined by the following formula (Willmott 1982):

$$d = 1 - \frac{\sum_i |p_i - a_i|^2}{\sum_i \left(|p_i - \bar{a}| + |a_i - \bar{a}|\right)^2} \quad (1)$$

where p_i refers to predicted values and a_i to observed ones, whereas $\bar{p}$ and $\bar{a}$ are denoted as the average of the predicted and observed values, respectively. The index of agreement ranges from 0.0 (theoretical minimum) to 1.0 (perfect agreement between observed and predicted values).

3 Results and discussion

The results presented are related to the clustering of the symptoms and to the estimation of the latter on a personalized level.

3.1 Clustering

The SOM and K-means clustering approach resulted in 14 clusters describing the structure of the joined dataset (pollen concentrations and allergic symptoms). Each cluster corresponds to a particular fraction of the overall data that are characterized by a great deal of similarity. Table 2 summarizes the pollen type and symptoms characteristics of each cluster, as well as the temporal and spatial stamp of the data assigned to each cluster.

Focusing on the symptoms described in each cluster, as well as the possible source, that is, pollen concentration, we identify three groups of clusters, which we describe below:

Table 2 Main characteristics of the 14 clusters summarizing the joined dataset

Cluster	Percentage of data in the cluster (%)	Day of year (avg ± SD)	Overall symptoms (avg ± SD)	Pollen type related to cluster (max concentration used)	Geographic region
C_1	5.74	136 ± 44	12.71 ± 3.10	–	All
C_2	8.26	101 ± 8	5.86 ± 4.02	*Acer, Betula, Carpinus, Fraxinus, Salix*	All
C_3	2.25	77 ± 9	4.64 ± 3.95	*Alnus, Corylus*	All
C_4	4.95	135 ± 9	2.08 ± 2.42	*Pinus, Rumex, Urtica*	All
C_5	3.70	172 ± 51	4.77 ± 3.11	–	All
C_6	8.77	182 ± 41	3.94 ± 1.76	–	All
C_7	12.74	75 ± 30	1.66 ± 2.08	–	All
C_8	1.50	82 ± 4	4.93 ± 3.88	*Curpessaceae, Populus*	East
C_9	10.59	162 ± 51	7.13 ± 2.43	–	All
C_{10}	7.93	170 ± 13	4.45 ± 3.55	*Plantago, Poaceae, Sambucus, Tilia*	All
C_{11}	1.26	117 ± 4	11.05 ± 2.89	*Quercus, Juglans, Platanus*	All
C_{12}	3.71	118 ± 4	3.30 ± 2.61	*Quercus, Juglans, Platanus*	All
C_{13}	4.69	244 ± 15	2.40 ± 3.13	*Ambrosia*	East
C_{14}	23.91	183 ± 42	0.55 ± 1.08	–	All

No or mild symptoms (C_7, C_{14}). These two clusters correspond to 36.6 % of the overall data records and describe the part of data that users report as not having symptoms or very mild ones. The temporal stamp of these clusters is diffuse (C7: winter and early spring, C14: spring and summer); that is, it corresponds to a large standard deviation of the variable "Day of Year." The pollen concentrations are low (on average), and there are no peak values of any pollen taxon associated with these clusters. Furthermore, patients assigned to these clusters follow the overall spatial distribution of data.

Pollen-induced symptoms (C_2, C_3, C_4, C_8, C_{10}, C_{11}, C_{12} and C_{13}). These clusters sum to 38.8 % of the overall data. The severity of the symptoms reported in these clusters ranges from mild to strong (C_{11}). Each of the clusters is associated with peak values of one or more pollen taxa. All of these clusters have a clear time stamp (period of 10–20 days), whereas clusters C_8 and C_{13} have additionally a clear spatial stamp, localized in the eastern part of the country.

Other symptoms (C_1, C_5, C_6 and C_9). These clusters correspond to 24.6 % of the overall data and indicate moderate to strong symptoms. The clusters have a diffuse time stamp, including data records from the whole flowering season. There is not a clear association with peak values of specific pollen taxa. The symptom patterns (Fig. 2) suggest that these clusters describe users with multiple allergies (C_1) or asthmatic users (C_5), due to distinct symptoms ("Lungs Asthma").

The distinct temporal and spatial patterns identified among different clusters are further clarified in Fig. 3. Cluster 2 (C_2, Day of Year: 101 ± 8), consists of data records closely scattered around the average "Day of Year" (Fig. 3a), whereas for C_{14} the distribution is much more diffuse spreading throughout the spring–summer period (Fig. 3b). The first pattern is characteristic of pollen-induced symptoms. Furthermore, in 2 cases (C_8 and C_{13}), there is a clear spatial pattern. Cluster 8 is presented in Fig. 3c. In contrast, Cluster 5 (Fig. 3d) includes data records following the spatial distribution of the database, that is, all regions of the country.

Moreover, it is worth noting that the clusters identified may help in the definition of threshold values per pollen taxon, on the basis of average scores. Thus, for example, C_{13} (identified to be related to *Ambrosia*) can lead to the linkage of (1) overall symptoms score (2.40 ± 3.13) and (2) percentages of individual symptoms (mainly nose symptoms) to a suggestion for a threshold value for the specific pollen taxa and country.

3.2 Symptoms forecasting

We used the CI methods presented in Sect. 2.3.1 to develop user-specific models for estimating the

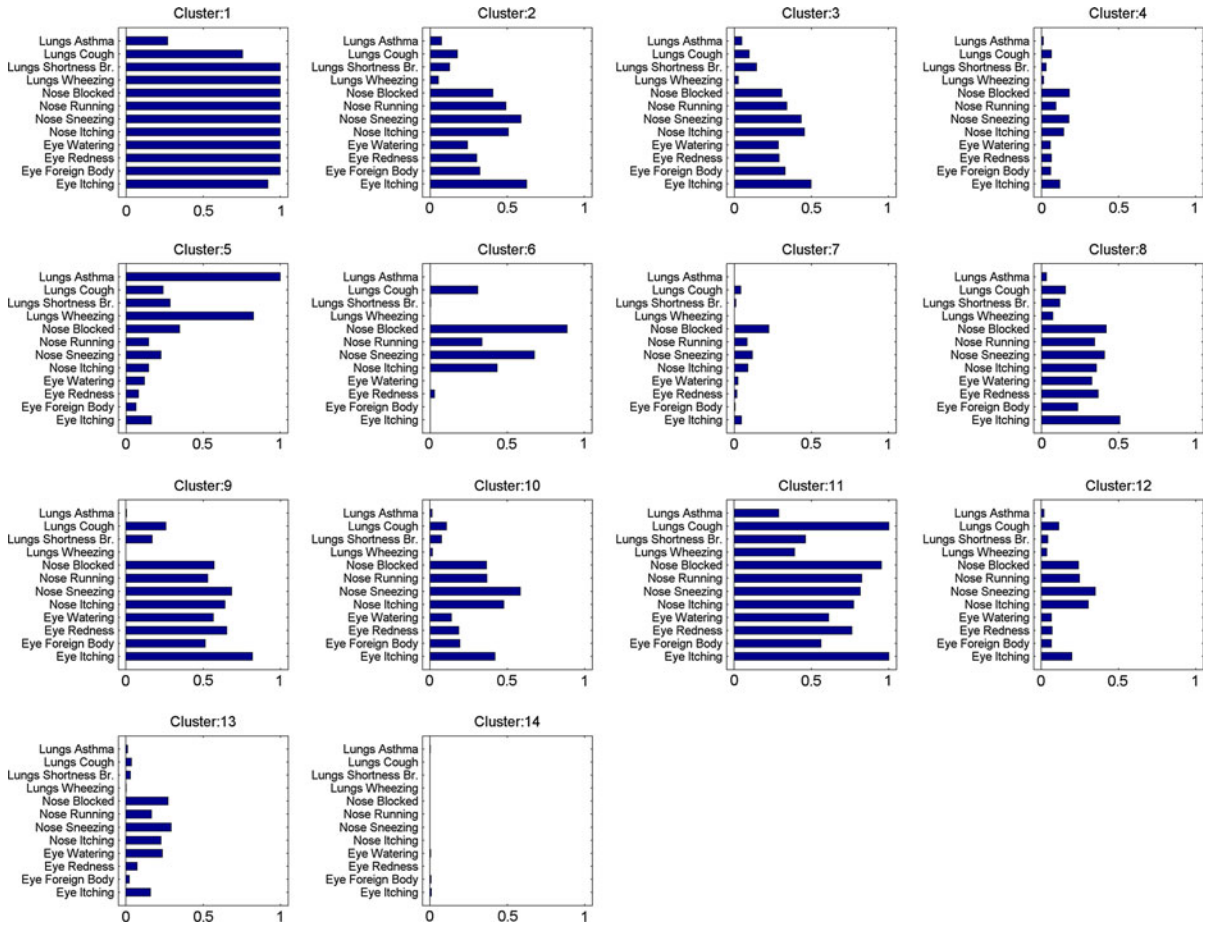


Fig. 2 Contribution of symptoms (*symptom patterns*) in each of the 14 clusters. The participation of specific symptoms is *graphed*, in a scale from zero to one (min: 0, max: 1). The symptoms follow the order presented in Table 1 (from *top* to *bottom*): *Lungs* {*Asthma, Cough, Shortness of Breath, Wheezing*}, *Nose* {*Blocked, Running, Sneezing, Itching*}, *Eye* {*Watering, Redness, Foreign Body, Itching*}

"overall symptoms" on the basis of pollen concentrations (Table 1). We focused on users with a large amount of data records; the threshold was set to at least 100 data records, thus resulting in a sample of 102 distinct users. The performance of the resulting models is summarized in Table 3. Results are presented as averages (and standard deviations) of the statistical indices used to evaluate model performance (Sect. 2.3.3).

As a first result, we identify the distinct performance of each model depending on the algorithm used to develop it. The nonparametric kNN models as well as the LS-SVR models indicate superior performance compared with the SVR and MLP models. This may be attributed to the limited number of training examples in the dataset (at least for the MLP models). Furthermore, all CI methods perform on average better than the reference (MLR) method. This indicates that modeling pollen-induced symptoms is a complex process and thus CI algorithms are more suitable.

The results presented in Table 3 indicate that modeling the process of pollen-induced symptoms on user level can result in models of acceptable performance and thus serve as the basis for operational information services. In contrast, an approach based on developing a unique model for all users was also tested, using the same data and algorithms. The performance of the resulting models ($d < 0.4$, $r < 0.25$, $rmse > 4$) suggests that such an approach is not successful. Furthermore, the results indicate that even within data assigned to the same cluster (Sect. 3.1), thus characterizing allergic symptoms to specific

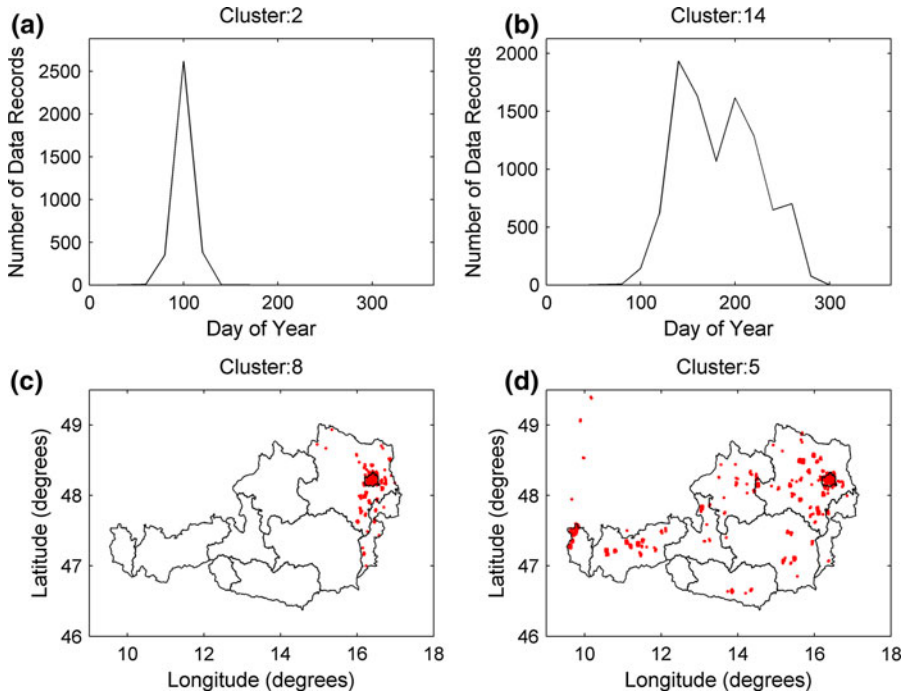


Fig. 3 Temporal and spatial distribution of data records for different clusters

Table 3 Performance (average and standard deviation) of 102 user-specific models

Model	*d*		*r*		RMSE	
	Avg	SD	Avg	SD	Avg	SD
MLR	0.74	0.14	0.61	0.20	2.26	1.05
MLP	0.76	0.15	0.63	0.22	2.17	0.93
kNN	0.80	0.11	0.67	0.17	2.07	0.86
SVR	0.74	0.14	0.64	0.19	2.11	0.88
LS-SVR	0.79	0.14	0.70	0.18	1.92	0.80

The unit of RMSE is the number of symptoms

taxa, satisfactorily performing models are not possible.

Etiopathogenesis of allergic diseases is multifactorial, since the immune system of each individual may react in different ways. Thus, the triggering of allergic symptoms in humans is a highly complex process that depends on several factors such as pollen concentrations, meteorological and chemical (i.e., air quality) weather conditions, people habits (outdoor activity, traveling, medication). Therefore, in some cases, the available data cannot produce accurate models. This can be demonstrated in more detail for the cases of two specific users (users with ID number 80 and 85, hereafter denoted as User#80 and User#85), with similar data records and maximum overall symptoms. The results of two user-specific models (Table 4) show that the allergic symptoms indicated by User#85 can be successfully modeled; however, the models were not as successful for User#80. Figure 4 presents the overall symptoms for both users, as a function of time. It is evident that the pattern of symptoms for both users is different and emphasizes the need for the production of user-specific models.

4 Conclusions and future work

In this paper, we elaborated and extended our previously published work (Voukantsis et al. 2011) to demonstrate the use of CI methods in order to analyze symptoms caused by pollen taxa of various types, and develop models capable of estimating pollen-induced symptoms on a personal level. The results demonstrate that there is no single or general mechanism linking pollen concentrations with symptoms. On the contrary, the personalization of symptomatology became more pronounced. On this basis, it is evident that the data collected by the PHD system can provide a solid platform that can help in identifying symptoms related to pollen taxa and geographic areas. The latter could lead to the identification of threshold values for

Table 4 Typical detailed performance of CI models for two users[a]

ID	Model	MLR	MLP	kNN	SVR	LS-SVR
User#80	*d*	0.67 ± 0.16	0.73 ± 0.16	0.74 ± 0.15	0.71 ± 0.13	0.79 ± 0.12
	r	0.54 ± 0.23	0.65 ± 0.26	0.59 ± 0.28	0.61 ± 0.19	0.68 ± 0.21
	rmse	4.51 ± 0.95	4.16 ± 1.33	4.44 ± 1.18	4.48 ± 1.08	3.83 ± 1.20
User#85	*d*	0.92 ± 0.09	0.95 ± 0.06	0.97 ± 0.02	0.94 ± 0.02	0.95 ± 0.03
	r	0.88 ± 0.15	0.93 ± 0.06	0.94 ± 0.04	0.92 ± 0.04	0.93 ± 0.05
	rmse	2.87 ± 0.90	2.30 ± 1.02	2.08 ± 0.78	2.70 ± 0.35	2.34 ± 0.67

[a] The results are presented as average and SD (tenfold cross-validation) of the statistical indices presented in Sect. 2.3.3. The unit of RMSE is the number of symptoms

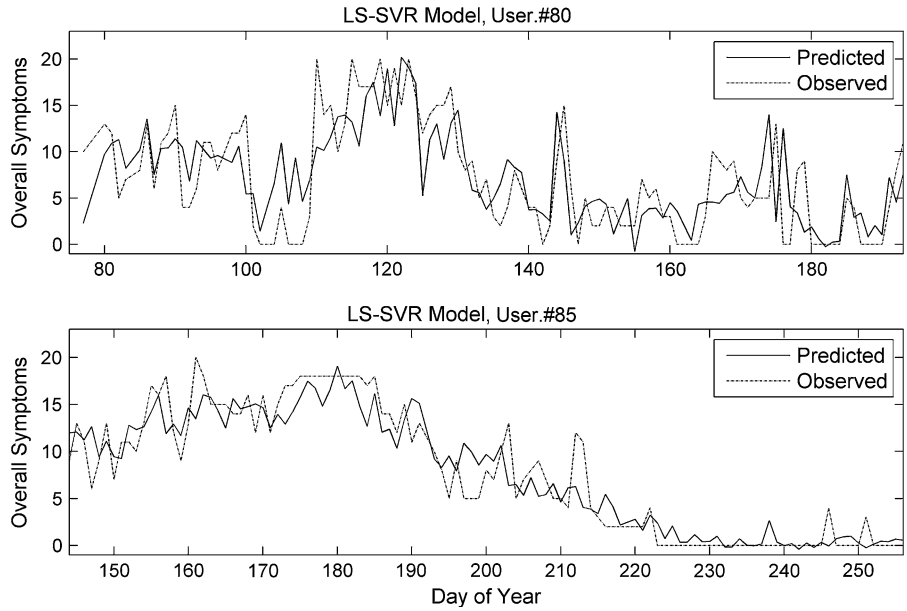


Fig. 4 Overall symptoms (*solid line*: predicted; *dash-dot line*: observed) for User#80 and User#85

regulatory and decision-making purposes. These data can also be used to build accurate user-specific models in most of the cases. The latter can serve as the basic modeling tool to develop personalized, health-related, quality-of-life information services.

Future work in this field includes the use of more data from the PHD system, involving more geographic locations and a longer time period. In addition, the utilization of the available meteorological databases, as well as other available databases, for example, air quality data, is expected to improve the performance of the existing models, by involving other environmental factors that affect the severity of pollen-induced symptoms. Furthermore, we expect that additional information can result in a more accurate categorization of users into several sensitized groups and identification of common characteristics. These goals will advance the accuracy and interpretability of the models and thus provide a more solid background for the development of personalized quality-of-life health services.

Acknowledgments The authors greatly acknowledge the contribution of all colleagues involved in the pollen concen tration data collection: Manfred Bobek, Sigmar Bortenschlager, Inez Bortenschlager, Uschi Brosch, Edith Bucher, Bernard Clot, Pramod Harvey, Veronika Kofler, Andreja Kofol-Seliger, Herta Koll, Sabine Kottik, Margit Langanger, Rudolf Litschauer, Karol Micieta, Anna Paldy, Ondrej Rybnicek, Hanna Schantl, Jutta Schmidt, Roland Schmidt, Ingrid Weichenmeier, Helmut Zwander. We would also like to greatly acknowledge all those PHD users contributing information via the PHD.

Ethical standard The authors of this paper state that they have taken into account the ethical standards laid down in the 1964 Declaration of Helsinki. All users of the PHD have given their consent prior to their inclusion in the PHD; in addition,

they had the option to give information anonymously. Details that might disclose the identity of the subjects participating in the study have been omitted. All data related to persons that have registered their symptoms and medication use to the PHD have been treated anonymously in this paper. On this basis, no ethical issues arise from this study and from the use of PHD data.

References

Annesi-Maesano, I. (1999). Epidemiological evidence of the occurrence of rhinitis and sinusitis in asthmatics. *Allergy, 54*, 7–13.

Beasley, R., Keil, U., von Mutius, E., Pearce, N., Ait-Khaled, N., Anabwani, G., Anderson, R., et al. (1998). Worldwide variation in prevalence of symptoms of asthma, allergic rhinoconjunctivitis, and atopic eczema: ISAAC. *Lancet, 351*, 1225–1232.

Berger, U., Jäger, S., & Bergmann, K. C. (2011). *PHD, the electronic patient's hayfever diary*. Amsterdam: ERS Congress.

Bousquet, J., Ansotegui, I. J., van Ree, R., Burney, P. G. J., Zuberbier, T., & van Cauwenberge, P. (2004). European Union meets the challenge of the growing importance of allergy and asthma in Europe. *Allergy, 59*, 1–4.

Bousquet, J., Chinn, S., Janson, C., Kogevinas, M., Burney, P., & Jarvis, D. (2007). Geographical variation in the prevalence of positive skin tests to environmental aeroallergens in the European Community Respiratory Health Survey I. *Allergy, 62*, 301–309.

Davies, D., & Bouldin, D. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 2*, 224–227.

de Monchy, J., Andersen, P. S., Bergmann, K. C., Chivato, T., Holm-Hansen, A., Jarisch, R., et al. (2004). Living and learning with allergy: A European perception study on respiratory allergic disorders. *Respiratory Medicine, 98*(5), 404–412.

Demoly, P., & Bousquet, J. (2006). The relation between asthma and allergic rhinitis. *Lancet, 368*, 711–713.

Emberlin, J. (2000). Aerobiology. In W. W. Busse & S. T. Holgate (Eds.), *Asthma and rhinitis* (Vol. 2, pp. 1083–1105). Oxford: Blackwell Science.

Emberlin, J., Mullins, J., Cordon, J., Jones, S., Millington, W., Brooke, M., et al. (1999). Regional variations in grass pollen seasons in the UK, long term trends and forecast models. *Clinical and Experimental Allergy, 29*, 347–356.

Galán, C., Cariñanos, P., Alcázar, P., & Dominguez-Vilches, E. (2007). *Spanish aerobiology network (REA) management and quality manual*. Cordoba: Servicio de Publicaciones Universidad de Córdoba.

Haykin, S. (1994). *Neural networks: A comprehensive foundation*. Upper Saddle River: Prentice Hall.

Hirst, J. M. (1952). An automatic volumetric spore trap. *Annals of Applied Biology, 39*(2), 257–265.

Huynen, M., Menne, B., Behrendt, H., Bertollini, R., Bonini, S., Brandao R., et al. (2003). Phenology and human health: Allergic disorders. In *Report of a WHO meeting*, Rome, Italy.

Jäger, S. (2000). Ragweed (Ambrosia) sensitisation rates correlate with the amount of inhaled airborne pollen. A 14-year study in Vienna, Austria. *Aerobiologia, 16*(1), 149–153.

Kohonen, T. (1997). *Self-organizing maps*. Berlin: Springer.

Kolehmainen, M., Martikainen, H., Hiltunen, T., & Ruuskanen, J. (2000). Forecasting air quality parameters using hybrid neural network modelling. *Environmental Monitoring and Assessment, 65*(1–2), 277–286.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. Le Cam & J. Neyman (Eds.), *The fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). California: University of California Press.

Malone, D. C., Lawson, K. A., Smith, D. H., Arrighi, H. M., & Battista, C. A. (1997). Cost of illness study of allergic rhinitis in the United States. *Journal of Allergy and Clinical Immunology, 99*, 22–27.

Mandrioli, P., (1995). Method for sampling and counting of airborne pollen and fungal spores. Institute of Atmospheric and Oceanic Sciences (ISAO). Accessed February 9, 2012 from http://www.polleninfo.org/upload/images/original/1995.pdf.

Mandrioli, P., Comtois, P., & Levizzani, V. (1998). *Methods in aerobiology*. Bologna: Pitagora Editrice.

Peden, D., & Reed, C. E. (2010). Environmental and occupational allergies. *Journal of Allergy and Clinical Immunology, 125*, 150–160.

Peng, H., Long, F., & Ding, C. (2005). Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(8), 1226–1238.

Piotrowska, K., & Weryszko-Chmielewska, E. (2006). Ambrosia pollen in the air of Lublin, Poland. *Aerobiologia, 2006*(22), 151–158.

Suykens, J. A. K., Van Gestel, T., De Brabanter, J., De Moor, B., & Vandewalle, J. (2002). *Least squares support vector machines*. Singapore: World Scientific.

Szeinbach, S. L., Seoane-Vazquez, E. C., Beyer, A., & Brock Williams, P. (2007). The impact of allergic rhinitis on work productivity. *Primary Care Respiratory Journal, 16*(2), 98–105.

Traidl-Hoffmann, C., Jakob, T., & Behrendt, H. (2009). Determinants of allergenicity. *Journal of Allergy and Clinical Immunology, 123*(3), 558–566.

Vapnik, V. (1995). *The nature of statistical learning theory*. New York: Springer.

Vesanto, J., & Alhoniemi, E. (2000). Clustering of the self-organizing map. *IEEE Transaction on Neural Networks, 11*(3), 586–600.

Voukantsis, D., Karatzas, K., Jaeger, S., & Berger, U. (2011). Personalized information services for quality of life: The case of airborne pollen induced symptoms. In L. S. Iliadis & C. Jayne (Eds.), *Proceedings of the 12 th international conference, engineering applications of neural networks 2011, and 7th IFIP WG 12.5 international conference, AIAI 2011*(pp. 509–515), Corfu, Greece, September 15–18, 2011, Part I.

Voukantsis, D., Karatzas, K., Rantio-Lehtimaki, A. & Sofiev, M. (2009). Investigation of relationships and interconnections between pollen and air quality data with the aid of computational intelligence methods. In *Proceedings of the 23 rd conference on environmental informatics and*

industrial environmental protection: Concepts, methods and tools (pp. 189–198). Berlin.

Willmott, C. J. (1982). Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society, 63*, 1309–1313.

World Health Organization (2006). Fact sheet No. 307, Asthma, Geneva. Accessed June 11, 2012 from http://www.who.int/mediacentre/factsheets/fs307/en/index.html.