

Analyzing the occurrence of an invasive aquatic fern in wetland using data-driven and multivariate techniques

Roghayeh Sadeghi · Rahmat Zarkami · Patrick Van Damme

Received: 1 February 2016 / Accepted: 2 February 2017 / Published online: 17 February 2017
© Springer Science+Business Media Dordrecht 2017

Abstract In the present study, the data-driven (classification trees and support vector machines) and multivariate techniques (principal component analysis and discriminant analysis) were applied to study the habitat preferences of an invasive aquatic fern (*Azolla filiculoides*) in the Selkeh Wildlife Refuge (a protected area in Anzali wetland, northern Iran). The applied database consisted of measurements from seven different sampling sites in the protected area over the study period 2007–2008. The cover percentage of the exotic fern was modelled based on various wetland characteristics. The predictive performances of the both data-driven methods were assessed based

on the percentage of Correctly Classified Instances and Cohen's kappa statistics. The results of the Paired Student's *t*-test ($p < 0.01$) showed that SVMs outperformed the CTs and thus yielded more reliable prediction than the CTs. All data mining and multivariate techniques showed that both physical-habitat and water quality variables (in particular some nutrients) might affect the habitat requirements of *A. filiculoides* in the wetland.

Keywords *Azolla filiculoides* (Lam.) · Data-driven methods · Invasive species · Management · Multivariate techniques · Predictive modelling

R. Sadeghi (✉) · P. Van Damme
Department of Plant Production, Faculty of Bio-Science Engineering, Ghent University, Coupure Links, 653, 9000 Ghent, Belgium
e-mail: roghayeh.sadeghipasvisheh@ugent.be;
Jila.sadeghi2003@yahoo.com

P. Van Damme
e-mail: Patrick.VanDamme@UGent.be

R. Zarkami
Department of Environmental Science, Faculty of Natural Resources, University of Guilan,
P.O. Box 1144, Sowmeh Sara, Guilan, Iran
e-mail: rzarkami2002@yahoo.co.uk

P. Van Damme
Faculty of Tropical Agrisciences, Czech University of Life Sciences Prague, Kamycka 129, Prague 6, Suchbátův 165 21, Czech Republic

Introduction

Freshwater exotic species are an issue of growing management concern (Vander Zanden and Olden 2008). They have one of the most harmful and least reversible impacts on natural ecosystems as they might change the local fauna and flora all around the world (Vitousek et al. 1996; Ricciardi and MacIsaac 2000). Exotic species are able to lower the ecological quality through changes in biological, chemical and physical properties of aquatic ecosystems (Olenin et al. 2007). On the top of this, the socio-economic damage is another important issue caused by invasive species.

Azolla filiculoides (Lam.) is one of the alien species that is already widely distributed in tropical, sub-

tropical and warm temperate regions, but particularly in Southeast Asia (Sweet and Hills 1971). It is a fast-growing fern with a doubling time of 2–5 days (Lumpkin and Plucknett 1982; Zimmerman 1985; Van Hove and Lejeune 2002; Taghi-Ganji et al. 2005). *Azolla* is a unique species among free-floating plants. In the scientific literature, however, various benefits have been mentioned for these floating ferns (e.g. nitrogen fixation and use of *Azolla* as green fertilizer), they are considered as troublesome weeds in invaded ecosystems.

A. filiculoides (Barreto et al. 2000) is a particular example of an alien species in many countries including Anzali wetland, northern Iran (Delnavaz and Ataei 2009). This floating water fern was intentionally brought from Philippine by Ministry of Agriculture for nitrogen fixation capacity. Since then, it has been spread in many wetlands in northern Iran particularly in Anzali wetland complex (JICA 2005). Due to its massive spread, some native macrophytes such as duck weed (*Lemna minor*) were completely wiped out and some of them are on the verge of extinction (JICA 2005).

For wetland restoration and conservation management, it is very important to get acquainted with the habitat preferences of exotic aquatic species. Therefore, it is necessary to examine the relationship between wetland characteristics and habitat requirements of exotic species. To achieve this goal, the use of suitable ecological methods is very important to successfully predict the habitat preferences of the target invasive species. Predictive modelling is one of the most important steps in the development of a standard habitat assessment protocol (Parsons et al. 2004). Ecological modelling can allow for the integration of physical, chemical and biological characteristics into measures, rather than just observations of causes and effects (Goethals and De Pauw 2001).

In recent years, various modelling techniques such as artificial neural network (ANN) (Robert 2003; Zhengfu and Fernando 2007; Gooyong et al. 2014), evolutionary polynomial regression (Elshorbagy and Baroudy 2009; Giustolisi and Savic 2009; Savic et al. 2009), classification trees (CTs) (Dzeroski et al. 2000; Dakou et al. 2007; Zarkami et al. 2010, 2014) and support vector machines (SVMs) (e.g. Zarkami 2011) have shown to be very powerful methods for assessing habitat requirements of organisms. CTs (Quinlan 1993) are such powerful tools allowing to predict

different characteristics such as presence/absence, biomass and abundance of various kinds of aquatic organisms. They are particularly useful to develop ecological data mining methods when dataset are limited (Goethals et al. 2007).

CTs can give insight in complex, unbalanced, non-linear ecological data where commonly used exploratory and statistical modelling techniques often fail to find meaningful ecological patterns (De'ath and Fabricius 2000). They yield reliable classifications with a transparent set of rules (Hoang et al. 2010). Due to their transparency and flexibility in use, CTs have recently gained in popularity (Hoang et al. 2010; Zarkami et al. 2012; Haghi Vayghan et al. 2015) and have been applied in a variety of ecological studies (Dakou et al. 2007; Everaert et al. 2011; Zarkami et al. 2012).

The SVMs (Vapnik 1995) is also a powerful method that implement a sequential minimal optimization (SMO) algorithm for training a support vector classifier using kernel functions (Platt 1998; Keerthi et al. 2001). They consist of input and output layers connected with weight vectors. SVMs maximise the margin around a hyperplane that separates two classes by mapping the input space into a high dimensional space or feature space. The mapping is determined by a kernel function such as polynomial kernel. SVMs have been applied for successful assessment of different types of organisms: e.g. macro-fauna community types (Akkermans et al. 2005), pike (*Esox lucius*) (Zarkami et al. 2012).

Multivariate analysis comprises a set of techniques dedicated to the analysis of data sets with more than one variable. This makes multivariate techniques suitable for analyzing ecological data which compose a number of environmental and biological data. Principal component analysis (PCA) is a multivariate data analysis technique, which is often used in different fields. In general, this method finds hypothetical variables (components) accounting for as much as possible of the variance in the multivariate data (Davis 1986; Harper 1999). Discriminant analysis (DA) method provides discriminant test for two or more groups (the latter is sometimes called canonical variates analysis). This module aims to find a transformation of input variables into latent variables (features) with maximum class separability (Fukunaga 1990). A scatter plot of specimens along the first two canonical axes produces maximal and second to

maximal separation between all groups. The axes are linear combinations of the original variables as in PCA, and eigenvalues indicate amount of variation explained by these axes.

Since exotic *Azolla* may cause possible threats to the rich original biological diversity, assessment on their habitat preferences can be helpful in order to restore, conserve and manage wetland ecosystems.

The present study primarily aims to compare the reliability of applied models (CTs and SVMs) using two performance criteria (CCI% and k). This would be an important issue to decide upon the most influential predictors deriving the habitat preferences of the target exotic species in the wetland. The main aim of the present research is to analyze the habitat preferences of *A. filiculoides* in Selkeh Wildlife Refuge (a protected area in Anzali wetland, north of Iran) using these two data-driven techniques. The aforementioned methods were chosen because they can both be used when dataset are not so big. Finally, the present paper aims to perform multivariate techniques using PCA and DA in order to assess the most important variables for the occurrence of *Azolla* in wetland.

More specifically, these types of researches will be useful for monitoring the most important variables for the target species in future.

Materials and methods

Description of study area

Anzali wetland complex is ecologically and internationally known as an important wildlife refuge particularly for a large number of birds. This wetland itself comprised three important areas including Siahkeshim Protected Area, Sorkahnkol Wildlife Refuge and Selkeh Wildlife Refuge (the study area). This international wetland consists of large, shallow, eutrophic freshwater lagoons (due to too much nutrients entering this part of wetland), shallow impoundments, marshes and seasonally flooded grasslands in the southwest Caspian lowlands. This wetland supports various species of fauna: breeding and wintering area of 77 migratory bird species (Mansoori 1995), a nursery and spawning habitat for 49 fish species (Abbasi et al. 1999), and a habitat for 31 species of mammals which reside in Anzali watershed (JICA 2005). More than 100 plant species occur in the

wetland. There are three types of plant communities (JICA 2005) which are commonly found in entire wetlands including: submerged communities (e.g. *Ceratophyllum demersum*), floating (e.g. *A. filiculoides* and *Nymphaea odorata*), and emerged (e.g. *Phragmites australis*). The whole area of the lagoon is covered by the submerged plant community. Some species of submerged plants such as *C. demersum* are very useful in Anzali wetland since they can accumulate the highest concentration of heavy metals (Ahmad et al. 2016). *A. filiculoides* forms a dense mat that covers approximately a quarter of Anzali wetland. Only the lagoon is spread of *A. filiculoides*. *Phragmites* is found in roughly a quarter of the wetland except for the lagoon. This kind of community generally lives in the shallow area of the eastern wetland and is widely distributed.

Various point and non-point sources of pollutants enter Anzali wetland. They originate from direct discharge of sanitary wastewater (produced from coastal cities), direct discharge of industrial wastewater (without or with less treatment), and agricultural activities in the surrounding areas. Pollutant loads negatively affect water quality and degrade structural characteristics of the wetland.

Figure 1 shows the study area (Selkeh Wildlife Refuge) which is an important refuge for migratory birds in Anzali wetland. The Selkeh has a total area of 360 ha which is located between latitude 37°22'58" and 37°26'51"N longitude 49°27' 09" and 49°28' 30"E. There are two rivers entering this area, namely Hendekhaleh and Trabkhaleh, both discharging various types of pollutants (e.g. untreated domestic, industrial and agricultural wastewaters) in this protected area (JICA 2005). Rice is one of the main crops which is cultivated in the surrounding area. Application of fertilizers, pesticides and herbicides (in the paddy fields) has caused eutrophication problem in this aquatic ecosystem. Moreover, various industrial and urban wastes have negatively impacted the wetland (JICA 2005).

Sampling sites and study design

Seven sampling sites were considered during 1 year study period. The selected sites were monthly sampled during 12 sampling campaigns (from October 2007 till September 2008) with taking 84 total dataset into account.

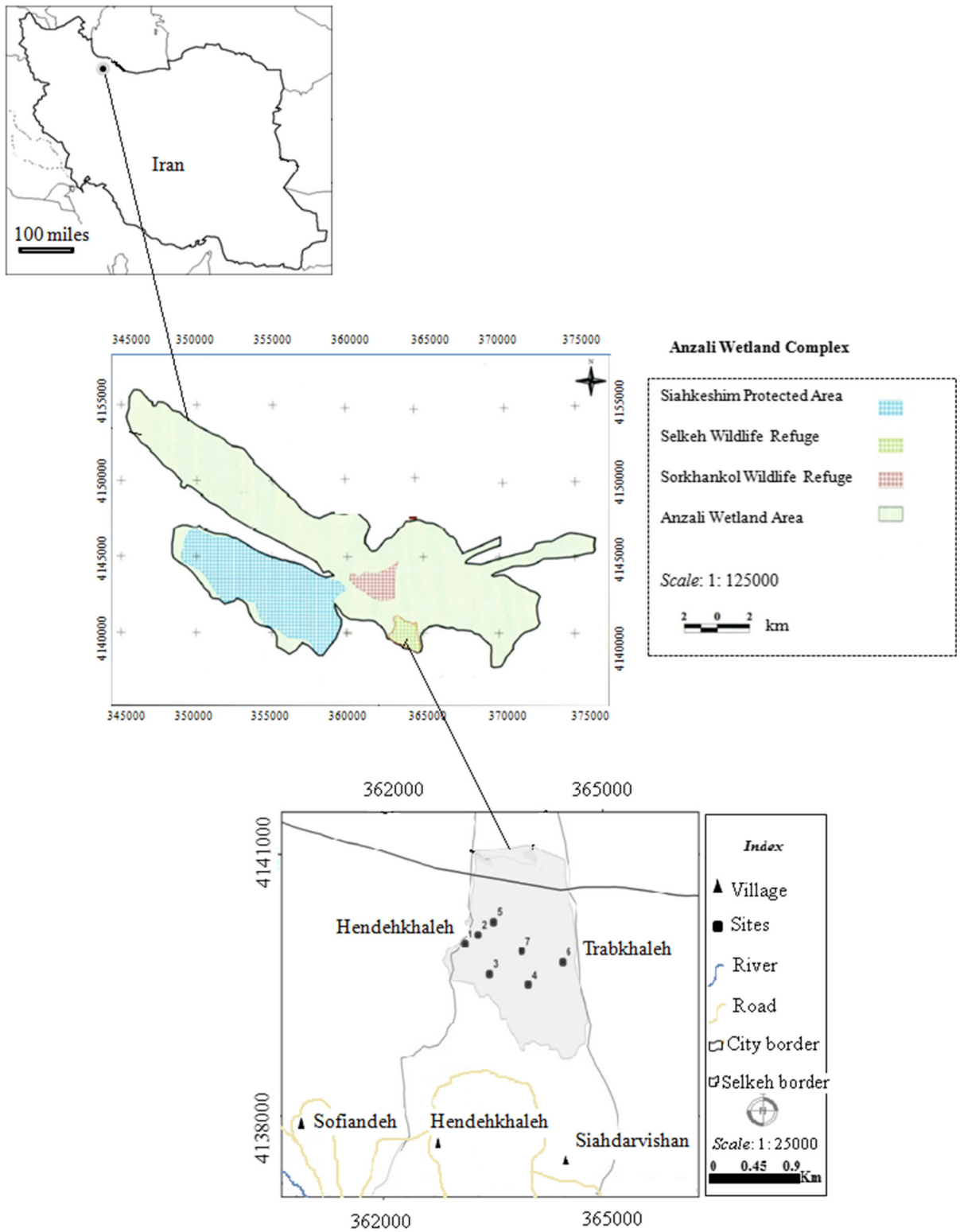


Fig. 1 Location of the sampling sites in Selkeh Wildlife Refuge, south of Anzali wetland, northern Iran (Source Guilan Environment Protection Bureau 2007)

The most important criteria that were considered for the site selection were based on variations in natural characteristics (e.g. drying out of wetland in dry season), anthropogenic influences (e.g. resulting from domestic, industrial and non-point sources), feasibility sampling (e.g. accessibility for boat transport) and ecological considerations (e.g. distribution of *Azolla* over the entire wetland). Systematic sampling (as described by Buckland et al. 2007) was used to take the samples at fixed intervals along the length of the wetland. To do so, two parallel transects, spaced at 500 m intervals, were run across the west-east gradient of the wetland (as illustrated in Fig. 1). The first transect consisted of sampling points of 1, 3 and 4, whereas the second transect comprised sampling points of 5, 7 and 6. Since point 1 was located in the river, one additional sampling point (no 2) was considered between points of 1 and 5 in order to get more wetland-related information. As a whole, the 7 sites completely differed from each other regarding ecological and geographical conditions. The sampling site 1 located near the river had no direct connection to the other sampling sites. The other six sites (the sites of first and second transects), however, were located in the wetland; they were separated from each other with a geographical barrier such as a ridge. The sites located within each transect were also independent to each other. For instance, in the first transect, one site was surrounded by the reeds (e.g. *Phragmites australis* and *Typha latifolia*) and another one was located in the open area and the third one was selected in the location dominated by *Trapa natans*. The independence of sites was also taken into account for the second transect.

As stated already, the distance between sampling sites was approximately chosen 500 m. The exact geographic location of each sampling site was determined using a GPS (GARMIN, etrex Vista). The selected sites were considered as the optimal sampling design for measuring the biotic (cover percentage of the exotic fern) and abiotic variables in the wetland. The other parts of wetland were not considered for site selection because of some reasons: dense cover of some floating (e.g. *Nelumbium nuciferum*) and emergent plants (e.g. *Phragmites* spp. and *Typha* spp.) caused the boat passage through the wetland impossible. On the other hand, the low depth of wetland (particularly in dry seasons) intensified the problem of boat transport through the wetland. On top of this,

sampling was impossible in some parts of wetland due to drying out of ecosystem in dry seasons.

At each sampling site, a set of wetland characteristics including the water quality and physical-habitat variables were measured (Table 1). Since the data-mining techniques applied in the present study (CTs and SVMs) are less affected with multicollinearity (high correlation between two or more predictor variables), the whole variables were statistically considered as inputs to the techniques. For the multivariate techniques (PCA and DA), one of the highly correlated variables were eliminated from the set of variables because they had no added values for analyzing the habitat preferences of the exotic fern. In total, 12 variables were ultimately used for the multivariate techniques.

Among the abiotic variables, it was assumed that air temperature and sunlight (the number of sunny hours) might stimulate the *Azolla*'s growth in the study area. As shown in Fig. 2, the highest air temperature and the maximum number of sunny hours were recorded in July and June, respectively.

The water quality samples were measured on a monthly basis near the wetland surface. In order to have a regular measurement, the samples were taken in the middle of each month. The chemical variables were analyzed based on standard methods (as described by APHA/AWWA/WEF 1998).

Field measurements for the physical-habitat variables were monthly measured by the Bandar-E Anzali Weather station, northern Iran (latitude 37°28'21"N and longitude 49°27'43"E). The water depth of wetland was measured using a yardstick.

Data related to *A. filiculoides* (cover percentage) was used as a response variable for all methods. This variable was simultaneously sampled with the abiotic data at each month. It was measured using a percentage cover class based on a modified Braun-Blanquet cover-abundance scale (Sumners and Archibold 2007). Cover percentage of invasive species was measured using 1 m² quadrat. In order to obtain an accurate estimate of cover percentage, the quadrat was divided into a grid of 100 squares. At each sampling site, the quadrats were repeated three times. Then, the average of three quadrats (observed cover percentage) was made in order to have an ultimate estimation. Depending on the extent of occurrence in the sampling sites and also to make an easy interpretation of the obtained results, the cover percentage of *A. filiculoides* was divided into three ordinal classes. The number of

Table 1 Summary statistics (minima, maxima, median, average and standard deviation) of the input variables used for modelling habitat requirements of *A. filiculoides* in sevensampling sites of Selkeh Wildlife Refuge (Anzali wetland, northern Iran) over a 12 months' study period (2007–2008) ($n = 84$)

Variables	Abbreviation	Unit	Min	Max	Median	Average \pm SD
Turbidity	–	FTU	0.74	125	15.25	17.29 \pm 17.46
Electric conductivity	EC	$\mu\text{S cm}^{-1}$	334	14,410	1360.50	2505.14 \pm 3337.98
Total hardness*	–	mg l^{-1}	112	1180	306	397.93 \pm 275.69
Humidity	–	%	73	90	84.50	78.92 \pm 3.53
Salinity	–	g l^{-1}	0.00	8.80	0.30	1.28 \pm 2.14
pH	–	–	6.92	8.73	7.56	7.63 \pm 0.31
Total dissolved solids	TDS	mg l^{-1}	279	4388	1009	1262.94 \pm 951.13
Total suspended solids	TSS	mg l^{-1}	0.40	165	27.40	40.50 \pm 37.66
Bicarbonate	HCO_3^-	mg l^{-1}	101.61	509.88	210.59	209.29 \pm 61.39
Carbonate	CO_3^{2-}	mg l^{-1}	0.17	10.01	0.65	1.12 \pm 1.28
Free CO_2	F- CO_2	mg l^{-1}	1.77	54.18	9.84	12.25 \pm 9.32
Total CO_2	T- CO_2	mg l^{-1}	89.58	453.11	185.53	184.78 \pm 54.39
Air temperature	T	$^\circ\text{C}$	6.70	26	15.90	16.18 \pm 6.78
Sunny hours	–	H	87.80	254.20	129.65	148.39 \pm 60.87
Ammonium	N- NH_4^+	mg l^{-1}	0.03	3.93	0.43	0.87 \pm 0.96
Nitrite	N- NO_2^-	mg l^{-1}	0.00	0.20	0.03	0.04 \pm 0.04
Nitrate	N- NO_3^-	mg l^{-1}	0.00	1.04	0.26	0.30 \pm 0.22
Organic nitrogen	Org-N	mg l^{-1}	0.01	4.33	0.56	0.95 \pm 1.05
Total nitrogen	T-N	mg l^{-1}	0.11	7.41	1.77	2.13 \pm 1.56
Dissolved oxygen	DO	mg l^{-1}	0.00	10.02	6.46	5.45 \pm 2.54
Biological oxygen demand	BOD ₅	mg l^{-1}	2	16	9	9.01 \pm 3.27
Chemical oxygen demand	COD	mg l^{-1}	13	98	44.80	50.80 \pm 22.04
Magnesium	Mg^{2+}	mg l^{-1}	4.86	221.75	39.87	64.69 \pm 61.26
Calcium	Ca^{2+}	mg l^{-1}	9.62	240.48	44.88	52.63 \pm 34.09
Freezing days	–	–	0.00	3.40	0.00	0.68 \pm 1.23
Sodium	Na^+	mg l^{-1}	19.29	473.97	105.30	115.45 \pm 77.27
Potassium	K^+	mg l^{-1}	0.07	15.86	4.04	5.86 \pm 4.35
Depth	–	m	0.30	1.20	0.50	0.53 \pm 0.20
Chloride	Cl^-	mg l^{-1}	1.10	232.88	19.88	46.35 \pm 60.07
Sulphate	SO_4^{2-}	mg l^{-1}	13.87	965.40	71.39	173.16 \pm 231.17
Orthophosphate	o- PO_4^{3-}	mg l^{-1}	0.01	0.21	0.05	0.07 \pm 0.04
Total phosphate	T- PO_4^{3-}	mg l^{-1}	0.03	1.56	0.37	0.42 \pm 0.28
Iron dissolved	Fe	mg l^{-1}	0.26	10.20	1.08	2.19 \pm 2.65

FTU formazin turbidity unit

* Expressed as $\text{CaCO}_3 \text{ mg l}^{-1}$

real observations for the classes of low, medium and high accounted for 40, 22 and 22, respectively.

Data processing and analysis methods

Before applying the data driven and multivariate techniques, data were primarily tested for statistically

normality. The data of some variables such as SO_4^{2-} , NH_4^+ , organic nitrogen and o- PO_4^{2-} were not normally distributed (the outliers of these variables were shown with the box—and—whisker plots). Since the outcomes of classification tools such as CTs and SVMs are not strongly influenced by wide ranges of data values, no transformation of the data was applied

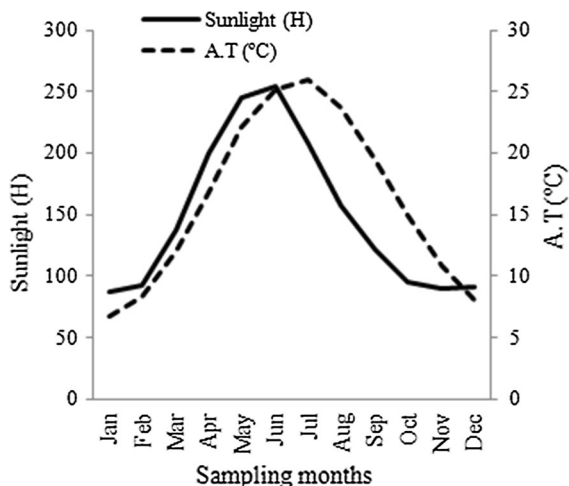


Fig. 2 The trend of A.T °C (air temperature), sunlight (the number of sunny hours) and depth in the Selkeh Wildlife Refuge (Anzali wetland, north of Iran)

for the aforementioned models. In contrast, log-transformation was merely made for the multivariate techniques applied in the present paper.

The Pearson correlation (*r*) was used to find the correlation between pairs of input variables and also between input variables and cover percentages of the target species. No collinear variables were dropped for the data mining techniques, while the removal of one of the highly correlated variables was only considered for the multivariate techniques.

Then data mining techniques (CTs and SVMs) (Witten et al. 2011, version 3.6.6) were used as the main models to analyze the habitat preferences of *A. filiculoides*. The output variable for the data-driven methods was the cover percentage of *A. filiculoides*, whereas the input variables were a set of water quality and physical-habitat characteristics of wetland. First, the method was applied based on CTs using all input predictors. For the training and validation of CTs, different fold cross-validation (from 2 to 10) were tested in order to get a reliable estimate of the model error and to avoid overfitting of the model. On the basis of this, CTs stability was best maximized using a 3-fold cross-validation. This was based on the highest predictive outcomes of CTs resulted from the predicted vs. observed values. The J48 with different intensities of pruning confidence factor (PCFs) were induced by changing the confidence factor into 0.01, 0.10, 0.25 and 0.50 values.

Akin to CTs, the output and input values of SVMs was the cover percentage and wetland characteristics, respectively. In the present research, the polynomial kernel was used. The parameter settings were default values except for the exponents of the polynomial Kernel (the exponents were tested from 1 to 10) (Witten et al. 2011).

The percentage of correctly classified instances (CCI%) and Cohen’s kappa statistic (*k*) (Cohen 1960) were used for assessment of the two techniques. Both criteria were based on the confusion matrix (the observed values vs. predicted ones) (Table 2). The predictive performances of the both techniques were identified with true positive (TP = *a*), false positive (FP = *b*), False Negative (FN = *c*) and True Negative (TN = *d*) cases obtained from each model (Fielding and Bell 1997) (Table 3). The degree of agreement in *k* was based on the following ranges (Landis and Koch 1977): ≤0 (poor); 0–0.20 (slight); 0.20–0.40 (fair); 0.40–0.60 (moderate); 0.60–0.80 (substantial) and 0.80–1 (almost perfect) and also the models with a CCI > 70% and *k* > 0.40 were considered to be reliable.

Paired Student’s *t*-tests (a two-tailed test with a 95% confidence interval) were conducted for the comparison of the predictive performance of the two applied methods.

Table 2 The confusion matrix as a basis for the performance measures with true positive values (TP), false positives (FP), false negatives (FN) and true negative values (TN)

	Observed	
	+	–
Predicted		
+	a (TP)	b (FP)
–	c (FN)	d (TN)

Table 3 Measures of predictive accuracy calculated from a confusion matrix

Performance measure	Formula
Cohen’s Kappa (<i>k</i>)	$\frac{[(a+d) - ((a+c)(a+b) + (b+d)(c+d)) / N]}{[N - ((a+c)(a+b) + (b+d)(c+d)) / N]}$
CCI	$(a + d) / N$

True positive (TP = *a*), false positives (FP = *b*), false negatives (FN = *c*) and true negative (TN = *d*)

Multivariate techniques (PCA and DA) were used to examine the environmental variables affecting the prevalence of *Azolla* in the study area. These two techniques were applied from the program package PAST (Paleontological Statistics, version 3) (Hammer 2013). PCA based on the first and second components were used to determine the most important variables for exploring the occurrence of the exotic fern in the Selkeh Wildlife Refuge. DA was performed for the specimens with three groups of *Azolla*'s cover classes to find the relation between the cover classes of *Azolla* and different seasons and sampling sites in the study area.

Results

Correlation analysis

The correlation analysis showed that *Azolla*'s cover percentage is strongly and positively correlated with the number of sunny hours ($r = 0.79$; $p < 0.01$) and air temperature ($r = 0.86$; $p < 0.01$) so the highest and lowest cover percentage of *Azolla* in relation to the aforementioned variables were observed in the dry (e.g. in summer) and wet seasons (e.g. in winter), respectively. The correlation analysis of wetland characteristics also showed that some chemical variables were strongly correlated. For instance, the high correlation was found between total nitrogen and NH_4^+ ($r = 0.71$; $p < 0.01$), total nitrogen and organic nitrogen ($r = 0.73$; $p < 0.01$).

Among the entire variables recorded in the wetland, the mild and extreme outliers were mainly recognized in water quality variables such as SO_4^{2-} , NH_4^+ , organic nitrogen, and o-PO_4^{3-} (Fig. 3). As illustrated here, the outliers are skewed to the upper part of the box plots. For the most values of these nutrients, the measurements were probably accurate due to the contaminated sites by the nutrients.

Comparison of predictive performances of CTs and SVMs

Among nine times cross-validation tested (from 2 to 10), the best predictive performance of CTs for *Azolla* were obtained with the 3-fold cross-validation. On the basis of this, the given value was used for testing different PCF levels to the entire data. The outcomes

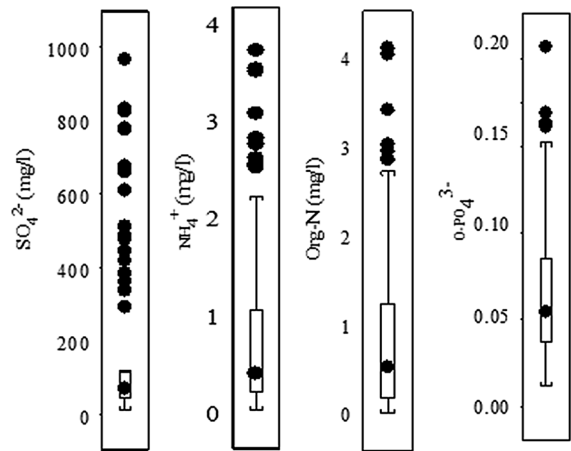


Fig. 3 The box and whisker plots (with minimum, median, maximum, the lower quartile ($Q1 = 0.25(N + 1)$) and upper quartile ($Q3 = 0.75(N + 1)$) representing distribution ranges of the water quality variables recorded in the Selkeh Wildlife Refuge ($n = 84$) (Org-N organic nitrogen)

of the CTs with 4 pruning levels are presented in Table 4. As seen here, different pruning affected the number of leaves, tree size and performance criteria.

However, when PCFs were applied at four levels (0.50, 0.25, 0.10 and 0.01), the predictive results were relatively stable over four different PCF levels. Based on results obtained, average CCI (%) for analyzing the habitat needs of *Azolla* ($\text{CCI} = 71 \pm 1.40\%$) indicated that more than 70% of instances were correctly classified. This implicitly demonstrated that the CCI gave a reliable outcome. Average k was also reliable since it met the threshold value ($k > 0.40$) to yield a trustworthy results ($k = 0.55 \pm 0.02$).

In addition to predictive performances, the number of leaves and sizes of each tree were also considered in order to test the complexity of the induced model. A very complex tree was constructed at PCF 0.50. Here, number of leaves (12) and tree size (23) revealed that the induced tree was difficult to use for interpreting *Azolla*'s habitat requirements. More pruning of the induced trees resulted in an easy and better understanding of habitat requirements of the species. In contrast, at PCF 0.01, only five leaves (with tree size of 9) were constructed but this level was not considered for the interpretation because it provided less information on *Azolla* in sampling sites. The PCF 0.10 was less reliable in terms of the two evaluation criteria used here. For this reason, it was not used for the interpretation of outcomes. However, there was no a

Table 4 The predictive outcomes of J48 for *Azolla* at Selkeh Wildlife Refuge (Anzali wetland, northern Iran) with four pruning confidence factors (PCFs)

PCFs	CCI (%)	k	# Leaves	Size of trees
0.50	71.40 ± 0.84	0.55 ± 0.01	12	23
0.25*	72.00 ± 0.65	0.60 ± 0.01	10	19
0.10	69.10 ± 1.60	0.50 ± 0.03	8	15
0.01	71.40 ± 2.52	0.55 ± 0.04	5	9
Mean ± SD	71.00 ± 1.40	0.55 ± 0.02	8.75 ± 2.99	16.5 ± 5.97

CCI (%) correctly classified instances, k Cohen's kappa, SD standard deviation, #leaves the number of leaves

* PCFs: Pruning Confidence Factors; PCF = 0.25 was used for the eventual calculation of habitat requirements of *A. filiculoides*

significant difference over the four levels of pruning, the PCF 0.25 was finally selected as the optimal pruning level since the highest values of the predictive results were obtained at this level of pruning (CCI = 72.00 ± 0.65% and $k = 0.60 ± 0.01$). Therefore, this value was used for subsequent model application and evaluation of occurrence of *Azolla* in the sampling sites. The confusion matrix (Fig. 4a) presents the performance of the induced tree by visualizing the distribution of the instances around the diagonal of the matrix. About 72% of instances were correctly classified and the remaining of instances (28%) was misclassified.

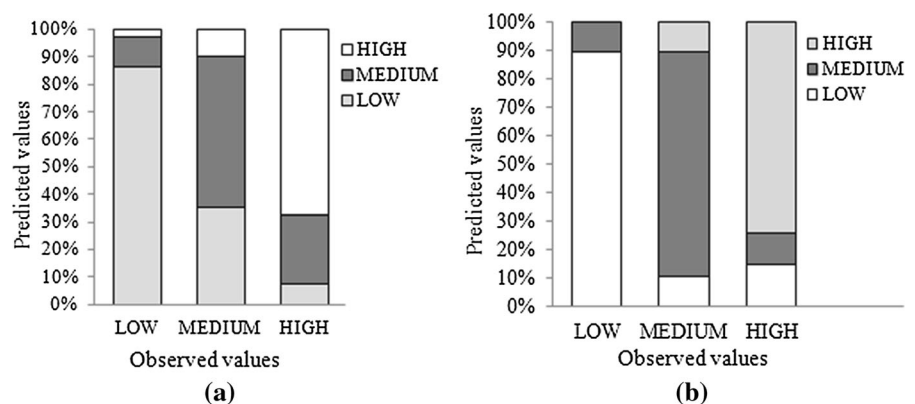
In SVMs, the best and highest predictive results (according to the two criteria) was obtained in the fold of 4. This cross-validation was ultimately checked based on the application of different exponents (from one to ten). The experimental results showed that the performance of SVMs was sensitive to the various exponents. The best and highest predictive outcomes were obtained using an exponent of one (using a linear support vector machine). Testing other exponents of

the polynomial kernel caused a higher risk of over-fitting. This led to lower the predictive performances of the applied models. Therefore, SMO with $\exp = 1$ was then run ten times after randomisation in order to check for robustness and reproducibility.

The outcomes of model showed that the habitat preferences of *Azolla* were successfully predicted by SVMs (CCI = 82.10 ± 3.35% and $k = 0.72 ± 0.05$). As illustrated in the confusion matrix (Fig. 4b), more instances were correctly classified with the SVMs.

Paired Student's t -tests (a two-tailed test with a 95% confidence interval) were conducted for the comparison of the predictive performance of models based on the two applied methods. The results of Paired Student's t -tests (p value = 0.001) showed that there was a significant difference between the CTs and SVMs regarding the two predictive criteria. The test showed that the SVMs performed significantly better than CTs for both criteria. In other words, a better predictive result was obtained for *A. filiculoides* with the SVMs.

Fig. 4 a, b Confusion matrices of J48 (a) (PCF 0.25) and SMO (b) ($\exp = 1$) for *Azolla*'s cover percentage classes



Rule induction with the CTs (J48 algorithm)

In total, among whole variables introduced to CTs, only nine predictors (four physical-habitat and five water quality variables) were constructed by the induced trees as the most important predictors to explain the occurrence of *Azolla* (Fig. 5). As shown here, tree was split in two portions based on air temperature. The left side of induced tree is related to the wet seasons (when the air temperature is dropping below 12 °C), while the right side of tree is more related to the dry seasons when air temperature is gradually increasing. This means that air temperature was the most important driving predictor for the habitat needs of *Azolla* in the wetland. The aquatic fern would prefer almost a medium range of air temperature (12 °C) for its growth in the Anzali wetland (mean A.T °C \pm SD = 16.18 \pm 6.78 °C). When air temperature falls below 12 °C, the relative humidity would then play the key role for its occurrence in this valuable ecosystem. In higher humidity (>82%), then prevalence of *Azolla* would be low, while if relative humidity was \leq 82%, then total suspended solids (TSS) concentration would be more important. Here, *Azolla* showed a low prevalence in the sampling sites

when TSS concentration exceeded 8 mg/l, while with a TSS \leq 8 mg/l, it had medium prevalence.

Depth of water layer also had a great importance in the study area. This variable grew after air temperature confirming its high ecological significance in the wetland. All other variables were related to the water depth. The minimum and maximum depth recorded in Selkeh Wildlife Refuge was 0.30 and 1.20 m, respectively (mean depth \pm SD = 0.53 \pm 0.20 m). Apparently, *Azolla* preferred a medium depth for its growth in the wetland (0.50 m). In the deeper parts of wetland (>0.50 m), the assessment of *Azolla* cover was only dependent on water quality variables (such as organic nitrogen and chemical oxygen demand). Increasing the organic nitrogen concentration will in turn lead to a medium prevalence of this species. In addition, rising chemical oxygen demand concentration in the wetland (>35 mg/l) would lower *Azolla* distribution and reverse. In shallow water (<0.50 m), the amount of sunlight contributed to the prediction. The invasive aquatic fern might expand its growth in the wetland when the amount of light intensity surpassed 207.8 h per month. But when light intensity dropped below 207.8 h, the occurrence of *Azolla* became dependent on orthophosphate and nitrate concentrations,

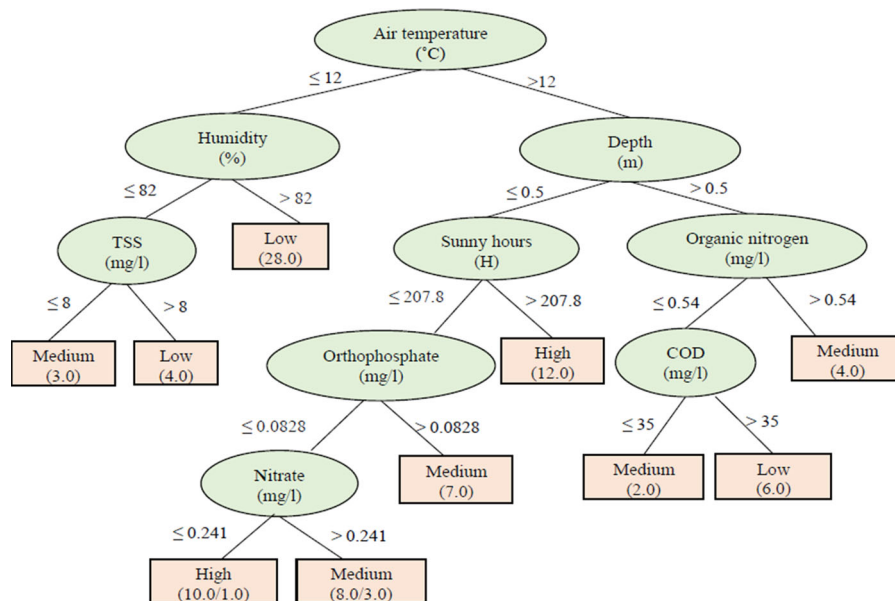


Fig. 5 J48 (PCF 0.25) for predicting the habitat preferences of *A. filiculoides* in Selkeh wildlife refuge (COD chemical oxygen demand, TSS total suspended solids). L, M and H represent low, medium and high cover percentage of *A. filiculoides*, respectively). Values between brackets in the rectangles indicate instances in which rules are true/false

medium and high cover percentage of *A. filiculoides*, respectively). Values between brackets in the rectangles indicate instances in which rules are true/false

respectively. So orthophosphate concentration is higher than 0.08 mg/l, the invasive fern will show a medium prevalence. When light intensity is below 207.8 h the amount of nitrate will then play an important role for *Azolla* in the study area. *Azolla* tended to show a high prevalence when nitrate concentration reached below 0.24 mg/l.

Weighing attributes using SMO

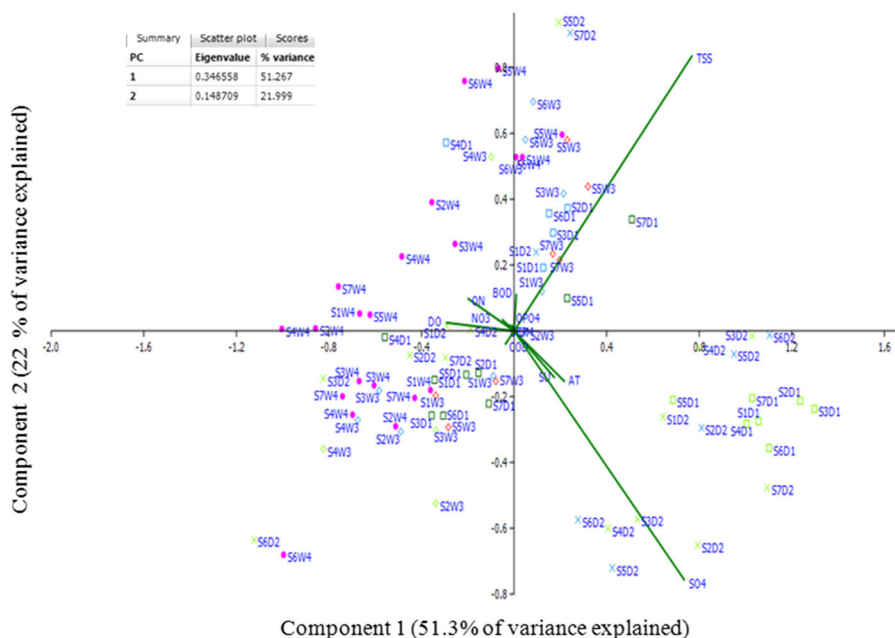
SVMs models gave different attribute weights for each input variable. Variables with an absolute weight value >0.50 were considered as very important predictors. The depth of wetland, dissolved oxygen, sulphate, air temperature, the number of sunny hours, humidity, orthophosphate, biological oxygen demand and TSS played an important role for assessing the habitat needs of *Azolla* in the study area (each predictor with an absolute weight >0.50). Salinity, ammonium and total nitrogen had an intermediary contribution to the prediction (0.40 < attribute weight < 0.50). Other variables (number of freezing days, total dissolved solids (TDS), organic nitrogen, potassium and total phosphate) had a less effect on the growth of this invasive species (each variable with a weight <0.10). Based on the attribute weights provided by the SVMs, most habitat variables (except freezing days) were more important variables than the water quality ones.

PCA

To examine the impact of spatio-temporal patterns on habitat conditions of *Azolla*, a PCA biplot (using log transformed values) was performed using 12 environmental variables in different sampling sites in dry and wet seasons (Fig. 6). The distribution of the samples showed clear seasonal patterns so that based on the first and second principal components, seasonal variations showed dissimilarity between dry and wet seasons. The first component describes 51.3% (with eigenvalue of 0.35) and second component explains 22% (with eigenvalue of 0.15) of total variations, respectively.

From the biplot, it can be observed that the most important environmental variables influencing the habitat preferences of the exotic fern in the Selkeh wetland were related a combination of physical-habitat and water quality variables. TSS, sulphate, air temperature and the number of sunny hours were among others affecting the occurrence of *Azolla* population mostly in spring and summer periods. In reverse, low cover percentage of *Azolla* was mainly attributed in the wet seasons (particularly in the winter period). In contrast to temporal patterns, there are no clear spatial patterns for the invasive fern in the different sampling sites. Nevertheless, among different sites, the sites of five and seven provided relatively

Fig. 6 PCA biplot of the samples collected in the dry (D) and the wet (W) seasons in the Selkeh Wildlife Refuge using log transformed values of water quality and habitat variables. D1, D2, W3 and W4 represent the sampling seasons which took place in spring, summer, autumn and winter, respectively. Si represents the number of sampling sites from 1 to 7. The first component explains 51.3% (with eigenvalue of 0.35) and second component describes 22% (with eigenvalue of 0.15) of total variations



a suitable habitat condition for *Azolla* in the wetland particularly in dry seasons.

DA

Akin to PCA, a DA was performed with 12 environmental variables (using log transformed values) collected in the same sampling sites and seasons (Fig. 7). Multivariate DA showed that the three groups of *Azolla*'s cover class were distinctly classified. The first discriminant axis describes 82.35% of the total variability and the second axis 17.65% of all eigenvalues, respectively. The projection of the two Canonical Variates displays a clear separation of low cover class with only small area of overlap with the cluster of high cover class while medium cover class has relatively a big area of overlap with low and high cover classes. From the DA biplot, it can be seen that the most important environmental variables affecting the high occurrence of *Azolla* population in dry seasons (as depicted with the green convex hulls) were related to TSS, SO_4^{2-} , air temperature and the

number of sunny hours. On the contrary, wet seasons (in particular winter period) supported less population of *Azolla* (low prevalence of *Azolla* is shown with the red convex hulls). Chemical variables such as the dissolved oxygen, organic nitrogen and nitrate concentrations and physical-habitat variables (e.g. depth and humidity) were observed in the opposite of high cover class.

Discussion

The outcomes of CTs and SVMs showed that the methods were able to provide a reliable prediction for the habitat preferences of *A. filiculoides* in the protected area. Many studies (Hoang et al. 2010; Zarkami et al. 2010; Zarkami 2011) stated that models should meet at least a threshold value of $CCI \geq 70\%$ and $k \geq 0.40$ in order to have a reasonable assessment for the target organism.

The use of optimal tree pruning (so-called PCF) in the J48 reduced the complexity of tree. By doing so,

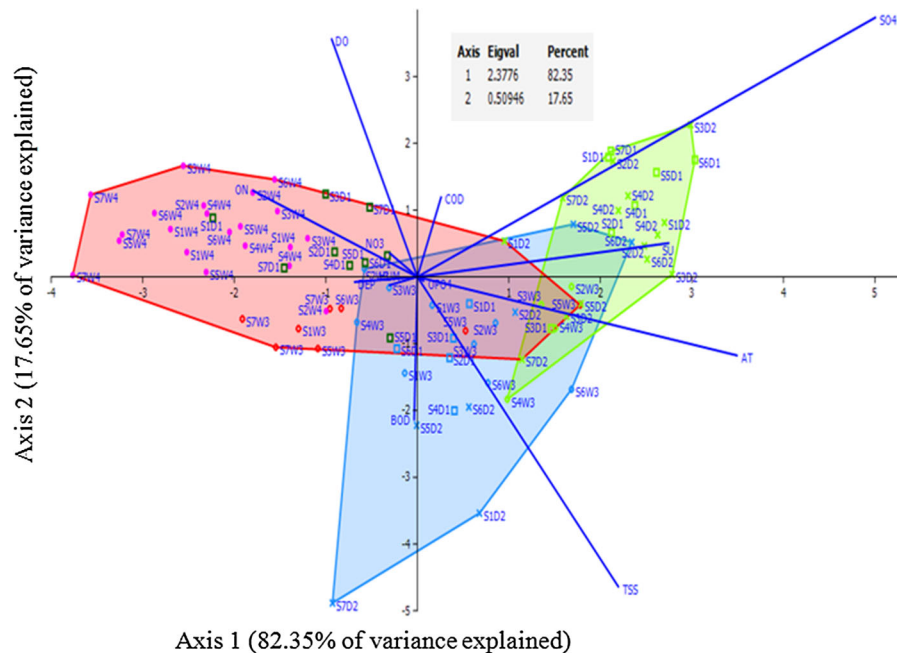


Fig. 7 Scatter plot of DA showing three groups of *Azolla*'s cover class based on 12 water quality and physical-habitat variables (green, blue and red convex hulls represent high, medium and low cover classes of *Azolla*, respectively). The first axis describes 82.35% of the total variability (with eigenvalue of 2.4) and second axis explains 17.65% of the total variability (with eigenvalue of 0.5), respectively. The specimens belonging

to each group are represented with different symbols. TSS total suspended solids, SO_4 sulphate, AT air temperature, SU the number of sunny hours, Dep depth, DO dissolved oxygen, ON organic nitrogen, SiDi represents the number of sampling sites from 1 to 7 in dry seasons (spring and summer), SiWi the number of sampling sites in wet seasons (autumn and winter). (Color figure online)

the induced model could become more transparent to assess the habitat preferences of organisms (Zarkami 2011). Consequently, the constructed trees allowed for a correct and easy ecological interpretation of selected variables for the exotic fern in the study area. In addition, the use of attribute weights in SVMs could provide more reliable predictive results than CTs for analyzing the habitat requirements of organisms proving that the SVMs are more robust than CTs to solve a range of problems which are observed as a noise in datasets (Hoang et al. 2010). However, for management of wetland ecosystems, the CTs are preferred over SVMs due to a better visualization of the selected predictors by wetland managers.

All data-driven (CTs and SVMs) and multivariate methods (PCA and DA) showed that the both physical-habitat and water quality variables might influence the habitat preferences of *A. filiculoides* in the Selkeh wetland. However, according to decision rule made by J48 and attribute weight of SMO, the physical-habitat variables might have to some extent more impact than water quality ones on the occurrence of the exotic fern in the wetland (Sadeghi et al. 2012). The biplot of multivariate techniques showed that the habitat preferences of the exotic fern in the wetland are more influenced by the water quality variables than physical-habitat ones.

Therefore, according to the both data-driven methods, air temperature, humidity, the number of sunny hours and the depth of wetland were important predictors in the wetland. There is a close and positive relationship between air temperature and the high abundance of *Azolla* in tropical and subtropical climates (Hill 2003; Van Der Heide et al. 2006). This predictor is a main factor in stimulating the growth of some invasive aquatic plants like water hyacinth (*Eichhornia crassipes*) and *A. filiculoides* in tropical regions (Kannaiyan and Somporn 1989). This implies that seasonality (the effect of different seasons) can have an important impact on the growth of *A. filiculoides* in the Anzali wetland. When temperature goes above 30 °C or drops to below −4 °C, the growth of this aquatic fern is highly restricted (Serag et al. 2000; Liu et al. 2008; Fernández-Zamudio et al. 2010). The given ranges were almost in accordance with the mean range recorded in the Anzali wetland.

Another important predictor in determining the growth of *Azolla* at the Selkeh Wildlife Refuge was the depth of wetland. Very deep or very shallow waters in the wetland can restrict the growth and biomass

production of the *A. filiculoides* (Biswas et al. 2005). This means that this aquatic fern requires a suitable depth for its optimal growth and biomass production. Apparently, many parts of Anzali wetland offer a suitable depth for the establishment of *A. filiculoides* (JICA 2005). The optimal depth for *A. filiculoides* is the places where emergent plants (e.g. cattails, *Typha latifolia*, common reed, *Phragmites australis*, and sedges, *Carex* spp..) become established at the edge of the ecosystem. According to the biplots of PCA and DA, the deeper part of wetland can restrict the prevalence of *Azolla* in wet seasons (low prevalence of *Azolla* is mainly found in wet seasons when the depth of wetland is increasing).

Other physical-habitat variables (except the number of freezing days) were also important factors to meet the habitat needs of the exotic fern in the sampling sites. The number of freezing days had almost no contribution to the evaluation of occurrence of the exotic species in the Selkeh wetland. The fact that the number of days dropping to below freezing is very short in the Anzali wetland, therefore this predictor cannot be considered as a determinant factor for the growth of *A. filiculoides*. Since there is enough humidity in the northern part of Iran, Anzali wetland is considered as a suitable place for the survival of *A. filiculoides* so that this variable was also considered as driving predictor in the wetland. Seemingly, this invasive species cannot survive in other parts of Iran due to a relative humidity <60%. In general, low humidity (<60%) makes *Azolla* very weak because an increase in *Azolla* biomass is somewhat dependent on air humidity as at less than 60% of relative humidity, the fern becomes dehydrated and fragile (Bocchi and Malgioglio 2010). Although, it has to be noted that very high humidity might also play an inhibitory role for the growth of *Azolla*. This is obvious from decision made by the constructed trees where low cover percentage of *Azolla* coincides with very high humidity (>82%). Also based on the ordination techniques, the low cover percentage of *Azolla* can be attributed to the winter season where the maximum humidity is found in the given period.

In addition to the physical-habitat variables, some chemical predictors such as organic nitrogen, TSS, SO_4^{-2} , DO and o-PO_4^{-3} concentrations (according to both data-driven methods) as well as a combination of these variables (based on PCA and DA methods) can also enhance *A. filiculoides*'s growth and sporulation

(Janes 1998). Particularly the lack of phosphorus might decrease or even stop *Azolla*'s growth (e.g. Watanabe and Espinas 1976). In contrast, low concentration of nitrate cannot restrict the growth of this exotic fern (Sadeghi et al. 2012). This was very well confirmed by the applied methods in the present study. The application of various types of fertilizers (in particular sulphate and phosphate) in rice fields is a main source of these nutrients in the wetland. High concentration of various types of nutrients in the Selkeh wetland might result in a significant decrease in dissolved oxygen concentration as confirmed with attribute weight of SVMs and also ordination techniques of PCA and DA in the dry seasons.

In contradiction of temporal patterns, no clear spatial patterns were found for the exotic fern in the different sampling sites. Though, some sites (e.g. 5 and 7) provided relatively a suitable habitat condition for *Azolla* in dry seasons in the wetland due to enough nutrients (e.g. sulphate), adequate sunny hours and air temperature. On top of this, these sites were in the vicinity of emergent plants (such as *Phragmites australis*, *Sparganium erectum* and *Typha latifolia*). These plants are able to create a windbreak for supporting of *Azolla* in the wetland (JICA 2005).

However, all applied techniques presented the most important explanatory wetland characteristics for analyzing the habitat preferences of *A. filiculoides* in the wetland, the selected variables might not be exactly the only ones for assessing the habitat requirements of the target organism (Ambelu et al. 2010). In other words, the methods (based on available information) would give the priority to the most important predictors. If those physical-habitat variables selected by the data-driven methods (particularly air temperature, humidity, depth...) were excluded from the models, other variables (particularly nutrients) would be more important ones. Sometimes the correlation in the data is ecologically meaningful so some variables (e.g. nutrients) were strongly correlated. However, the SVMs and CTs are generally less affected by correlated variables (Zarkami et al. 2012), those very high correlated variables might to some extent prevent the methods for selecting both variables so that any r above 0.20 would cause such data-driven models unstable (Goethals 2005).

However, gathering a big dataset might make the models more reliable, the historical datasets will not necessarily improve the predictive performances of

models if the additional series are noisy or unrelated to the target variable (Boivin and Ng 2006). On the top of this, if there are not huge variations of environmental gradients over a historical period, 1 year data collection (albeit with considering monthly sampling into account) will be sufficient to enhance the predictive power of data-driven models (as it is already performed for CTs and SVMs in the present research).

Moreover, the study of the habitat needs of invasive aquatic ferns is to some extent difficult because the biotic and abiotic variables that influence their growth are complex. Particularly, in contrast to other *Azolla* species, the assessment of habitat requirements of *A. filiculoides* is not so easy because this fern can somewhat tolerate a variety of environmental conditions (Karatayev et al. 2009). On top of this, invasive species are more tolerant than native ones to environmental pollutants (Devin and Beisel 2007).

Conclusions

Based on the outcomes of present work, it is concluded that when datasets are limited the SVMs might yield more trustable outcomes than CTs for predicting the habitat requirements of exotic species. Nevertheless, CTs (due to a better visualization of the outcomes for wetland managers) can be a promising tool over SVMs in order to relate the wetland characteristics to habitat preferences of the exotic species. Yet, historical data gathering would further improve the prediction accuracy of models and hereby lead to a better decision on the habitat preferences of *A. filiculoides* in the wetland. According to the results of multivariate techniques (PCA and DA), it is concluded that combination of these ordination techniques with data-driven ones might yield better outcomes to decide on the most important variables deriving the habitat preferences of exotic ferns in wetlands. The results of data-driven and multivariate techniques suggest that for the future monitoring, one has to take the both physical-habitat (e.g. depth and humidity) and water quality variables (e.g. orthophosphate) into account while focusing more on the water quality characteristics of the wetland since most of the physical-habitat variables are unmanageable.

Acknowledgements We would like to thank Guilan Environmental Protection Bureau and Selkeh Wildlife Refuge centre for giving opportunity to take samples in the field.

References

- Abbasi FM, Brar DS, Carpena AL, Fukui K, Khush GS (1999) Detection of autosyndetic and allosyndetic pairing among A and E genomes of *Oryza* through genomic in situ hybridization. *Rice Genet Newsl* 16:24–25
- Ahmad SS, Reshi ZA, Shah MA, Rashid I, Ara R, Andrabi SMA (2016) Heavy metal accumulation in the leaves of *Potamogeton natans* and *Ceratophyllum demersum* in a Himalayan RAMSAR site: management implications. *Wetl Ecol Manag* 24(4):469–475
- Akkermans W, Verdonschot PFM, Nijboer RC, Goedhart PW, Braak CJF (2004) Predicting macro-fauna community types from environmental variables by means of support vector machines. In: Lek S, Scardi M, Verdonschot PFM, Descy JP, Park YS (eds) Modelling community structure in freshwater ecosystems. Springer, Berlin, p 518
- Ambelu A, Lock K, Goethals PLM (2010) Comparison of modelling techniques to predict macroinvertebrate community composition in rivers of Ethiopia. *Ecol Inform* 5:147–152
- APHA/AWWA/WEF (1998) Standard methods for the examination of water and wastewater, 19th ed. Washington
- Barreto R, Charudattan A, Pomella A, Hanada R (2000) Biological control of neotropical aquatic weeds with fungi. *Crop Prot* 19:697–703
- Biswas M, Parveen S, Shimozawa H, Nakagoshi N (2005) Effects of *Azolla* species on weed emergence in a rice paddy ecosystem. *Weed Biol Manag* 5:176–183
- Bocchi S, Malgioglio A (2010) *Azolla-Anabaena* as a biofertilizer for rice paddy fields in the Po Valley, a temperate rice area in northern Italy. *Int J Agron*. doi:10.1155/2010/152158
- Boivin J, Ng S (2006) Are more data always better for factor analysis? *J Econ* 132:169–194
- Buckland ST, Borchers DL, Johnston A, Henrys PA, Marques TA (2007) Line transect methods for plant surveys. *Biometrics* 63:989–998
- Cohen J (1960) A coefficient of agreement for nominal scales. *Educ Psychol Meas* 20:37–46
- Dakou E, D'heygere T, Dedecker A, D'heygere T, Goethals PLM, De Pauw N, Lazaridou-Dimitriadou M (2007) Decision tree models for prediction of macroinvertebrate taxa in the river Axios Northern Greece. *Aquat Ecol* 41:399–411
- Davis JC (1986) Statistics and data analysis in geology. Wiley, New York
- De'ath G, Fabricius KE (2000) Classification and regression trees: a powerful yet simple technique for ecological data analysis. *Ecology* 81:3178–3192
- Delnavaz B, Ataei A (2009) Alien and exotic *Azolla* in northern Iran. *Afr J Biotechnol* 8:187–190
- Devin S, Beisel JN (2007) Biological and ecological characteristics of invasive species, a gammarid study. *Biol Invasions* 9:13–24
- Dzeroski S, Demsar D, Grbovic J (2000) Predicting chemical parameters of river water quality from bioindicator data. *Appl Intell* 13:7–17
- Elshorbagy A, El-Baroudy I (2009) Investigating the capabilities of evolutionary data-driven techniques using the challenging estimation of soil moisture content. *J Hydroinf* 11:237–251
- Everaert G, Boets P, Lock K, Džeroski S, Goethals PLM (2011) Using classification trees to analyze the ecological impact of invasive species in polder lakes in Flanders, Belgium. *Ecol Model* 222:2202–2212
- Fernández-Zamudio R, García-Murillo P, Cirujano S (2010) Germination characteristics and sporeling success of *A. filiculoides* Lam., an aquatic invasive fern, in a Mediterranean temporary wetland. *Aquat Bot* 93:89–92
- Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. *Environ Conserv* 24:38–49
- Fukunaga K (1990) Introduction to statistical pattern recognition. Academic Press, New York
- Giustolisi O, Savic DA (2009) Advances in data-driven analyses and modelling using EPR-MOGA. *J Hydroinf* 11:225–236
- Goethals PLM (2005) Data driven development of predictive ecological models for benthic macroinvertebrates in rivers. PhD thesis. University of Ghent, p 377
- Goethals PLM, De Pauw N (2001) Development of a concept for integrated river assessment in Flanders, Belgium. *J Limnol* 60(1):7–16
- Goethals PLM, Dedecker AP, Gabriels W, Lek S, De Pauw N (2007) Applications of artificial neural networks predicting macroinvertebrates in freshwaters. *Aquat Ecol* 41:491–508
- Gooyong L, Sangeun L, Heekyung P (2014) Improving applicability of neuro-genetic algorithm to predict short-term water level: a case study. *J Hydroinf*. doi:10.2166/hydro.2013.011
- Haghi Vayghan A, Zarkami R, Sadeghi R, Fazli H (2015) Modelling habitat preferences of Caspian kutum, *Rutilus frisii kutum* (Kamensky, 1901) (Actinopterygii, Cypriniformes) in the Caspian Sea. *Hydrobiologia*. doi:10.1007/s10750-015-2446-3
- Hammer Ø (2013) Paleontological statistics (PAST). Natural History Museum, University of Oslo, Oslo, p 221
- Harper DAT (1999) Numerical palaeobiology. Wiley, New York
- Hill MP (2003) The impact and control of alien aquatic vegetation in South African aquatic ecosystems. *Afr J Aquat Sci* 28:19–24
- Hoang TH, Lock K, Mouton A, Goethals PLM (2010) Application of classification trees and support vector machines to model the presence of macroinvertebrates in rivers in Vietnam. *Ecol Inform* 5:140–146
- Janes R (1998) Growth and survival of *A. filiculoides* in Britain. 1. Vegetative reproduction. *New Phytol* 138:367–376
- Japan International Cooperation Agency (JICA) (2005) The study on integrated management of the Anzali Wetland in the Islamic Republic of Iran-final report, vol 2. p 222
- Kannaiyan S, Somporn C (1989) Effect of high temperature on growth, nitrogen fixation, and chlorophyll content of five species of *Azolla-Anabaena* symbiosis. *Biol Fertil Soils* 7:168–172
- Karatayev AY, Burlakova LE, Padilla DK, Mastitsky SE, Olein S (2009) Invaders are not a random selection of species. *Biol Invasions* 11:2009–2019
- Keerthi SS, Shevade SK, Bhattacharyya C, Murthy KRK (2001) Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Comput* 13:637–649

- Landis JR, Koch GG (1977) The measurement of observer agreement for categorical data. *Biometrics* 33:159–174
- Liu X, Min C, Xia-shi L, Chungchu L (2008) Research on some functions of *Azolla* in CELSS system. *Acta Astronaut* 63:1061–1066
- Lumpkin TA, Plucknett DL (1982) *Azolla* as a green manure, use and management in crop production. Westview tropical agriculture, vol 5. Westview Press, Boulder
- Mansoori J (1995) Islamic Republic of Iran. In: Scott DA (ed) A directory of wetlands in the Middle East. IUCN, Slimbridge
- Olenin S, Minchin D, Daunys D (2007) Assessment of biopollution in aquatic ecosystems. *Mar Pollut Bull* 55:379–394
- Parsons M, Thoms MC, Horris RH (2004) Development of a standard approach to river habitat assessment in Australia. *Environ Monit Assess* 98:109–130
- Platt J (1998) Fast training of support vector machines using sequential minimal optimization. In: Schoelkopf B, Burges C, Smola A (eds) *Advances in Kernel methods—support vector learning*. MIT Press, Cambridge
- Quinlan JR (1993) C4.5, program for machine learning. Morgan Kaufmann Publishers, San Francisco, p 302
- Ricciardi A, MacIsaac HJ (2000) Recent mass invasion of the North American Great Lakes by Ponto-Caspian species. *Trends Ecol Evol* 15:62–65
- Robert JA (2003) Neural network rainfall-runoff forecasting based on continuous resampling. *J Hydroinf* 5:51–61
- Sadeghi R, Zarkami R, Sabetraftar K, Van Damme P (2012) Application of classification trees to model the distribution pattern of a new exotic species *Azolla filiculoides* (Lam.) in Selkeh Wildlife Refuge, Anzali wetland, Iran. *Ecol Model* 243:8–17
- Savic DA, Giustolisi O, Laucelli D (2009) Asset deterioration analysis using multi-utility data and multi-objective data. *J Hydroinf* 11:211–224
- Serag MS, El-Hakeem A, Badway M, Mousa MA (2000) On the ecology of *A. filiculoides* Lam. in Damietta District, Egypt. *Limnologica* 30:73–81
- Summers WH, Archibold OW (2007) Exotic plant species in the southern boreal forest of Saskatchewan. *For Ecol Manag* 251:156–163
- Sweet AR, Hills LV (1971) A study of *A. pinnata* R. brown. *Am Fern J* 71:1–14
- Taghi-Ganji M, Khosravi M, Rakhshae R (2005) Biosorption of Pb (²I), Cd (²I), Cu (²I) and Zn (II) from the wastewater by treated *A. filiculoides* with H₂O₂/MgCl₂. *Int J Environ Sci Technol* 1:265–271
- Van Der Heide T, Roijackers RMM, Peeters ETHM, Van Nes EH (2006) Experiments with duckweed–moth systems suggest that global warming may reduce rather than promote herbivory. *Freshw Biol* 51:110–116
- Van Hove C, Lejeune A (2002) The *Azolla–Anabaena* symbiosis. *Biol Environ* 102:23–26
- Vander Zanden MJ, Olden JD (2008) A management framework for preventing the secondary spread of aquatic invasive species. *Can J Fish Aquat Sci* 65:1512–1522
- Vapnik V (1995) *The nature of statistical learning theory*. Springer, New York, p 187
- Vitousek PM, D’Antonio CM, Loope LL, Westbrooks R (1996) Biological invasions as global environmental change. *Am Sci* 84:468–478
- Watanabe I, Espinase CR (1976) Potential of nitrogen fixing *Azolla–Anabaena* complex as fertilizer in paddy soil. IIRRI Saturday seminar
- Witten IH, Frank E, Hall MA (2011) *Data mining, practical machine learning tools and techniques*, 3rd edn. Morgan Kaufmann, San Francisco, p 629
- Zarkami R (2011) Application of classification trees-J48 to model the presence of roach (*Rutilus rutilus*) in rivers. *CJES* 9:189–198
- Zarkami R, Goethals PLM, De Pauw N (2010) Use of classification tree methods to study the habitat requirements of tench *Tinca tinca*. L., 1758. *CJES* 8:55–63
- Zarkami R, Sadeghi R, Goethals PLM (2012) Use of fish distribution modelling for river management. *Ecol Model* 230:44–49
- Zarkami R, Sadeghi R, Goethals PLM (2014) Modelling occurrence of roach “*Rutilus rutilus*” in streams. *Aquat Ecol* 48:161–177
- Zhengfu R, Fernando A (2007) Use of an artificial neural network to capture the domain knowledge of a conventional hydraulic simulation model. *J Hydroinf* 9(1):15–24
- Zimmerman WJ (1985) Biomass and pigment production in three isolates of *Azolla*: II. Response to light and temperature stress. *Ann Bot Lond* 56:701–709