**UNIVERSIDAD DE LA REPÚBLICA**
**FACULTAD DE AGRONOMÍA**

# MODELOS ADITIVOS GENERALIZADOS APLICADOS A LA PREDICCIÓN DE VEGETACIÓN ACUÁTICA SUMERGIDA EN UN HUMEDAL CONSTRUIDO

**por**

**Andrea Lucía GARAY DELBONO**

TESIS presentada como uno de los requisitos para obtener el título de *Magister* en Ciencias Agrarias opción Bioestadística

MONTEVIDEO
URUGUAY
setiembre 2016

Tesis aprobada por el tribunal integrado por Ing Agr. (Ms.C.) Alejandra Borges, Lic. Biol. (Ph.D.) Carolina Crisci, e Ing. Agr. (Ph.D.) Mónica Cadenazzi, el 22 de setiembre de 2016. Autora: Lic. Biol. Andrea Garay. Directora Ing. Agr. (Ph.D.) Lucía Gutiérrez.

## AGRADECIMIENTOS

Quiero agradecer a todos los que participaron de una u otra manera en hacer esto realidad. A mi tutora, Lucía Gutiérrez, quien me enseñó mucho en este proceso, tanto a nivel académico como personal, gracias por tu tiempo y dedicación. Al gran grupo humano del Departamento de Biometría, Estadística y Computación (DBEC) de Facultad de Agronomía, en especial a Nati, Ale, Inés, Bettina, Pablo y Virginia, gracias por las charlas, y los consejos. A Carolina Crisci, Alejandra Borges y Mónica Cadenazzi por sus aportes como tribunal de este trabajo. Al DBEC que nos da la valiosa posibilidad de continuar nuestra formación en el departamento. A mis padres y mi hermana Lety qué están siempre para apoyar, para sumar y para dar los mejores consejos. A mis abuelos que siempre están presentes, y a mi hermosa familia que me reconforta y llena el alma. A Gabriel por hacerme el aguante en todas las etapas, por impulsarme, por aconsejarme, por querer lo mejor para mí,...feliz de compartir mi vida contigo.

**TABLA DE CONTENIDO**

Página

# RESUMEN

Los Modelos Aditivos Generalizado (GAM) son extensiones semi-paramétricas de los modelos lineales generalizados, en donde el predictor lineal no es simplemente una combinación lineal de las variables explicativas, sino que es una combinación lineal de funciones suavizadas de las variables explicativas. Los GAM pueden ser utilizados para predecir la distribución y abundancia de la vegetación acuática. Las macrófitas son un componente clave en los humedales artificiales construidos para mejorar la calidad del agua. El objetivo de este trabajo fue aplicar GAM para predecir la abundancia de la vegetación acuática sumergida en un humedal construido. Se estudió el efecto de variables topográficas y el día del año sobre la abundancia de la vegetación en tres años de estudio. Para evitar la sobredispersión de la variable abundancia debida a la gran proporción de ceros, se trabajó primero con los modelos de ocurrencia (presencia-ausencia) y abundancia dado que la especie está presente por separado. Luego, para combinar las predicciones de ambos modelos, se comparó: (1) el uso de umbrales de ocurrencia para convertir las probabilidades predichas de ocurrencia a datos de presencia/ausencia utilizando la predicción de abundancia condicional para los sitios donde se predijo la presencia de la especie, con (2) el producto de las predicciones de los modelos de ocurrencia y de abundancia condicional. Ambas aproximaciones fueron comparadas basándose en la suma de cuadrados residual. Cada modelo ajustado se utilizó para predecir dentro de cada año y para años desconocidos. Las predicciones se evaluaron mediante validación cruzada, a partir del área bajo la curva para los modelos de ocurrencia, y el coeficiente de correlación para abundancia condicional. Se obtuvieron buenas predicciones para predecir dentro de un mismo año. Sin embargo, los modelos no fueron muy buenos para predecir años desconocidos. En este trabajo fue posible realizar predicciones de vegetación acuática sumergida y evaluar variables que afectan su distribución y abundancia.

**Palabras clave**: modelos aditivos generalizados, ocurrencia, abundancia, predicción conjunta

# GENERALIZED ADDITIVE MODELS APPLIED TO PREDICT ABUNDANCE AND OCCURRENCE OF SUBMERGED AQUATIC VEGETATION IN A CONSTRUCTED WETLAND

## SUMMARY

The Generalized Additive Models (GAM) are semi-parametric extensions of Generalized Linear Models, where the linear predictor is not simply a linear combination of the explanatory variables, but is a linear combination of smooth functions of the explanatory variables. GAM can be used to predict macrophytes distribution and abundance. Macrophytes are key components in artificial wetlands constructed to improve water quality. The aim of this study was to apply GAM to predict the abundance of submerged aquatic vegetation in a constructed wetland. The effect of topographical variables and day of the year on the vegetation abundance during three years was studied. To avoid overdispersion of the abundance variable due to de high proportion of zeros, we initially worked with the occurrence (presence or absence) and the abundance given the specie is present models separately. Later, to combine both model predictions, we compared: (1) the use of occurrence thresholds to convert the predicted probabilities of occurrence to presence/absence data using the prediction of conditional abundance for the sites where a presence was predicted, to (2) the method of multiplying the predictions by the conditional abundance predictions. Both approaches were compared based on the sum squared residuals. Each adjusted model was used to predict within a year and to predict unknown years. Predictions were evaluated with cross-validation, estimating the area under the curve for occurrence models, and the correlation coefficient for conditional abundance models. We obtained good predictions to predict within a year. However, models were not as good in making predictions for unknown years. In this work it was possible to make good predictions of submerged aquatic vegetation and to identify the variables that affect their distribution and abundance.

**Keywords:** generalized additive models, occurrence, abundance, joint prediction

# 1. INTRODUCCIÓN

## 1.1. LOS HUMEDALES ARTIFICIALES Y LA VEGETACIÓN ACUÁTICA

Los humedales son áreas inundadas de forma temporal o permanente, funcionan como sumidero de nutrientes y zonas de amortiguación (Greenway, 2007). Los humedales construidos son diseñados para simular los procesos que naturalmente tienen lugar en los humedales. En este sentido, los humedales artificiales permiten mejorar la calidad del agua (Vymazal *et al.,* 2006). Un factor de suma importancia para incrementar la calidad del agua es la interacción entre los componentes bióticos y abióticos de estos sistemas, favoreciendo procesos de remoción, reciclaje, o acumulación de contaminantes y nutrientes a través de procesos físicos, químicos, y biológicos (Reddy y D'Angelo, 1997; Mitsch y Gosselink, 2000; Wetzel, 2001; Williams, 2002). Específicamente, la vegetación acuática es un componente clave en la remoción de nutrientes del agua. Los tallos y hojas de las plantas reducen la velocidad y turbulencia del agua, causando filtración y sedimentación de las partículas orgánicas e inorgánicas, y remoción de nutrientes. Además, las macrófitas proporcionan una superficie para el desarrollo de perifiton y microorganismos que colaboran en el consumo de los nutrientes presentes en el agua (Greenway, 2007). La comprensión de la distribución y abundancia de la vegetación acuática ayuda a comprender el funcionamiento de los humedales construidos y mejorar el diseño de dichos sistemas. Sin embargo, la modelación de la vegetación no es un fenómeno trivial.

## 1.2. MODELOS ADITIVOS GENERALIZADOS

Los modelos de regresión han sido ampliamente utilizados en ecología para predecir la vegetación de diversos ecosistemas. Sin embrago, los análisis más clásicos muchas veces no se ajustan a la naturaleza de los datos en ecología. Los modelos lineales generalizados (GLM) incluyen características de los modelos lineales, pero permiten trabajar bajo distribuciones no normales y con estructuras de varianza no constante (Hastie y Tibshirani, 1990). En los GLM se asume una relación llamada función de enlace entre la media de la variable de respuesta y la combinación lineal de las variables explicativas. Por otro lado, los modelos aditivos generalizados (GAM) son extensiones semi-paramétricas de los GLM, y presentan una mayor flexibilidad para modelar diferentes estructuras de datos. Los GAM proveen un marco general para extender los modelos lineales permitiendo funciones no lineales de las variables explicativas, lo cual permite potencialmente hacer predicciones más precisas y más realistas para la variable de respuesta. En lugar de utilizar una relación a priori entre los parámetros, son modelos dirigidos por los datos, ya que se basan en los propios datos para especificar la forma que tiene el modelo (Yee y Mitchell, 1991). La función de enlace en estos modelos establece una relación entre la media de la variable de respuesta y una función suavizada de la variable explicativa (Hastie y Tibshirani, 1986, 1990). Los GAM se pueden escribir de la siguiente manera:

$$Y_i = \beta_0 + \sum_{j=1}^{p} f(x_{ij}) + \varepsilon_i$$

$$= \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip}) + \varepsilon_i$$

$$g(E(Y)) = \beta_0 + f_1(x_{i1}) + f_2(x_{i2}) + \cdots + f_p(x_{ip})$$

donde $Y$ es la variable de respuesta con una distribución perteneciente a la familia exponencial, $\beta_0$ el intercepto, $f(x_{ij})$ la función suviazada de las variables explicativas y $\varepsilon_i$ los residuales. Son modelos aditivos porque las funciones para cada variable explicativa se estiman de forma separada, y luego se suman todas sus contribuciones. Esto permite examinar el efecto de cada variable de respuesta sobre la variable de respuesta de forma individual mientras que se mantienen las otras variables fijas. Los GAM proporcionan un compromiso de gran utilidad entre los modelos lineales y los totalmente no paramétricos.

## 1.3. MODELADO DE VEGETACIÓN ACUÁTICA

La variable abundancia de vegetación usualmente presenta una gran proporción de ceros (Barry y Welsh, 2002). La presencia de una gran proporción de ceros es muy común en datos de campo, y es una forma particular de sobre-dispersión (McCullagh y Nelder, 1989). Para hacer frente a este problema se han desarrollado diversos enfoques como ser: el uso de distribuciones delta lognormal (Aitchison y Brown, 1957; Pennington, 1983), métodos delta de aproximación de la varianza (Stefánsson, 1996), la distribución binomial negativa (Warton, 2005), y las distribuciones cero-inflados (Welsh *et al.*, 1996). Welsh *et al.* (1996) modelaron la ocurrencia y la abundancia dado que la especie está presente por separado para hacer frente a la gran proporción de ceros. Luego, Barry y Welsh (2002) propusieron combinar las predicciones de los modelos de ocurrencia y abundancia condicional a través del producto de las predicciones. Por otro lado, uno de los temas que aún no está resuelto en los modelos de ocurrencia es cuál es la mejor estrategia para seleccionar un umbral de ocurrencia para transformar las predicciones de ocurrencia en datos de presencia/ausencia. Existen muchas aproximaciones para determinar umbrales, por ejemplo métodos arbitrarios en donde un valor fijo se establece como umbral, o métodos que buscan maximizar la concordancia entre los valores observados y predichos para selección de un umbral (Liu *et al.,* 2005; Jiménez-Valverde y Lobo, 2007). Los GAM son una herramienta eficaz para predecir la abundancia de vegetación, y es posible aplicar la partición de los datos de abundancia en ocurrencia y abundancia condicional para hacer frente a la gran proporción de ceros de estos datos.

## 1.4. OBJETIVOS DE INVESTIGACIÓN

El objetivo principal de este trabajo es predecir la abundancia de la vegetación acuática sumergida en un humedal construido en una cuenca agrícola intensiva utilizando modelos aditivos generalizados.

Como objetivos específicos, se buscó:
a) implementar modelos aditivos generalizados para modelar vegetación acuática con gran proporción de ceros.
b) comparar estrategias para incorporar predicciones de ocurrencia y de abundancia condicional para predecir la abundancia de la vegetación.
c) evaluar el desempeño de las diferentes estrategias en base a ajuste y capacidad predictiva de los modelos
d) predecir vegetación acuática sumergida dentro de años y para años desconocidos
Esta tesis consta de un capítulo de introducción, un artículo de investigación, y los resultados, discusión y conclusiones generales del trabajo de investigación. Se planifica enviar el artículo a la revista Ecological Modelling (http://www.journals.elsevier.com/ecological-modelling/).

# 2. GENERALIZED ADDITIVE MODELS APPLIED TO PREDICT ABUNDANCE AND OCCURRENCE OF SUBMERGED AQUATIC VEGETATION IN A CONSTRUCTED WETLAND

## 2.1. SUMMARY

The Generalized Additive Models (GAM) are semi-parametric extensions of Generalized Linear Models, where the linear predictor is not simply a linear combination of the explanatory variables, but is a linear combination of smooth functions of the explanatory variables. GAM can be used to predict macrophytes distribution and abundance. Macrophytes are key components in artificial wetlands constructed to improve water quality. The aim of this study was to apply GAM to predict the abundance of submerged aquatic vegetation in a constructed wetland. The effect of topographical variables and day of the year on the vegetation abundance during three years was studied. To avoid overdispersion of the abundance variable (due to de high proportion of zeros) we initially worked with the occurrence (presence or absence) and the conditional abundance (abundance data when there is the specie is present) separately. Later, to combine both model predictions, we compared: (1) the use of occurrence thresholds to convert the predicted probabilities of occurrence to presence/absence data using the prediction of conditional abundance for the sites where a presence was predicted, and (2) the method of multiplying the predictions of ocurrence by the conditional abundance predictions. Both approaches were compared based on the sum squared residuals. Each adjusted model was used to predict within a year and to predict unknown years. Predictions were evaluated with cross-validation, estimating the area under the curve for occurrence models, and the correlation coefficient for conditional abundance models. We obtained good predictions to predict within a year. However, models were not as good in making predictions for unknown years. In this work it was possible to make good predictions of submerged aquatic vegetation and to identify the variables that affect their distribution and abundance.

**Keywords:** generalized additive models, occurrence, abundance, joint prediction

## 2.2. RESUMEN

Los Modelos Aditivos Generalizado (GAM) son extensiones semi-paramétricas de los modelos lineales generalizados, en donde el predictor lineal no es simplemente una combinación lineal de las variables explicativas, sino que es una combinación lineal de funciones suavizadas de las variables explicativas. Los GAM pueden ser utilizados para predecir la distribución y abundancia de la vegetación acuática. Las macrófitas son un componente clave en los humedales artificiales construidos para mejorar la calidad del agua. El objetivo de este trabajo fue aplicar GAM para predecir la abundancia de la vegetación acuática sumergida en un humedal construido. Se estudió el efecto de variables topográficas y el día del año sobre la abundancia de la vegetación en tres años de estudio. Para evitar la sobredispersión de la variable abundancia debida a la gran proporción de ceros, se trabajó primero con los modelos de ocurrencia (presencia-ausencia) y abundancia dado que la especie está presente por separado. Luego, para combinar las predicciones de ambos modelos, se comparó: (1) el uso de umbrales de ocurrencia para convertir las probabilidades predichas de ocurrencia a datos de presencia/ausencia utilizando la predicción de abundancia condicional para los sitios donde se predijo la presencia de la especie, con (2) el producto de las predicciones de los modelos de ocurrencia y de abundancia condicional. Ambas aproximaciones fueron comparadas basándose en la suma de cuadrados residual. Cada modelo ajustado se utilizó para predecir dentro de cada año y para años desconocidos. Las predicciones se evaluaron mediante validación cruzada, a partir del área bajo la curva para los modelos de ocurrencia, y el coeficiente de correlación para abundancia condicional. Se obtuvieron buenas predicciones para predecir dentro de un mismo año. Sin embargo, los modelos no fueron muy buenos para predecir años desconocidos. En este trabajo fue posible realizar predicciones de vegetación acuática sumergida y evaluar variables que afectan su distribución y abundancia.

**Palabras clave**: modelos aditivos generalizados, ocurrencia, abundancia, predicción conjunta

## 2.3. INTRODUCTION

### 2.3.1. CONSTRUCTED WETLANDS AND AQUATIC VEGETATION

Wetlands are recognized for their functions as a sink of nutrients and buffer zones (Greenway, 2007). Constructed wetlands are engineering systems designed to simulate the processes that occurs in natural wetlands, having a more controlled environment to improve water quality (Vymazal *et al.*, 2006). The interaction between biotic and abiotic components of wetlands are vital to achieve an increase in water quality due to the removal, recycling, or accumulation of contaminants and nutrients by physical, chemical, and biological processes (Reddy and D'Angelo, 1997; Mitsch and Gosselink, 2000; Wetzel, 2001; Williams 2002). Specifically, macrophytes are key components in removing nutrients because the stems and leaves of plants reduce the water speed and turbulence, causing filtration and sedimentation of organic and inorganic particles, and therefore removing nutrients. Furthermore, macrophytes provide a surface for the development of periphyton and microorganisms that in turn remove nutrients (Greenway, 2007). Understanding the distribution and abundance of aquatic vegetation will allow us to comprehend constructed wetlands and improve wetland design. However, vegetation modelling is not a trivial phenomenon.

### 2.3.2. GENERALIZED ADDITIVE MODELS

Regression analyzes have been widely used in ecology to predict vegetation in diverse ecosystems. Generalized Linear Models (GLMs; McCullagh and Nelder, 1989) include characteristics of linear models but allow working under non-normal distributions with non-linear models and with non-constant variance structures (Hastie and Tibshirani, 1990). A link function between the mean of the response variable and the linear combination of the explanatory variables is therefore assumed. Generalized Additive Models (GAMs; Hastie and Tibshirani, 1990) are semi-parametric extensions of the GLMs, performing the parameterization process similar to GLMs, but allowing predictors to be modelled as either parametric or non-parametric (Guisan *et al.*, 2002).

Instead of using an *a priori* relationship between explanatory variables, GAMs are data-driven models where a smoothing function is estimated from the data (Yee and Mitchell, 1991). The link function of GAM establishes a relationship between the mean of the response variable and the smooth function of the explanatory variable (Hastie and Tibshirani 1986, 1990). Therefore, GAM have more flexibility for modeling different data structure like bi-modal, skewed or zero-inflated distributions. These models can be written as:

$$Y_i = \beta_0 + \sum_{j=1}^{p} f(x_{ij}) + \varepsilon_i$$

where $Y_i$ is the response variable with an exponential family distribution, $\beta_0$ the intercept, $\sum_{j=1}^{p} f(x_{ij})$ the sum of the smoothed functions of the j-th explanatory variables and $\varepsilon_i$ the residuals.

### 2.3.3. MODELING AQUATIC VEGETATION

Abundance data have usually a large proportion of zeros (Barry and Welsh, 2002). The problem of extra zeros is very common in field data, and is a particular form of over-dispersion (McCullagh and Nelder, 1989). To deal with this problem many approaches have been developed; the use of lognormal delta distributions (Aitchison and Brown, 1957; Pennington, 1983), delta method approximation of variance (Stefánsson, 1996), negative binomial distribution (Warton, 2005), and zero inflated distributions (Welsh *et al.*, 1996). Specifically, Welsh *et al.* (1996) modeled separately occurrence and abundance given the species is present. Later, Barry and Welsh (2002) proposed a method to combine occurrence and conditional abundance predictions by using their product. One of the issues that still remains largely unsolved is the best strategy to select an occurrence threshold to transform the probabilities predictions into presence/absence data. There are many approaches to determining thresholds, for example arbitrary fixed values such as a threshold of 0.5, or methods that maximize the agreement between observed and predicting distribution for choosing the threshold value (Liu *et al*., 2005; Jiménez-Valverde and Lobo, 2007). However, when combining occurrence and conditional abundance to estimate total abundance, methods have disregarded the thresholding approach. The objective of this work was to extend generalized additive models for predicting total abundance of submerged aquatic vegetation species in a constructed wetland, to determine the most relevant explanatory variables to predict total abundance and to evaluate its predictive ability by modelling both occurrence and conditional abundance in a single model evaluating thresholding alternative.

## 2.4. MATERIALS AND METHODS

### 2.4.1. Study area and sampling strategy

The study area was a constructed wetland in an intensive agricultural watershed in the State of Iowa, USA (4667000 Northing, 442700 Easting). The mean area is 300 m$^2$ and the mean depth is 0.54 m. Aquatic vegetation of the wetland was implanted, and the predominant submerged aquatic vegetation species are: *Potamogenton illinoensis*, *Ceratophyllum demersum*, *Chara* sp., and *Potamogeton natans*. Repeated surveys over years and dates were conducted surveying four times in 2010, two in 2011, and three in 2012 (Table 2). At each survey, we set-up approximately 30 transects perpendicular to the direction of the water flow. Within each transect, several sampling locations were established, conforming a data set of 987 sampling points. At each station, a 20 by 20 cm quadrat was established and submerged aquatic vegetation (SAV) cover percentage and water depth were recorded. Aquatic vegetation was modeled only for species presented in at least 10% of total stations.

### 2.4.2. Calculated variables

Terrain, curvature, and relative position variables such as aspect, ruggedness, littoral slope, topographic index and roughness were calculated from water depth data using

R software (R Development Core Team 2014, Table 1). The survey day was calculated as a Julian day independently for each year.

**Table 1.** Calculated variables as predictors.

| Variable | Description |
|---|---|
| Littoral slope | It is the maximum rate of change of slope between each cell and its neighbours. |
| Aspect | Orientation of the slopes measure by Indices of northness and eastness. |
| Roughness | Measure of terrain complexity that represents the ratio of surface area to planar area. |
| Ruggedness | Terrain complexity grid which combines variation in slope and aspect into a single measure. |

### **2.4.3. Data analysis**

We adjusted a GAM for occurrence and conditional abundance separately following Welsh *et al.* (1996). Later, we combined the predictions from occurrence and conditional abundance to obtain total abundance predictions.

The predictor variables used in the models were: topographic variables (water depth, aspect, ruggedness, littoral slope, topographic index and roughness) included as smooth terms using penalized splines, and Julian days used as a linear term. For the smooth functions we used thin plate regression splines and the degree of smoothness of the smooth function was estimated by generalized cross-validation for each case. Analyses were conducted in R project (R Development Core Team, 2014) and GAMs were adjusted using *mgcv* library (Wood, 2006).

#### 2.4.3.1. Modeling occurrence

To model species occurrence (presence or absence), a GAM with a binomial distribution and logit link function was used. The adjusted GAM was: $Y_i = \alpha + \sum_{j=1}^{n} f(X_{ij}) + \varepsilon_i$, where $Y_i = logit(p) = \log\left(\frac{p}{1-p}\right)$ and p is the probability of a species being present, $\alpha$ is the intercept, and $\sum_{j=1}^{n} f(X_{ij})$ the sum of smoothed functions of the explanatory variables. Because our final goal was to model total abundance, we used the same set of explanatory variables for occurrence and conditional abundance. Therefore, variable selection was performed following a forward selection process based on the sum of the Akaike Information Criterion (AIC) values from the occurrence and conditional abundance models (Akaike, 1973). The AIC takes into account the complexity and goodness of fit of the model.

Model predictions were evaluated with a 10-fold cross-validation. Data were randomly divided into training (90%) and testing (10%) sets. The training set was used to adjust the models and to make predictions for the test set. This process was simulated 1000 times. The ability of the occurrence models to discriminate among sampling stations with and without species was compared using the area under the receiver operating characteristic curve (ROC). Briefly, the ROC is a graphical representation of the sensitivity as a function of (1-specificity). The area under the ROC (AUC) is an indicator of the predictive power of the model. If the AUC is less than 0.5 it indicates that the model ranks poorly, and if it is equal to 1 it indicates a perfect classification. In this case a perfect classification means that the model predicts the presence when the species is really present and absences when it is not present. The AUC value was calculated for each iteration and a confidence interval was constructed using the 0.025 and 0.975 quantile of the empirical distribution of the AUC.

#### 2.4.3.2. Modeling conditional abundance

To model the abundance given the specie is present, the same general procedure as above was followed. However, a Gaussian distribution with a log link function was used instead. The adjusted GAM was: $Y_i = \alpha + \sum_{j=1}^{n} f\left(X_{ij}\right) + \varepsilon_i$, where $Y_i$ is the percent of cover of submerged aquatic vegetation at the sampling quadrat, $\alpha$ is the intercept, and $\sum_{j=1}^{n} f\left(X_{ij}\right)$ the sum of smoothed functions of the explanatory variables. Explanatory variables were again selected with the forward approach using the sum of AIC values from occurrence and conditional abundance models. To assess the predictive ability of the model we estimated the correlation coefficient ($r_{(Y_i,\widehat{Y_i})}$) between observed and predicted conditional abundance. The correlation coefficient was estimated for each iteration and a confidence interval was constructed using the 0.025 and 0.975 quantile of the empirical distribution of the coefficient of correlation.

#### 2.4.3.3. Modeling total abundance

To provide a unique set of predictions for total abundance generated from the selected models of occurrence and conditional abundance, two approaches were proposed. As a first approach, predictions of total abundance were estimated as the product of the estimated occurrence probability and conditional abundance predictions. Therefore, if the prediction of occurrence is zero, total abundance is zero. However, if the prediction of occurrence is different of zero total abundance is weighted by the probability of occurrence. The second approach for estimating the predictions of total abundance use different thresholds of occurrence. We evaluated a sequence of thresholds between 0 and 1 to determine the best threshold. If the prediction of occurrence is less than the threshold value, total abundance is zero. Otherwise, total abundance takes the value of the predicted conditional abundance. The Sum of Square Residuals (SSR) estimated as the sum of squares of the difference between observed abundance and predicted values was used to evaluate model performance. SSR was estimated for the product of predictions (first approach) and for each threshold between zero and one for the second approach. We selected the approach and the threshold with the lowest SSR.

The predictions of the best approach were evaluated based on a 10-fold cross-validation simulated 1000 times was done. We estimated the correlation coefficient ($r_{(Y_i, \widehat{Y_\imath})}$) between observed and predicted total abundance and the root mean squared error prediction estimated as: $RMSEP = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - \widehat{Y_\imath})^2}{n}}$, where $Y_i$ is the observed value and $\widehat{Y_\imath}$ is the prediction. For both indicators, a confidence interval was constructed using the 0.025 and 0.975 quantile of their empirical distribution.

#### 2.4.3.4 Data sets

A within year predictions of total abundance was obtained for each species present in at least 10% of the total sample stations, and for total submerged aquatic vegetation (SAV). SAV occurrence was estimated as the occurrence of any species at each location, while abundance was estimated as the sum of individual species abundance. Models were therefore fit for *Potamogeton* sp. in 2010, 2011, and 2012, *C. demersum* in 2010, and for SAV in 2010, 2011, and 2012. Additionally, a global model for all the years was adjusted for *Potamogeton* sp., *C. demersum*, and SAV, including year as a linear predictor. The abundance of the remaining species was evaluated in the global model as a smoothed term for *Potamogeton* sp. and *C. demersum*. Finally, models for each year were evaluated to predict unknown years, and the same cross-validation structure was used to evaluate model performance.

## 2.5. RESULTS

### 2.5.1. Sampling characterization

A total of 993 locations were registered during the three-year survey. A summary of the explanatory variables is shown in Table 2. Briefly, the mean depth was similar for all the surveyed years, and the topographic variables were of the same order, but littoral slope showed a smaller mean in 2010 than 2011 and 2012. Submerged aquatic vegetation (SAV) was observed at 62% of the registered stations in 2010, at 61% in 2011, and at 90% in 2012 (Figure 1). In 2012 not only the observed occurrence but also the mean abundance of SAV was higher than the years before. In 2010 the species *C. demersum* and *Potamogeton* sp. were dominant. *C. demersum* was observed at 34% of the total locations, while *Potamogeton* sp. at 30% of the locations in this year (Table 2). The stations in 2010 with *Potamogeton* sp. had a mean water depth of 0.58 m and a maximum of 1.37 m, and *C. demersum* a mean water depth of 0.82 m and a maximum of 1.96 m. *Potamogeton* sp. was the dominant specie in 2011 and 2012 being observed at 59% and 85% of the locations respectively (Table 2). *C. demersum* declined its observed frequency and abundance through the years, while *Potamogeton* sp. increased both occurrence and abundance throughout the years. In 2011, the stations with *Potamogeton* sp. had a mean water depth of 0.45 m and a maximum of 1.19 m, and in 2012 the mean depth of the stations with *Potamogeton* sp. was 0.54. Other SAV species recorded during the surveys were: *P. natans* (1%) and *Chara* sp. (1%) in 2010, *P. natans* (0.5%) in 2011, and *Chara* sp. (10%), *P. natans* (2%) and *C. demersum* (2%) in 2012.

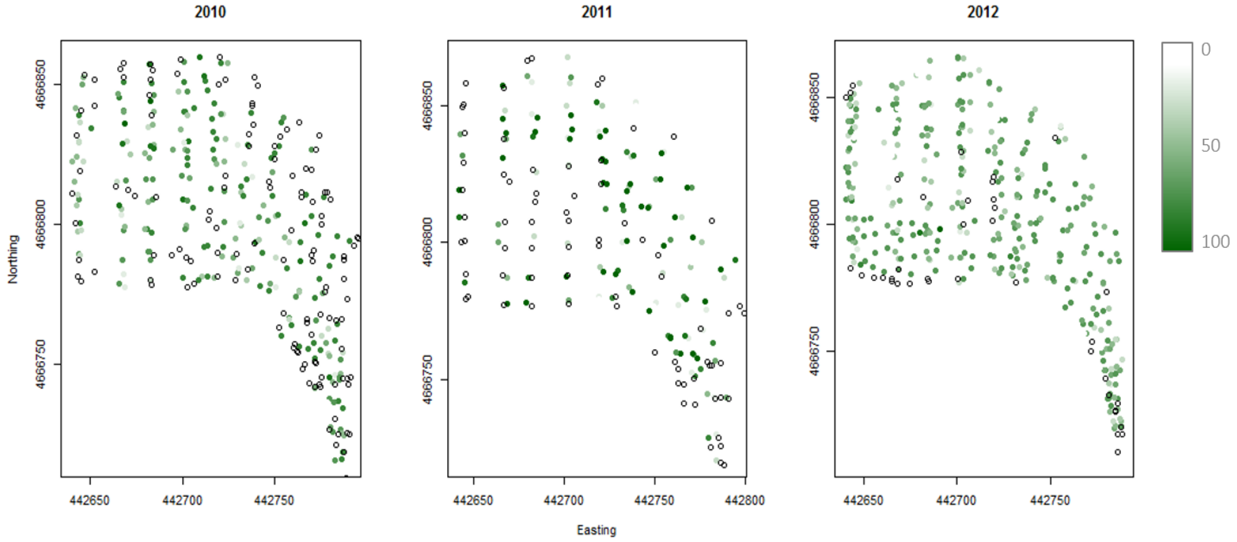

**Figure 1.** Abundance of submerged aquatic vegetation (SAV) observed in the years 2010, 20111 and 2012. Sampling locations with the observed percentage of cover of SAV is shown.

**Table 2.** Survey description: sampling characterization, topographic variables characterization, and observed species occurrence frequency and conditional abundance (mean and standard deviation between brackets).

| | 2010 | 2011 | 2012 |
|---|---|---|---|
| Surveys (julian day) | 154, 194, 243 and 283 | 141 and 182 | 67, 97 and 127 |
| # total locations | 395 | 208 | 390 |
| # location with vegetation | 243 | 127 | 351 |
| *Topographic variables* | | | |
| Depth (m) | 0.579 (0.369) | 0.520 (0.338) | 0.519 (0.307) |
| Littoral slope (degrees) | 0.0589 (0.0504) | 2.843 (2.403) | 2.506 (2.972) |
| Aspect (degrees) | 176.2 (102.7) | 173.4 (104.1) | 184.5 (111.9) |
| Ruggedness (m) | 0.047 (0.041) | 0.040 (0.033) | 0.047 (0.042) |
| Roughness (m) | 0.1527 (0.128) | 0.1310 (0.107) | 0.150 (0.128) |
| Topographic index | 0.0024 (0.0211) | - | 0.0008 (0.015) |
| *Occurrence* † | | | |
| Po | 0.30 | 0.59 | 0.85 |
| Cd | 0.34 | 0.05 | 0.02 |
| SAV | 0.62 | 0.61 | 0.90 |
| *Conditional abundance* † | | | |
| Po | 58.98 (32.26) | 59.05 (34.59) | 72.58 ( 30.64) |
| Cd | 50.07 (26.24) | 31.14 (36.68) | 6.94 (6.82) |
| SAV | 59.12 (28.07) | 59.61 (34.45) | 76.4 (29.4) |

† Cd: *C. demersum*, Po: *Potamogeton* sp., SAV: total submerged aquatic vegetation.

### 2.5.2. Occurrence and conditional abundance models

The selected models for occurrence and conditional abundances for the two species and total SAV in each year included sampling day as a linear term, and depth, day by depth, and the topographic variables aspect or ruggedness as smoothing terms (Table 3, Figure 2 and 3). These models had high predictive values, with AUC values larger than 0.5, and large correlation values for abundance prediction. The smallest predictive values were obtained for *Potamogeton* sp. in 2012 (Table 3). The best approach for modelling occurrence and conditional abundance together was the product of the predictions of occurrence and conditional abundance, as is shown for *C. demersum* in 2010 (Figure 4). The thresholding approach had larger SSR values than the product predictions approach for all the situations. The correlation between the observed abundance and the predictions obtained by the product had values between 0.42 for *Potamogeton* sp. in 2012 and 0.74 for *Potamogeton* sp. in 2010. The observation and the predictions are shown in Figure 5.

The global models (Table 3) for *Potamogeton* sp., *C. demersum* and SAV included either year as a linear term, or the linear term day and its interaction with year. Models also included a smooth function of depth and the topographic index (Table 3). Although the abundance of *C. demersum* was included as a smooth function for the model adjusted in *Potamogeton* sp. and *viceversa*, the abundance of the other species was not included as a explanatory variable in global models after the forward selection process. The approach used for calculating the total abundance predictions in the global models was also the product of the predictions of occurrence and conditional abundance. We also found very good predictions for models combining all years (Table 3). Higher values of AUC (higher than 0.50) were observed for the occurrence model. The estimated correlation coefficients were higher for total abundance than for conditional abundance, but always larger than 0.31 (observed in the *Potamogeton* sp. global model).

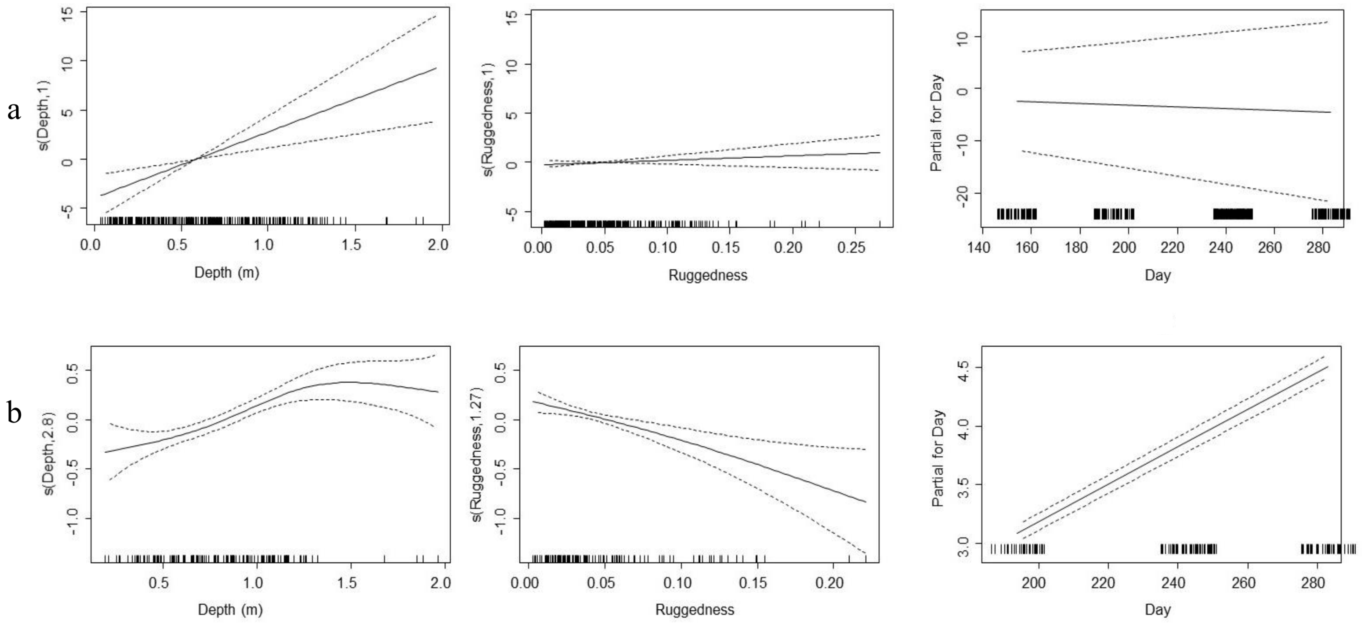

**Figure 2.** Partial additive terms of the adjusted models for occurrence (a) and conditional abundance (b) for *C. demersum* in 2010 for each one of the selected explanatory variables: depth, ruggedness and day. Dashed lines indicate 95% confidence intervals for occurrence and conditional abundance prediction.

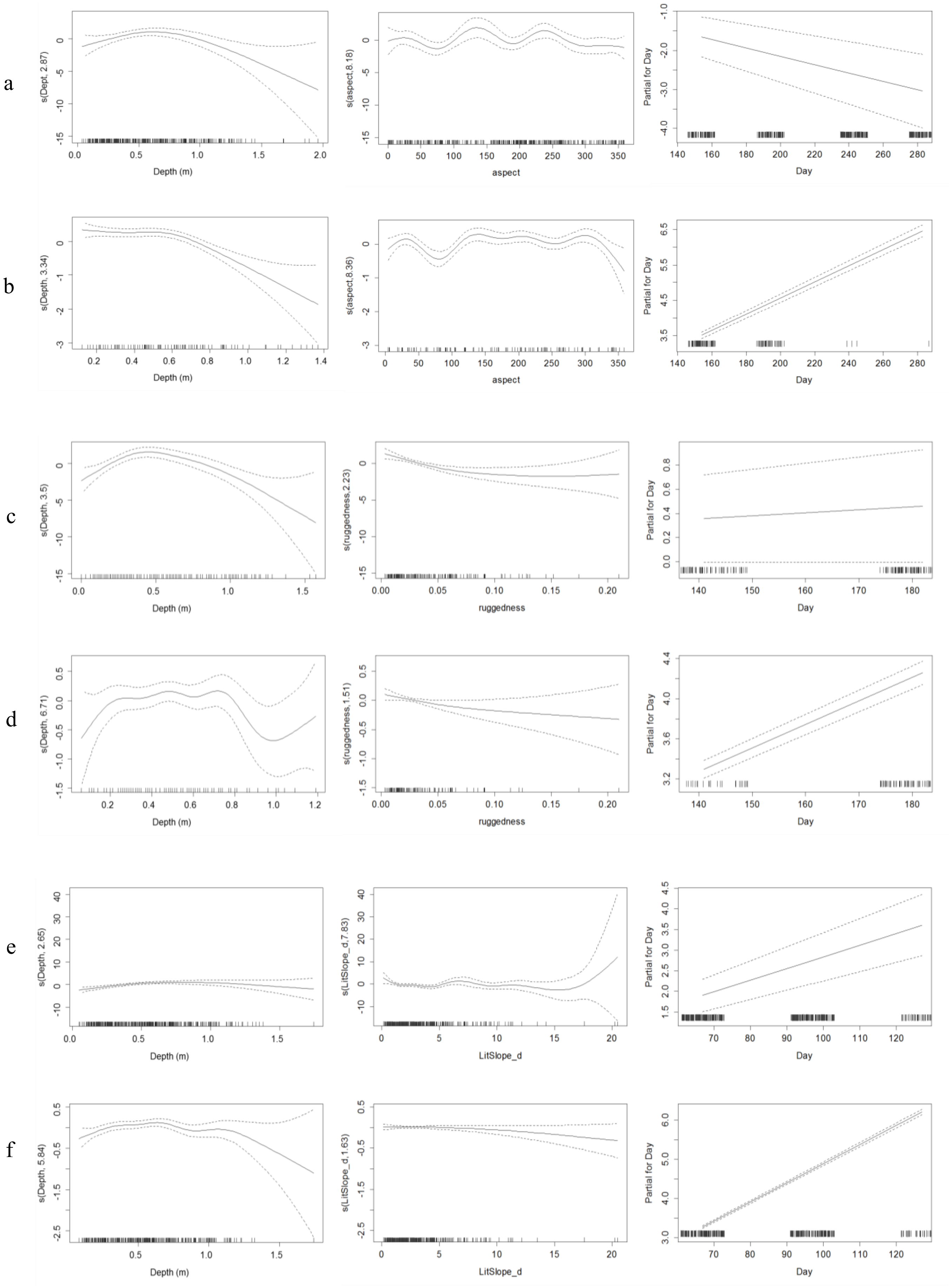

**Figure 3**. Partial additive terms of the adjusted models for occurrence (a, c, e) and conditional abundance (b, d, f) for *Potamogeton* sp. in 2010 (a, b), 2011 (c, d) and 2012 (e, f) for each one of the selected explanatory variables: depth, aspect, ruggedness, littoral slope, and day. Dashed lines indicate 95% confidence intervals for occurrence and conditional abundance prediction.

**Table 3.** Selected models for the combined analysis of occurrence and conditional abundance. For occurrence, model fit is indicated with the area under the curve (AUC). The correlation of observed values and cross-validation predicted values ($r_{(Y_i,\widehat{Y_i})}$) is reported for the conditional abundance model and for the combined model of total abundance. Rooted mean squared error prediction (RMSEP) is reported for total abundance. Mean values of the 1000-iterations are shown, confidence intervals from the empirical distributions are indicated between square brackets.

| | | | | Occurrence model | Conditional abundance model | Total abundance model | |
|---|---|---|---|---|---|---|---|
| Year | Species [†] | # | Selected variables [‡] | AUC | $r_{(Y_i,\widehat{Y_i})}$ | $r_{(Y_i,\widehat{Y_i})}$ | RMSEP |
| 2010 | Cd | 136 | Day + s(Dep) + s(Dep*Day) + s(Rug) | 0.78 [0.77; 0.78] | 0.53 [0.46; 0.57] | 0.71 [0.69; 0.72] | 20.00 [19.55; 20.62] |
| 2010 | Po | 120 | Day + s(Dep) + s(Dep*Day) + s(Asp) | 0.78 [0.77; 0.78] | 0.56 [0.53; 0.59] | 0.74 [0.73; 0.75] | 21.77 [21.40; 22.23] |
| 2010 | SAV | 243 | Day + s(Dep) + s(Dep*Day) + s(Asp) | 0.76 [0.75; 0.76] | 0.45 [0.42; 0.47] | 0.64 [0.63; 0.65] | 27.85 [27.56; 28.16] |
| 2011 | Po | 122 | Day + s(Dep) + s(Dep*Day) + s(Rug) | 0.75 [0.74; 0.76] | 0.52 [0.42; 0.58] | 0.71 [0.69; 0.72] | 27.73 [27.12; 28.49] |
| 2011 | SAV | 127 | Day + s(Dep) + s(Dep*Day) + s(Rug) | 0.75 [0.75; 0.76] | 0.58 [0.54; 0.61] | 0.72 [0.71; 0.74] | 27.32 [26.75; 28.02] |
| 2012 | Po | 332 | Day + s(Dep) + s(Dep*Day) + s(Lit) | 0.62 [0.61; 0.63] | 0.27 [0.23; 0.30] | 0.42 [0.40; 0.45] | 34.75 [34.27; 35.33] |
| 2012 | SAV | 351 | Day + s(Dep) + s(Dep*Day) + s(Rug) | 0.64 [0.63; 0.65] | 0.33 [0.30; 0.36] | 0.53 [0.51; 0.54] | 30.79 [30.40; 31.25] |
| All | Cd | 156 | Year + s(Dep) + s(Top) | 0.72 [0.71; 0.72] | 0.42 [0.39; 0.45] | 0.58 [0.57; 0.60] | 48891 [175077; 181017] |
| All | Po | 574 | Year + Day + s(Dep) + s(Top) + Year*Day | 0.78 [0.78; 0.79] | 0.31 [0.28; 0.33] | 0.67 [0.65; 0.68] | 30.94 [30.46; 31.78] |
| All | SAV | 721 | Year + s(Dep) + s(Top) | 0.69 [0.69; 0.70] | 0.34 [0.31; 0.36] | 0.52 [0.51; 0.52] | 33.82 [33.73; 33.97] |

[†] Cd: *C. demersum*, Po: *Potamogeton* sp., SAV: total submerged aquatic vegetation.

[‡] Dep: depth, Rug: ruggedness, Asp: aspect, Lit: littoral slope, Top: topographic index.

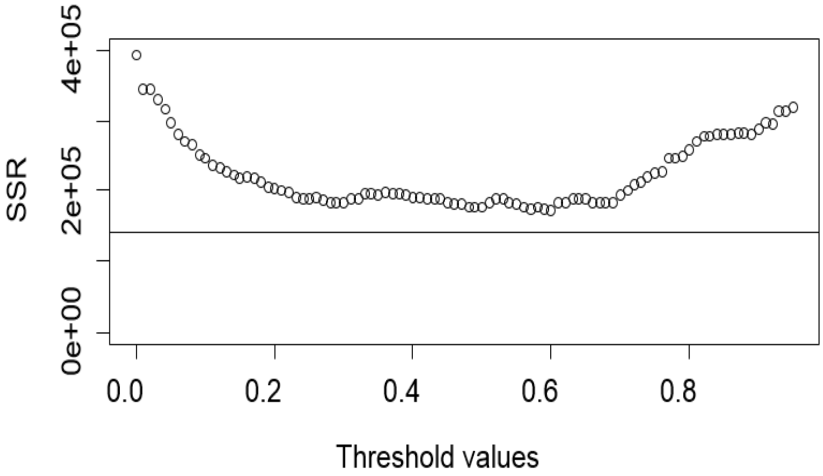

**Figure 4.** Model comparison for *C. demersum* abundance in 2010: Residuals sum of squares (SSR) as a function of different threshold values for occurrence. The total predictions were calculated either as the product of the predictions of occurrence and conditional abundance (solid line), or the abundance predictions for different occurrence threshold values (dots).

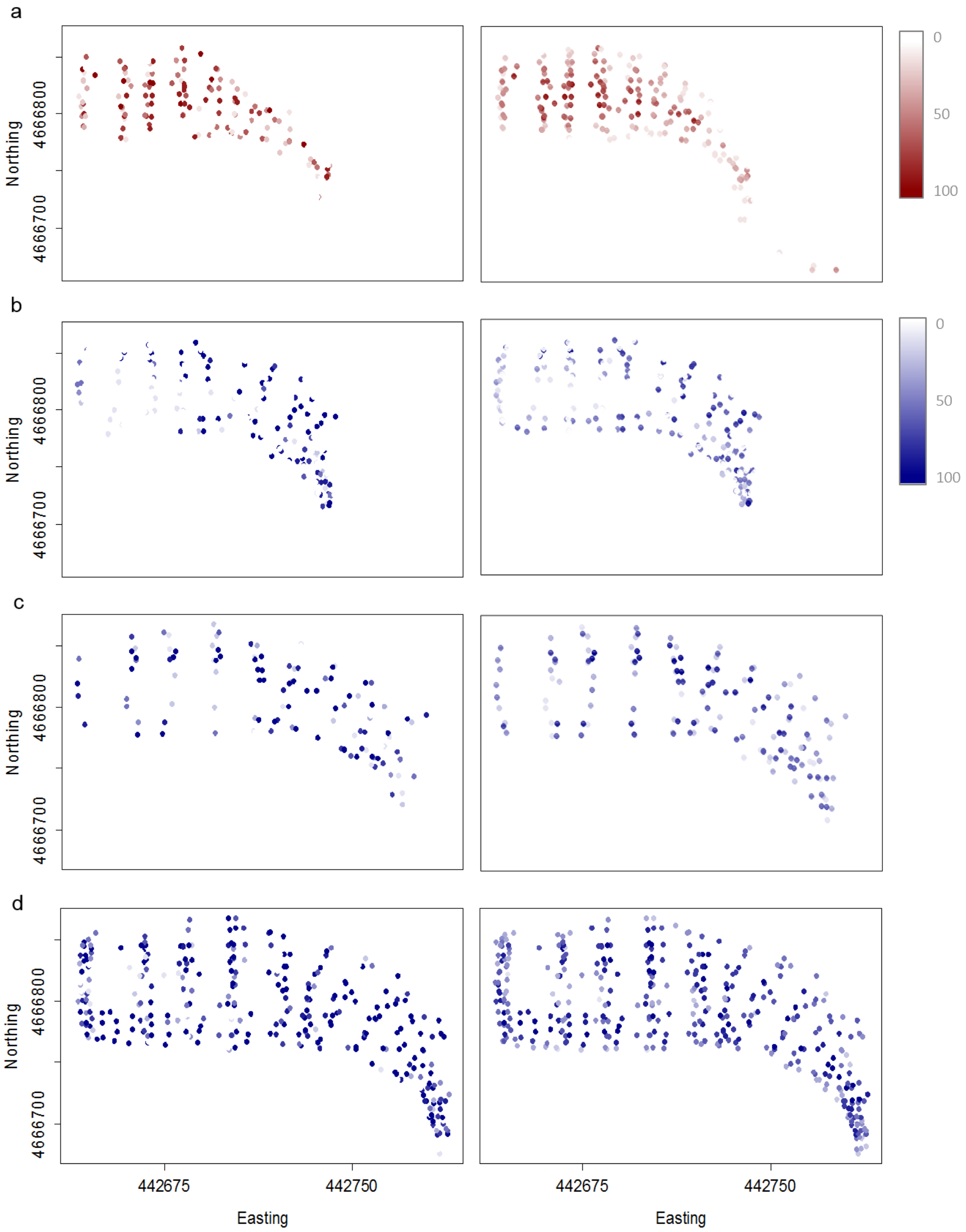

**Figure 5.** Observation (first column) and predictions (second column) of total abundance of *C. demersum* in 2010 (a), and *Potamogeton* sp. in 2010 (b), 2011 (c) and 2012 (d).

### 2.5.3. Unknown year predictions

We obtained reasonable good predictions of occurrence for unknown years (Table 4). However, we found a value of AUC smaller than 0.5 for *C. demersum*-2010 predicting *C. demersum*-2012. We had good predictions for conditional and total abundance when using the year 2010 predicting 2011 for *C. demersum*, but not for *Potamogeton* sp.

**Table 4.** Unknown year predictions. Models where estimated for one year and used to predict an unknown year. For occurrence, model fit is indicated with the area under the curve (AUC). The correlation of observed values and cross-validation predicted values ( $r_{(Y_i,\widehat{Y_i})}$) is reported for the conditional abundance model and for the combined model of total abundance. Rooted mean squared error prediction (RMSEP) is reported for total abundance. Mean values of the 1000-iterations are shown, confidence intervals from the empirical distributions are indicated between square brackets.

| | | | Occurrence model | Conditional abundance model | Total abundance model | |
|---|---|---|---|---|---|---|
| Specie[†] | Training year | Predicted year | AUC | r | r | RMSEP |
| SAV | 2010 | 2011 | 0.67 [0.67; 067] | 0.39 [ 0.34; 0.43] | 0.61 [ 0.59; 0.63] | 32.07 [31.61; 32.56] |
| | | 2012 | 0.50 [0.50; 0.50] | 0.11 [ 0.05; 0.16] | 0.24 [ 0.17; 0.28] | 76.09 [75.72; 76.29] |
| SAV | 2011 | 2012 | 0.56 [0.56; 0.56] | 0.18 [ 0.01; 0.23] | 0.19 [ 0.17; 0.20] | 76.58[‡] [76.39; 76.76] |
| Po | 2010 | 2011 | 0.55 [0.52; 0.57] | -0.09 [-0.17; -0.05] | 0.18 [ 0.11; 0.22] | 46.02 [44.76; 48.56] |
| | | 2012 | 0.54 [0.51; 0.56] | 0.05 [-0.01; 0.09] | 0.03 [-0.04; 0.06] | 52.45 [47.82; 73.85] |
| Po | 2011 | 2012 | 0.56 [0.54; 0.57] | 0.17 [-0.08; 0.24] | 0.16 [ 0.12; 0.19] | 72.00 [71.49; 71.79] |
| Cd | 2010 | 2011 | 0.56 [0.56; 0.57] | 0.60 [ 0.21; 0.82] | 0.19 [ 0.12; 0.26] | 11.38 [10.76; 12.07] |
| | | 2012 | 0.47 [0.45; 0.50] | 0.07 [-0.67; 0.72] | -0.01 [-0.02; -0.01] | 1.45 [ 1.45; 1.45] |

[†] Cd: *C. demersum*, Po: *Potamogeton* sp., SAV: total submerged aquatic vegetation.

[‡] An extreme value of 1.081e+10 was omitted.

## 2.6. DISCUSSION

The two step approach to modelling data with high proportion of zeros provides a way to generate abundance prediction when there is overdispersion, we modeled abundance only when the species is present and the occurrence separately (Welsh *et al.*, 1996).

We compared two approaches to obtaining total abundance predictions (the product of the predictions approach and the thresholding approach), both of them combined the occurrence and conditional abundance predictions. The thresholding approach is a very well-known method, where a specific threshold is chosen to transform occurrence probabilities into presence/absence data. However, choosing the appropriate threshold method remains challenging (Liu *et al.*, 2005; Jiménez-Valverde and Lobo, 2007). The product of the predictions approach is a recent method proposed in 2002 (Barry and Welsh, 2002). In this work, we compared the thresholding approach using all possible thresholds between 0 and 1, to the product of occurrence and conditional abundance predictions. Although the use of thresholding is a well-established method and highly successful, we found that weighting the conditional abundance by probabilities of occurrence was better for all the adjusted models. Although the product of predictions approach was proposed by Barry and Welsh (2002) and used for example by Jensen *et al.* (2005), Gómez and Defeo (2012), Lauria *et al.* (2015) and Parra *et al.* (2016), our work is the first that compared both approaches.

With this work we were able to obtain very good predictions of total abundance of the aquatic vegetation for different years and in the global models using all the years. The high degree of correlation between the predicted and observed values indicates that we can reliably predict total abundance using this models. This was also seen by Jensen and collaborators (2005). They also found that inter-annual predictions (unknown years) were not as good as intra-annual predictions as seen in this work. This could be because of differences between years that are not considered in the variables used in the models. As seen in Drexler and Ainsworth (2013) GAMs are successfully for predicting abundance, additionally to their exploratory use and a tool for identifying influential environmental variables. We had better predictions for total abundance when the occurrence and conditional abundance predictions are combined, having better predictive ability than conditional abundance. The predictions produced a spatial map very similar to that based on observations, although it was generated using unknown data within the year.

Predicting for new years was challenging. The very low correlation coefficients observed in the models of total submerged aquatic vegetation and *Potamogeton* sp. adjusted in both 2011 and 2012 to predict 2010 could be a consequence that in 2011 and 2012 there is only one dominant species and is difficult to predict stations where there are two possible dominant species.

Other methods can be used to make predictions, one of these is the classification tree method. However, when working with fine scales as in this work, GAM are preferred (Thuiller *et al.*, 2003). Classification tree analysis showed lower accuracy than generalized methods (linear and additive models), especially at small scales. However, the performance of GAM was constant for all scales (Thuiller *et al.,* 2003). This comparison was made for occurrence data. It is desirable to compare other methods to the performance that we have in this work, for example comparing classification trees to evaluate their use for zero-inflated data.

Species occurrence and their abundance could be affected by environmental variables such as water temperature, salinity, pH, light penetration (water clarity), fetch or nutrients concentration (Al-Kenzawi, 2007, 2009; Herb and Stefan, 2006; Spencer and Ksander, 1991 which were not considered in this work and could improve models for make better predictions. It is desirable to adjust a model for more than one wetland, that includes other information regarding to each wetland as climatic and location data, to have a model capable to predict in more than one situation, and to explore the capacity of the model of making predictions for unknown years if we include new explanatory variables in the models not considered in this work.

## 2.7. REFERENCES

Aitchison, J., Brown, J.A.C., 1957. The Lognormal Distribution. Cambridge University Press, Cambridge, UK.

Akaike, H., 1973. Information theory as an extension of the maximum likelihood principle, in: Petrov, B.N., Csaki, F. (Eds.), Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 267–281.

Al-Kenzawi, M.A.H., 2009. Seasonal changes of nutrient concentrations in water of some locations in southern Iraqi marshes, after restoration. Baghdad Sci. J. 6(4), 711–718.

Al-Kenzawi, M.A.H., 2007. Ecological study of aquatic macrophytes in the central part of the marshes of Southern Iraq. M.Sc. Thesis. Baghdad University-College of Science for Women, Irak.

Barry, S.C., Welsh, A.H., 2002. Generalized additive modelling and zero inflated count data. Ecol. Model. 157, 179–188.

Drexler M., Ainsworth C.H., 2013. Generalized additive models used to predict species abundance in the gulf of Mexico: An Ecosystem Modelling Tool. Plos one. 8(5), 1–7.

Gómez, J., Defeo, O., 2012. Predictive distribution modeling of the sandy-beach supralittoral amphipod *Atlantorchestoidea brasiliensis* along a macroscale estuarine gradient. Estuar. Coast. Shelf Sci. 98, 84–93.

Greenway, M., 2007. The role of macrophytes in nutrient removal using constructed wetlands, in: Singh, S.N., Tripathi, R.D. (Eds.), Environmental bioremediation technologies. Springer, Berlin, pp. 331–351.

Guisan, A., Edwards, T.C., Hastie, T., 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecol. Model. 157, 89–100.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized additive models, Chapman & Hall, USA.

Hastie, T.J., Tibshirani, R.J., 1986. Generalized additive models. Stat. Sci. 1, 297–318.

Herb, W.R., Stefan, H.G., 2006. Seasonal growth of submersed macrophytes in lakes: the effects of biomass density and light competition. J. Ecol. Model. 193, 560–574.

Jensen, P.O., Seppelt, R., Miller, T.J., Bauer, L.J., 2005. Winter distribution of blue crab *Callinectes sapidus* in Chesapeake Bay: application and cross-validation of a two-stage generalized additive model. Mar. Ecol. Prog. Ser. 299, 239–255.

Jiménez-Valverde, A., Lobo, J.M., 2007. Threshold criteria for conversion of probability of species presence to either–or presence–absence. Acta Oecol. 31, 361–369.

Lauria, V., Gristina, M., Attrill, M.J., Fiorentino, F., Garofalo, G., 2015. Predictive habitat suitability models to aid conservation of elasmobranch diversity in the central Mediterranean Sea. Sci. Rep. 5, 13245.

Liu, C., Berry, P.M., Dawson, T.P., Pearson, R.G., 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography. 28, 385–393.

McCullagh, P., Nelder, J.A., 1989. Generalized linear models. Chapman & Hall, London.

Mitsch, W.J., Gosselink, J.G., 2000. Wetlands, John Wiley and Sons Ltd., New York.

Parra, H.E., Phama, C.K., Menezesa, G.M., Rosa, A., Tempera, F., Morato, T., 2016. Predictive modelling of deep-sea fish distribution in the Azores. Deep Sea Res. 5, 1-12.

Pennington, M., 1983. Efficient estimators of abundance, for fish and plankton surveys. Biom. 39: 281–286.

R Development Core Team, 2014. R: A language and environment for statistical computing, reference index version. Available at http://www.R-project.org (verified 20 Feb. 2014). R Foundation for Statistical Computing, Vienna, Austria.

Reddy, K.R., D'Angelo, E.M., 1997. Biogeochemical indicators to evaluate pollutant removal efficiency in constructed wetlands. Water Sci. Technol. 35, 1–10.

Spencer, D., Ksander, G., 1991. Influence of temperature and light on early growth of *Potamogeton gramineus* L. J. Freshw. Ecol. 6, 227–235.

Stefánsson, G., 1996. Analysis of Groundfish Survey Abundance Data: Combining the GLM and Delta Approaches. J. Mar. Sci. 53: 577–588.

Thuiller, W., Araújo, M.B., Lavorel, S., 2003. Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. J. Veg. Sci. 14: 669-680.

Vymazal, J., Greenway, M., Tonderski, K., Brix, H., Mander, Ü., 2006. Constructed wetlands for wastewater treatment, in: Verhoeven, J.T.A., Beltman, B., Bobbink, R., Whigham, D.F. (Eds.), Wetlands and natural resource management. Springer, Berlin, pp. 69–96.

Warton, D.I., 2005. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Env. 16: 275–289.

Watson, J.T., Sherwood, S.C., Kadlec, R.H., Knight, R.L., Whitehouse, A.E., 1989. Performance expectations and loading rates for constructed wetlands, in: Hammer, D.A. (Ed.), Constructed wetlands for wastewater treatment: municipal, industrial and agricultural. CRC Press, Chelsea, pp. 319–351.

Welsh, A.H., Cunningham, R., Donnelly, C., Lindenmeyer, D., 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecol. Model. 88, 297–308.

Wetzel, R.C., 2001. Fundamental processes within natural and constructed wetland ecosystems: short-term versus long-term objectives. Water Sci. Technol. 44 (11–12), 1–8.

Williams, J.B., 2002. Phytoremediation in wetland ecosystems: progress, problems, and potential. Crit. Rev. Plant Sci. 21(6): 607–635.

Wood, S.N., 2006. GAMs in practice: mgcv, in: Wood, S.N. (Ed.), Generalized Additive Models: An Introduction with R. Chapman & Hall, London, pp. 217–271.

Yee, T.W., Mitchell, N.D., 1991. Generalized additive models in plant ecology. J. Veg. Sci. 2: 587–602.

# 3. RESULTADOS

## 3.1. CARACTERIZACIÓN DEL MUESTREO

Un total de 993 localidades se registraron durante los tres años de estudio. La profundidad media fue similar para todos los años estudiados, y las variables topográficas eran del mismo orden para todos los años, pero la pendiente litoral mostró una media menor de 2,011 en 2010 y 2012. La vegetación acuática sumergida (SAV) se observó en el 62% de las estaciones registradas en 2010, en el 61% en 2011, y en el 90% en 2012. En 2012 no sólo la ocurrencia observada de SAV fue superior a los años anteriores, sino también la abundancia media. En 2010 las especies *C. demersum* y *Potamogeton* sp. fueron dominantes. Se observó *C. demersum* en el 34% de las estaciones totales, mientras que *Potamogeton* sp. en el 30% de las estaciones de este año. En 2010 las estaciones con *Potamogeton* sp. tenían una profundidad media de 0,58 m y un máximo de 1,37 m, y *C. demersum* una profundidad de 0,82 m, y un máximo de 1,96 m. *Potamogeton* sp. fue la especie dominante en 2011 y 2012, se observó en el 59% y el 86% de las estaciones respectivamente. *C. demersum* disminuyó su frecuencia y abundancia observada a través de los años, y *Potamogeton* sp. aumentó su presencia y abundancia. En 2011, las estaciones con *Potamogeton* sp. tenían una profundidad media de 0,45 m y un máximo de 1,19 m, y en 2012 la profundidad media de las estaciones con *Potamogeton* sp. fue de 0,54 m. Otras especies de SAV registrados durante los muestreos fueron: *P. natans* (1%) y *Chara* sp. (1%) en 2010, *P. natans* (0,5%) en el año 2011, y *Chara* sp. (10%), *P. natans* (2%) y *C. demersum* (2%) en el año 2012.

## 3.2. MODELOS DE OCURRENCIA Y ABUNDANCIA CONDICIONAL

Los modelos seleccionados de ocurrencia y abundancia condicional para las dos especies y para SAV total en cada año incluyeron día de muestreo como un término lineal, y las funciones suavizadas de profundidad, día por profundidad, y las variables aspecto o robustez. Estos modelos presentaron valores de AUC mayores a 0.5 en todos los casos, y altos valores del coeficiente de correlación para los modelos de abundancia. Los valores predictivos más pequeños se obtuvieron para *Potamogeton* sp. en 2012. Para combinar las predicciones de ocurrencia y abundancia condicional el método que dio una menor suma de cuadrado residual fue el producto de las predicciones en todos los casos. El enfoque de umbrales siempre tuvo valores de SSR más grandes que el enfoque de producto de las predicciones. La correlación entre la abundancia observada y las predicciones de abundancia total (obtenidas como el producto) presentó valores entre 0.42 para *Potamogeton* sp. en 2012 y 0,74 para *Potamogeton* sp. en 2010.

Los modelos globales seleccionados de *Potamogeton* sp., *C. demersum* y SAV total incluyeron año y día como término lineal, además de la interacción entre ellos para *Potamogeton* sp. Los tres modelos incluyeron la función suavizada de profundidad y del índice topográfico. El enfoque utilizado para el cálculo de las predicciones de la abundancia total también fue el producto de las predicciones para los modelos globales como se observó anteriormente.

## 3.3. PREDICCIONES PARA AÑOS DESCONOCIDOS

Las predicciones para años desconocidos no fueron tan buena como las predicciones realizadas dentro de un mismo año. El coeficiente de correlación estimado para la abundancia condicional y total fue alta para el año 2010 prediciendo 2011 y 2012, para el año 2011 prediciendo 2012, y para 2012 prediciendo 2011. Se observaron valores bajos de estimación de los modelos ajustados en SAV-2011 para predecir SAV-2010 y en SAV 2012 para predecir SAV-2010. Los casos mencionados también están relacionados con altos valores de la RMSEP. Hay un valor extremo de RMSEP en SAV-2011 predecir SAV-2012 en este caso la mediana de la RMSEP se 77.0, 1.081e + 10 es un valor extremo observado en una de las 1000 iteraciones realizadas.

Los menores valores de AUC para *Potamogeton* sp. se observaron en los modelos ajustados en Po-2011 para predecir Po-2010 y Po-2012 para predecir Po-2010, esto también es visto en abundancia total con bajos coeficientes de correlación y valores altos de RMSEP. Para *C. demersum* el modelo ajustado en 2010 predice de forma diferencial los dos años, siendo mejores las predicciones para 2011.

A pesar de la abundancia de *C. demersum* se incluyó como predictora para el modelo ajustado de *Potamogeton* sp. y viceversa, la abundancia de la otra especie no fue incluida en la selección ya que no se veía un decremento del AIC.

## 4. DISCUSIÓN Y CONCLUSIONES

La partición de los datos totales en datos de ocurrencia y abundancia condicional proporciona un método simple para generar predicciones cuando existe una alta proporción de ceros a fin de evitar sobredispersión (Welsh *et al.*, 1996).

Se compararon dos enfoques que combinan las predicciones de los modelos de ocurrencia y abundancia condicional para la obtención de las predicciones de la abundancia total. La selección de un umbral específico para transformar las probabilidades de ocurrencia en datos de presencia/ausencia es un método muy conocido. Sin embargo, la elección del umbral adecuado sigue siendo un reto (Liu *et al.*, 2005; Jiménez-Valverde y Lobo, 2007). En este trabajo, en lugar de seleccionar el mejor umbral se compararon todos los umbrales posibles entre 0 y 1. El producto de las predicciones de ambos modelos fue propuesto por Barry y Welsh (2002). Aunque el uso de umbrales es un método bien establecido, se encontró que el producto de las predicciones de abundancia por las probabilidades de ocurrencia predicha fue mejor para todos los modelos ajustados (*C. demersum* en 2010, *Potamogeton* sp. y SAV en 2010, 2011 y 2012, y para los modelos globales). Aunque el producto de las predicciones fue propuesto por Barry y Welsh (2002) y ha sido utilizado en diversos trabajos como ser: Jensen *et al.* (2005), Gómez y Defeo (2012), Lauria *et al.* (2015) y Parra *et al.* (2016), nuestro trabajo es el primero que compara ambos enfoques.

Con este trabajo hemos sido capaces de obtener muy buenas predicciones de la abundancia total de la vegetación acuática para diferentes años y en los modelos globales que utilizan todos los años. El alto grado de correlación entre los valores predichos y observados indica que podemos predecir de forma fiable la abundancia total utilizando estos modelos. Esto también fue visto por Jensen *et al.* (2005). Sin embargo, ellos también encontraron que las predicciones inter-anuales (para años desconocidos) no son tan buenas como las predicciones intra-anuales. Esto podría ser debido a diferencias entre años que no se consideran en las variables utilizadas en los modelos. Como plantean Drexler y Ainsworth (2013) los GAM pueden ser utilizados con éxito para predecir abundancia, además de su uso exploratorio como una herramienta para la identificación de las variables ambientales influyentes.

La ocurrencia de especies y su abundancia podrían verse afectados por las variables ambientales tales como la temperatura del agua, salinidad, pH, penetración de la luz, concentración de nutrientes, entre otros (Al-Kenzawi, 2007, 2009; Herb y Stefan, 2006; Spencer y Ksander, 1991). Sería deseable incorporar dichas variables para mejorar futuras predicciones.

A futuro se planten comparar diferentes métodos de predicción para datos faltantes y comparar su desempeño; además de incorporar más humedales al análisis incluyendo datos climáticos y de localización, entre otros.

## 5. BIBLIOGRAFÍA

Aitchison J, Brown JAC. 1957. The Lognormal Distribution. Cambridge, UK: Cambridge University Press.

Akaike H. 1973. Information theory as an extension of the maximum likelihood principle. En: Petrov BN & Csaki F. (Eds.). Second International Symposium on Information Theory. Budapest: Akademiai Kiado. 267 – 281.

Al-Kenzawi MAH. 2009. Seasonal changes of nutrient concentrations in water of some locations in southern Iraqi marshes, after restoration. Baghdad Science Journal, 6(4): 711 – 718.

Al-Kenzawi MAH. 2007. Ecological study of aquatic macrophytes in the central part of the marshes of Southern Iraq. M.Sc. Thesis. Baghdad, Iraq. Baghdad University-College of Science for Women.

Barry SC, Welsh AH. 2002. Generalized additive modelling and zero inflated count data. Ecological modelling, 157: 179 – 188.

Drexler M, Ainsworth CH. 2013. Generalized additive models used to predict species abundance in the gulf of Mexico: An Ecosystem Modelling Tool. Plos one, 8(5): 1 – 7.

Gómez J, Defeo O. 2012. Predictive distribution modeling of the sandy-beach supralittoral amphipod *Atlantorchestoidea brasiliensis* along a macroscale estuarine gradient. Estuarine, Coastal and Shelf Science, 98: 84 – 93.

Greenway M. 2007. The role of macrophytes in nutrient removal using constructed wetlands. En: Singh SN, Tripathi RD. (Eds.). Environmental bioremediation technologies. Berlin: Springer. 331 – 351.

Guisan A, Edwards TC, Hastie T. 2002. Generalized linear and generalized additive models in studies of species distributions: setting the scene. Ecological modelling, 157: 89 – 100.

Hastie TJ, Tibshirani RJ. 1990. Generalized additive models. USA: Chapman & Hall.

Hastie TJ, Tibshirani RJ. 1986. Generalized additive models. Statistical Science, 1: 297 – 318.

Herb WR, Stefan HG. 2006. Seasonal growth of submersed macrophytes in lakes: the effects of biomass density and light competition. Journal of Ecological Modelling, 193: 560 – 574.

Jensen PO, Seppelt R, Miller TJ, Bauer LJ. 2005. Winter distribution of blue crab *Callinectes sapidus* in Chesapeake Bay: application and cross-validation of a two-stage generalized additive model. Marine Ecology Progress Series, 299: 239 – 255.

Jiménez-Valverde A, Lobo JM. 2007. Threshold criteria for conversion of probability of species presence to either–or presence–absence. Acta Oecologica, 31: 361 – 369.

Lauria V, Gristina M, Attrill MJ, Fiorentino F, Garofalo G. 2015. Predictive habitat suitability models to aid conservation of elasmobranch diversity in the central Mediterranean Sea. Scientific Reports, 5: 13245.

Liu C, Berry PM, Dawson TP, Pearson RG. 2005. Selecting thresholds of occurrence in the prediction of species distributions. Ecography, 28: 385 – 393.

McCullagh P, Nelder JA. 1989. Generalized linear models. London: Chapman & Hall.

Mitsch WJ, Gosselink JG. 2000. Wetlands. New York: John Wiley and Sons Ltd.

Parra HE, Phama CK, Menezesa GM, Rosa A, Tempera F, Morato T. 2016. Predictive modelling of deep-sea fish distribution in the Azores. Deep Sea Research, 5: 1 – 12.

Pennington M. 1983. Efficient estimators of abundance, for fish and plankton surveys. Biometrics, 39: 281 – 286.

R Development Core Team. 2014. R: A language and environment for statistical computing, reference index version. Available at http://www.R-project.org (verified 20 Feb. 2014). Vienna, Austria: R Foundation for Statistical Computing.

Reddy KR, D'Angelo EM. 1997. Biogeochemical indicators to evaluate pollutant removal efficiency in constructed wetlands. Water science and technology, 35: 1 – 10.

Spencer D, Ksander G. 1991. Influence of temperature and light on early growth of Potamogeton gramineus L. Journal of Freshwater Ecology, 6: 227 – 235.

Stefánsson G. 1996. Analysis of Groundfish Survey Abundance Data: Combining the GLM and Delta Approaches. Journal of Marine Science, 53: 577 – 588.

Thuiller W, Araújo MB, Lavorel S. 2003. Generalized models vs. classification tree analysis: Predicting spatial distributions of plant species at different scales. Journal of Vegetation Science, 14: 669 – 680.

Vymazal J, Greenway M, Tonderski K, Brix H, Mander Ü. 2006. Constructed wetlands for wastewater treatment. En: Verhoeven JTA, Beltman B, Bobbink R, Whigham DF. (Eds.). Wetlands and natural resource management. Berlin: Springer. (Ecological Studies; 190). 69 – 96.

Warton DI. 2005. Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. Environmetrics, 16: 275 – 289.

Watson JT, Sherwood SC, Kadlec RH, Knight RL, Whitehouse AE. 1989. Performance expectations and loading rates for constructed wetlands. En: Hammer DA (Ed.). Constructed wetlands for wastewater treatment: municipal, industrial and agricultural. Chelsea: CRC Press. 319 – 351.

Welsh AH, Cunningham R, Donnelly C, Lindenmeyer D. 1996. Modelling the abundance of rare species: statistical models for counts with extra zeros. Ecological modelling, 88: 297 – 308.

Wetzel RC. 2001. Fundamental processes within natural and constructed wetland ecosystems: short-term versus long-term objectives. Water science and technology, 44 (11–12): 1 – 8.

Williams JB. 2002. Phytoremediation in wetland ecosystems: progress, problems, and potential. Critical reviews in plant sciences, 21(6): 607 – 635.

Wood SN. 2006. GAMs in practice: mgcv. En: Wood SN. (Eds.). Generalized Additive Models: An Introduction with R. London: Chapman & Hall. 217 – 271.

Yee TW, Mitchell ND. 1991. Generalized additive models in plant ecology. Journal of vegetable science, 2: 587 – 602.