



UPPSALA
UNIVERSITET



UPTEC W 16004

Examensarbete 30 hp
Mars 2016

Ensemblemodellering av piggvarens habitat utgående från provfiske- och miljödata

Mårten Erlandsson

SAMMANFATTNING

Ensemblemodellering av piggvarens habitat utgående från provfiske- och miljödata

Mårten Erlandsson

Piggvarens (*Scophthalmus maximus*) val av habitat i Östersjön har modellerats utifrån provfiskedata och miljövariabler. Vid totalt 435 stationer i Östersjön har data samlats in i form av provfiske, CTD-mätningar (konduktivitet, temperatur och djup) och videofilmer. Genom att analysera videofilmerna från havsbotten i Östersjön har den klassificerats efter fyra olika förklaringsvariabler: täckningsgrad mjukbotten, strukturbildande växter, övriga alger och täckningsgrad blåmusslor. Ytterligare sex förklaringsvariabler har samlats in från mätningar och befintliga kartor: bottensalinitet, botten temperatur, djup, siktdjup, vågexponering och bottenlutning. Dessa tio förklaringsvariabler har använts i tio olika enskilda statistiska modelleringssyften med förekomst/icke-förekomst av piggvar som responsvariabel. Nio av tio modeller visade på bra resultat ($AUC > 0,7$) där CTA (Classification Tree Analysis) och GBM (Global Boosting Model) hade bäst resultat ($AUC > 0,9$). Genom att kombinera modeller med bra resultat på olika sätt skapades sex ensemblemodeller för att minska varje enskild modells svagheter. Ensemblemodellerna visade tydligt fördelarna med denna typ av modellering då de gav ett mycket bra resultat ($AUC > 0,949$). Den sämsta ensemblemodellen var markant bättre än den bästa enskilda modellen.

Resultaten från modellerna visar att största sannolikheten för piggvarsförekomst i Östersjön är vid grunt (< 20 meter) och varmt (> 10 °C) vatten med hög vågexponering ($> 30\,000$ m²/s). Dessa tre variabler var de med högst betydelse för modellerna. Täckningsgrad mjukbotten och de två växtlighetsvariablerna från videoanalyserna var de tre variabler som hade lägst påverkan på piggvarens val av habitat. Med en högre kvalitet på videofilmerna hade de variablerna kunnat klassificeras i mer specifika grupper vilket eventuellt gett ett annat resultat. Generellt visade modellerna att denna typ av habitatmodellering med provfiske och miljödata både är möjlig att utföra.

Nyckelord: Piggvar, Habitatmodellering, Ensemblemodellering, Videoanalys

ABSTRACT

Ensemble modelling of the habitat of turbot based on video analyses and fish survey data

Mårten Erlandsson

The turbot's (*Scophthalmus maximus*) selection of habitat in the Baltic Sea has been modeled on the basis of fish survey data and environmental variables. At a total of 435 stations in the Baltic Sea, data was collected in the form of fish survey data, CTD (Conductivity, Temperature and Depth) measurements and videos. By analyzing the videos from the seabed of the Baltic Sea, four different explanatory variables have been classified: coverage of soft bottom, structure-forming plants, other algae and coverage of mussels. Another six explanatory variables have been collected from measurements and existing rasters: salinity, temperature, depth, water transparency, wave exposure and the bottom slope. These ten explanatory variables have been used in ten different species distribution modeling methods with the presence/absence of turbot as a response variable. Nine out of ten models showed good results ($AUC > 0.7$) where the CTA (Classification Tree Analysis) and GBM (Global Boosting Model) performed the best ($AUC > 0.9$). By combining the models with good performance in six different ensemble models each individual models' weaknesses were decreased. The ensemble models clearly showed strength as they gave a very good performance ($AUC > 0.94$). The worst ensemble model was significantly better than the best individual model.

The results of the models show that the largest probability of occurrence of turbot in the Baltic Sea is in shallow (< 20 m) and warm ($> 10^{\circ}\text{C}$) water with high wave exposure ($> 30,000\text{ m}^2/\text{s}$). These three variables were those with the highest significance for the models. Coverage of soft bottom and the two vegetation variables, from the video analyzes, had the lowest impact on the turbot's choice of habitat. A higher quality of the videos would have made it possible to classify these variables in more specific groups which might have given a different result. Generally, the models showed that this type of modeling of habitat is possible to perform with fish survey and environmental monitoring data and generates useful results.

Keywords: Turbot, Species distribution modeling, Ensemble Modeling, Video Analysis

FÖRORD

Denna studie har genomförts som det avslutande examensarbetet inom Civilingenjörsprogrammet i Miljö- och vattenteknik vid Uppsala universitet och Sveriges lantbruksuniversitet. Arbetet har utförts under 2015/2016 och motsvarar 30 hp. Handledare har varit Ann-Britt Florin och biträdande handledare har varit Ulf Bergström, båda vid institutionen för akvatiska resurser vid Sveriges lantbruksuniversitet. Examinator har varit Anna Sjöblom vid institutionen för geovetenskaper, Uppsala universitet och ämnesgranskare har varit Andreas Bryhn vid institutionen för akvatiska resurser, Sveriges lantbruksuniversitet.

Jag vill tacka Ann-Britt Florin för sin expertis och hjälpsamma vägledning genom hela projektet. Ulf Bergström vill jag tacka så mycket för all hjälp med alla frågor som har dykt upp under projektets gång. Jag vill även rikta ett tack till Alessandro Orio som var till stor hjälp under videoanalyserna.

Utöver dessa vill jag tacka Andreas Bryhn, Anna Sjöblom och Victor Eliasson för granskning och förbättring av slutrapporten.

Kartan i figur 3 i denna rapport är skapad med ArcGIS® mjukvara från Esri. ArcGIS® och ArcMap™ används i rapporten enligt licens från ESRI. Copyright © Esri. All rights reserved. För mer information om ESRI® programvara, besök www.esri.com

POPULÄRVETENSKAPLIG SAMMANFATTNING

Statistiska modeller är ett viktigt verktyg för att studera hur miljövariabler påverkar olika arters val av levnadsmiljö. De kan bland annat ge en bild över vilken typ av arbete som är viktigast att prioritera för att förbättra statusen för hotade arter. De är även viktiga instrument vid analyser av markanvändning och andra miljöförändringar. Genom en ökad kunskap om sambandet mellan en hotad art och miljön den lever i, kan artens bevarandearbete effektiviseras. Med teknologins och ekologins framsteg de senaste decennierna har modelleringen av arters levnadsmiljöer utvecklats kraftigt. Många nya, mer avancerade modelleringsmetoder med olika styrkor och svagheter har utvecklats. Sambandet mellan arter och miljövariabler är ofta komplexa och kan vara svåra att förenkla. Det leder till att det är mycket svårt att avgöra vilken typ av modell man bör använda. Denna studie har jämfört tio olika typer av statistiska modeller för att få en bättre bild av piggvarens val av levnadsmiljö i Östersjön.

Piggvarens finns längs hela Europas kust, från Medelhavet i söder upp till Ålands hav i Östersjön. Lokalt är piggvaren en viktig kommersiell resurs för fiskare. Internationella naturvårdsunionens rödlistningssystem har klassat piggvaren som ”nära hotad” i Östersjön. Motiveringen är högt fisketryck, förstörelse av uppväxtområden och syrebrist i djupare övervintringsområden på grund av övergödning.

Piggvaren har fiskats vid 435 stationer i Östersjön under två års tid. I samband med fisket har även videosekvenser spelats in från havsbotten och mätningar av vattnets egenskaper genomförts vid varje station. De variabler som mättes var djup, salthalt, temperatur och siktdjup. Genom att titta på videofilmerna från havsbotten avgjordes hur stor del av botten som täcktes av fyra olika variabler. Dessa fyra variabler var täckningsgrad mjukbotten (silt, sand och grus), större växter och alger, övriga alger och blåmusslor. De sista två variablerna som användes till modelleringen var vågexponering (hur utsatt stationen är för vind) och bottenlutning. Dessa hämtades från befintliga databaser.

Modellerna kopplade ihop sambandet mellan miljövariablerna vid varje station med om det hade fångats piggvar vid stationen. För att skapa modellerna användes data från 80 % av stationerna. För att avgöra hur bra en modell är testades den på de återstående 20 % stationerna. Eftersom antalet piggvarar är känt på de stationer där modellen testades kunde modellerna bedömas beroende på hur många stationer där piggvar förekom som den klassat rätt och hur många stationer där piggvar inte förekom som den klassat rätt.

Nio av de tio modellerna fick ett godkänt resultat, där två stack ut med mycket bra resultat. Dessa två var CTD (Classification Tree Analysis) och GBM (Global Boosting Model). Även om dessa modeller visade på bra resultat så har de fortfarande svagheter. Genom att kombinera alla modeller med bra resultat med olika matematiska metoder togs ytterligare sex modeller fram. De lyckades mycket framgångsrikt att beskriva piggvarens habitatval. Alla sex kombinerade modellerna hade bättre resultat än de bästa enskilda modellerna. Bäst lyckades en modell som rangordnade de olika enskilda modellerna utefter dess resultat och lät dem påverka den kombinerade modellen i olika grad beroende på dess rang. Slutsatsen från modelleringen blev således att det är möjligt att modellera piggvarens habitat i Östersjön utifrån miljövariabler. De kombinerade modellerna har potential att på ett mycket effektivt sätt minska de olika individuella modellernas svagheter.

Resultaten från modellerna visar att största sannolikheten för piggvarsförekomst i Östersjön var vid grunt (grundare än 20 meter) och varmt (varmare än 10 °C) vatten med hög vågexponering (mer än 30 000 m²/s). Djup, temperatur och vågexponering var de variabler som visade sig spela störst roll för piggvaren i dess val av habitat. Variablerna som kom från videoanalyserna visade sig vara de som spelade minst roll för piggvaren. Eventuellt skulle de påverka mer om man använde enskilda alger eller växter i modellerna istället för att gruppera in dem i större grupper. För att en sådan indelning ska vara möjlig behövs videosekvenser med hög videokvalitet vilken kan vara svårt på havsbotten med dess låga ljusstillgång.

ORDLISTA

ANN	Artificial Neural Network. En statistisk maskinlärande modell.
ArcGIS	Mjukvara från ESRI för analys och bearbetning av geografisk data
AUC	Area Under Curve. Areal under ROC-kurvan. Ger ett värde på hur bra en modell är.
BIOMOD2	BIOdiversity MODelling. Ett paket i statistikprogrammet R för habitatmodellering.
CTA	Classification Tree Analysis. En statistisk klassifikationsmodell.
CTD	Ett mätinstrument som mäter konduktivitet, temperatur och djup
Ensemblemodell	En kombination av flera enskilda modeller
FDA	Flexible Discriminant Analysis. En statistisk klassifikationsmodell.
GAM	Generalized Additive Models. En statistisk regressionsmodell.
GBM	Generalised Boosted Models. En statistisk maskinlärande modell.
GIS	Geografiska informationssystem.
GLM	Generalized Linear Models. En statistisk regressionsmodell.
Kollinearitet	Ett mått på samvariationen mellan variabler.
MARS	Multivariate Adaptive Regression Splines. En statistisk regressionsmodell
MAXENT	Maximum Entropy. En statistisk regressionsmodell.
R	Programvara för statistiska analyser.
RF	Random Forest. En statistisk maskinlärande modell.
ROC-kurva	Receiver Operating Characteristic curve. En grafisk presentation av hur en modell presterar.
Sensitivitet	Sannolikheten att en modell har klassat ett positivt värde rätt kallas för sensitivitet.
Specificitet	Sannolikheten för att ett negativt värde är rätt klassat kallas för specificitet.

SRE

Surface Range Envelope. En statistisk envelope-modell.

VIF

Variance Inflation Factor. Ett mått på hur mycket flera variabler korrelerar med varandra.

INNEHÅLLSFÖRTECKNING

1	Introduktion	1
1.1	Syfte	3
2	Metod.....	4
2.1	Insamling av data	4
2.2	Videoanalyser	5
2.3	Behandling av data.....	6
2.4	Statistiska analyser.....	7
2.4.1	Kollinearitetstest.....	7
2.4.2	Statistisk modellering	8
2.5	Beskrivning av modeller	9
2.5.1	GLM - Generalized Linear Models	9
2.5.2	GAM - Generalized Additive Models	9
2.5.3	MARS – Multivariate Adaptive Regression Splines	9
2.5.4	CTA – Classification Tree Analysis	9
2.5.5	FDA – Flexible Discriminant Analysis	10
2.5.6	ANN – Artificial Neural Network	10
2.5.7	GBM – Generalised Boosted Models.....	10
2.5.8	RF – Random Forest.....	10
2.5.9	MAXENT – Maximum Entropy.....	10
2.5.10	SRE – Surface Range Envelope	10
2.5.11	Ensemblemodeller	10
2.6	Modellutvärdering	11
3	Resultat	12
3.1	Videoanalyser	12
3.2	Kollinearitet	12
3.3	Modellresultat	13
3.3.1	Förklaringsvariabler.....	13
3.3.2	Modellutvärdering	18
4	Diskussion	21
5	Slutsatser.....	24
6	Referenser	25
7	Bilagor	28

1 INTRODUKTION

Geografiska prediktionsmodeller är ett viktigt populationsekologiskt verktyg. De kan bland annat ge en bild över vilken typ av bevarandearbete som är viktigast att prioritera för hotade arter. Förutom deras fördelar inom populationsekologin är de även viktiga instrument vid analys av markanvändning och andra miljöförändringar (Guisan & Zimmermann, 2000). För att optimera effekten av eventuella åtgärder i bevarandearbetet av hotade arter är det viktigt med kunskap om samband mellan arten och olika klimat- och miljövariabler. Denna typ av samband har under lång tid använts för att förklara olika arters utbredning och ligger som grund för dagens geografiska prediktionsmodeller (Guisan & Zimmermann, 2000). Behovet av en ökad förståelse av populationers dynamik och distribution tillsammans med datorteknologins framsteg och tillgången av miljödata har lett till en kraftig utveckling av arters habitatmodellering (Sundblad, 2010; Reiss, et al., 2011).

Det finns idag ett flertal olika typer av modelleringstekniker som olika frekvent används vid habitatmodellering. Trots att nyare maskinlärande habitatmodelleringstekniker som Maximum Entropy (MAXENT), Global Boosting Model (GBM) och Random Forest (RF) har visat sig vara mycket användbara och ofta ger ett bättre resultat än mer klassiska metoder som Generalized Linear Models (GLM), används dessa fortfarande relativt sällan (Elith, et al., 2006; Bucas, et al., 2013). Genom att kombinera resultatet från enskilda modeller till en gemensam modell skapas en så kallad ensemblemodell. En stor fördel med ensemblemodeller är att de kan utnyttja de enskilda modellernas styrkor och minimera deras svagheter vilket kan leda till bättre prediktioner (Arujo, 2007).

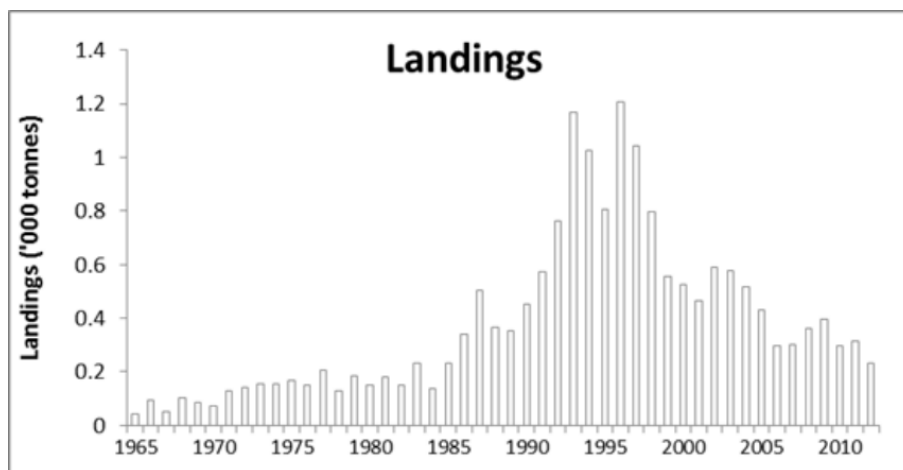
Med den senaste versionen av paketet BIOMOD2 i statistikprogrammet R finns det möjlighet att använda tio olika modelleringstyper samtidigt med samma indata (Thullier, et al., 2015). Habitatmodeller kan generellt delas in i fyra olika grupper: regressionsmodeller (till exempel GLM, Generalized Additive Models (GAM) och Multivariate Adaptive Regression Splines (MARS)), klassifikationsmodeller (till exempel Flexible Discriminant Analysis (FDA) och Classification Tree Analysis (CTA)), maskinlärande modeller (till exempel Artificial Neural Network (ANN), GBM, RF och MAXENT) och så kallad envelope-teknik (till exempel Surface Range Envelope (SRE)).

Piggvaren (*Scophthalmus maximus*) är en art som intressant för habitatmodellering då den är viktig både ur kommersiell- och bevarandesynpunkt (HELCOM, 2013). Piggvaren är en plattfisk (Figur 1) som föredrar sand- eller stenbotten. Födan består i huvudsak av andra bottenlevande fiskar som smörbultar och tobis, men den äter även musslor och kräftdjur (FishBase, 2015). Piggvarens utbredningsområde sträcker sig längs hela Europas kust, från Medelhavet i söder upp till polcirkeln i norr (Florin & Höglund, 2007). I Östersjön finns populationer upp till Ålands hav. Norr om Ålands hav är salthalten för låg för piggvaren (Florin & Höglund, 2007). Piggvaren i Östersjön är anpassad till den lägre saliniteten då dess ägg har en optimal utveckling vid 15 psu (Florin, 2005) till skillnad från i Nordsjön där det ligger på 20 till 35 psu (Karås & Klingsheim, 1997). Studier har visat på en viss genetisk skillnad mellan piggvaren i Nordsjön och Östersjön (Vandamme, et al., 2014).



Figur 1. Vuxen piggvar (*Scophthalmus maximus*). Foto från FishBase av Trevor Meyer enligt licens CC-BY-NC.

Kommersiellt är piggvaren en lokalt viktig resurs för kustfisket i Östersjön (Nissing, et al., 2012). Det kommersiella fisket av piggvar i Östersjön tog fart i mitten på 80-talet med de största fångsterna i mitten av 90-talet, därefter har fångsterna minskat kraftigt (Figur 2). ICES (International Council for the Exploration of the Sea) har sedan 2012 rekommenderat att fångsterna bör reduceras (ICES, 2014).



Figur 2. Total mängd landad piggvar (1000 ton) i Östersjön under 1965-2012 (ICES, 2014).

De största hoten mot piggvaren i Östersjön är idag högt fisketryck, förstörelse av uppväxtområden och syrebrist i djupare övervintringsområden på grund av minskad frekvens av stora, syrerika saltvatteninbrott från Kattegatt samt eutrofiering (HELCOM, 2013). Av dessa anledningar har HELCOM (Baltic Marine Environment Protection Commission, även

Helsinki Commission) klassat piggvaren som *Near Threatened*, det vill säga ”nära hotad” enligt Internationella naturvårdsunionens rödlistningssystem (HELCOM, 2013). De anger att det finns ett behov av en förvaltningsplan för att skydda piggvaren i vissa områden under vissa perioder. Vidare föreslår de att viktiga lek- och uppväxtområden bör skyddas från exploatering. Habitatmodellering ger en möjlighet att identifiera dessa skyddsvärda områden.

1.1 SYFTE

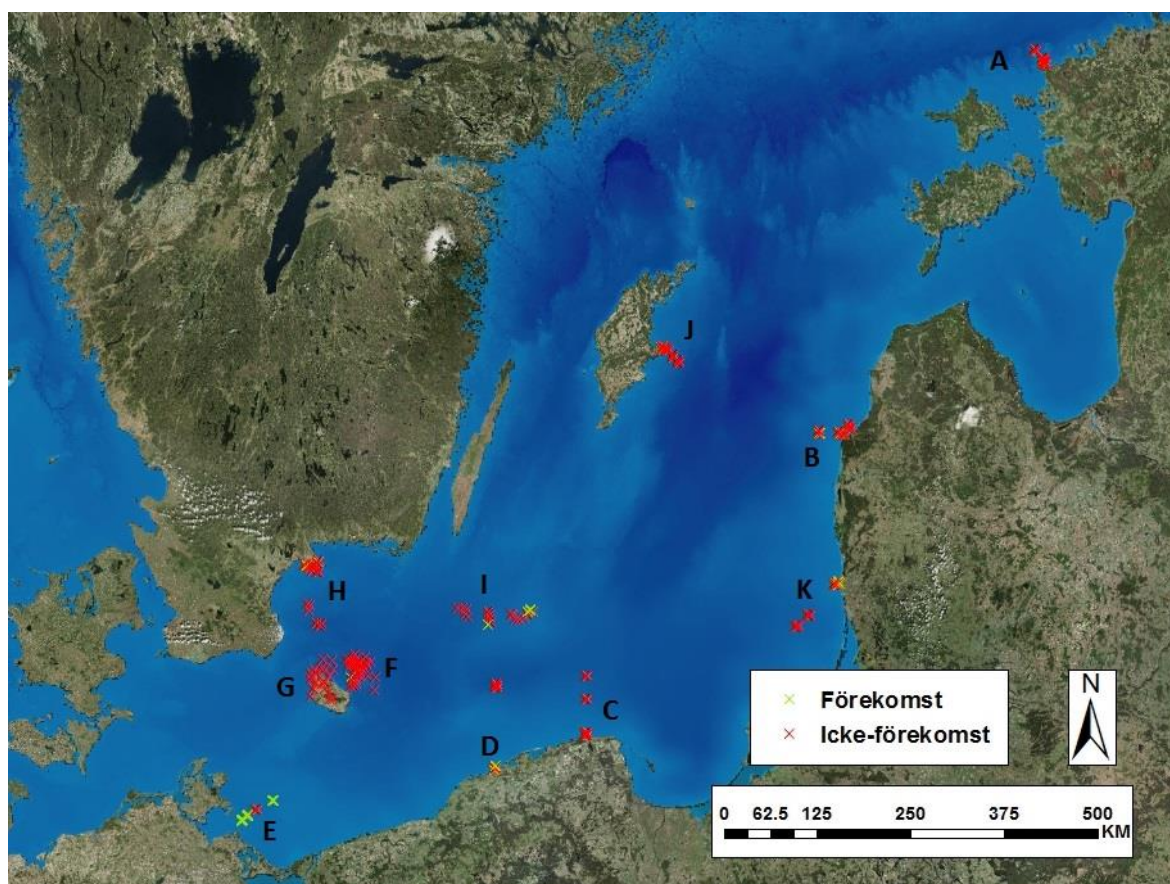
Denna studie har syftat till att ta fram och jämföra olika habitatmodeller över piggvarens förekomst i Östersjön. Ett flertal olika typer av modeller med samma indata har jämförts. De jämfördes även med ensemblemodeller som är en kombination av de framtagna modellerna. Modellerna visade vilka variabler som är viktigast för piggvarens val av habitat. Det syftar till att ge en ökad kunskap över vilka typer av bevarandearbete som effektivast gynnar piggvaren.

Ytterligare ett syfte var att undersöka användbarheten av videoanalyser. För att få en bild över hur viktigt det är att samma person utför alla klassificeringarna har studien undersökt hur personberoende skattningarna av habitatvariablerna är i videoanalyserna. Därutöver har betydelsen av de förklaringsvariablerna som är framtagna ur videoanalyserna testats i modellerna för att avgöra om videoanalyser är en användbar metod vid denna typ av modellering.

2 METOD

2.1 INSAMLING AV DATA

De data som användes för modelleringen kom huvudsakligen från BONUS-INSPIRE-projektet. Det är ett EU- och FORMAS-finansierat forskningsprojekt som syftar till att öka kunskapen om rumslik ekologi för ett effektivare bevarandearbete av arter i Östersjön (Ojaveer, et al., 2014). Projektet är ett samarbete mellan sju länder med kust mot Östersjön. I projektet har provfiske med garn skett på elva olika transekter fördelade över Egentliga Östersjön (Figur 3). Varje transekt består av stationer på olika djup mellan 5-70 meter, där varje djupnivå omfattar fem olika stationer (Ojaveer, et al., 2014).



Figur 3. Stationer i Egentliga Östersjön där data på piggvarens förekomst är insamlat från. Bokstaven anger vilket transekt stationerna tillhör och färgen anger om piggvar fångades vid stationen enligt provfiskedata från BONUS-INSPIRE. Kartan är skapad enligt licens med mjukvaran ArcGIS® av Esri.

Förutom provfiske gjordes även CTD-mätningar (konduktivitet, temperatur och djup) vid varje station. Slutligen spelades en filmsekvens av botten in vid varje station med syfte att få information angående bottensubstrat, vegetation och bottenlevande djur (Ojaveer, et al., 2014). Piggvarens habitat modellerades utifrån de provfisken som skedde på våren 2014 och 2015.

Varje station och år där piggvar fångades klassades som förekomst, övriga stationer klassades som icke-förekomst (Tabell 1).

Tabell 1. Sammanfattning över INSPIRE-projektets stationer och piggvarsförekomst. Skillnaden i antalet stationer mellan 2014 och 2015 beror på att vissa stationer togs bort på grund av avsaknad av data.

Land	Transekt	Antal stationer		Stationer med piggvar	
		2014	2015	2014	2015
Estland	A	23	25	2	1
Lettland	B	20	20	6	3
Polen	C	25	25	2	0
	D	25	24	2	5
Tyskland	E	20	20	14	17
Danmark	F	19	14	1	0
	G	13	9	0	0
Sverige	H	18	22	3	1
	I	10	3	1	1
	J	25	25	4	2
Litauen	K	25	25	9	4
Totalt	11	223	212	44	34
		435 stationer		78 stationer (17,93 %)	

Vattnets fysikaliska egenskaper mättes med en CTD-sond. De egenskaper som mättes var salinitet, temperatur, syrgashalt, djup, konduktivitet och tryck. Dessa transformerades i den mån det var möjligt till samma enhet, se kapitel 2.3. Då projektets länder hade olika typer av utrustning skiljde sig även vilka egenskaper som mättes.

I modelleringen användes även variablerna vågexponering och bottenlutning. Dessa variabler hämtades från kartraster över hela Östersjön. Vågexponeringsrastret hämtades från EMODnet (European Marine Observation and Data Network) (EMODnet, 2012). Det hade en upplösning på 25 m * 25 m och var skapat i mjukvaran WaveImpact 1.0 (Wennberg & Lindblad, 2006). Vågexponeringen beräknas utifrån medelvinden och vindens stryklängd i 16 riktningar. Rastret för havsbottens lutning hämtades från HELCOM (HELCOM, 2015). Det hade upplösningen 200 m * 200 m och lutningen mättes i procent. Utifrån mätstationernas koordinater hämtades ett värde från vågexponerings- och bottenlutningsrastret i GIS med funktionen "Extract values to point" för varje station (ArcMap, Version 10.2.2).

Alla förklaringsvariabler som användes i studien har i tidigare studier visat sig påverka förekomsten av olika fiskarter (Bergström, et al., 2011; Florin, et al., 2009).

2.2 VIDEOANALYSER

Genom att en kamera sänktes ner i vattnet vid varje station spelades videosekvenser av havsbotten in. De användes för att kategorisera havsbotten med avseende på potentiella förklaringsvariablers täckningsgrad. För att bedöma precisionen av variablerna från kategoriseringen analyserades filmer från 72 stationer separat av två personer. Detta skedde efter en

interkalibrering mellan personerna och en gemensam inläring om arter och substratstorlekar. Resultatet från dessa kategoriseringar jämfördes sedan för att avgöra om osäkerheten i bedömningen för respektive variabel var tillräckligt låg.

Täckningsgraden av havsbottnens olika substratstorlekar delades in enligt följande kategorier (Karlsson & Hansbo, 1984):

- Silt och lera (< 0,06 mm)
- Sand (0,06 – 2 mm)
- Grus (6 – 60 mm)
- Sten (60-600 mm)
- Block (> 600 mm)
- Berggrund

För modelleringen slogs dessa kategorier samman till två kategorier: mjukbotten (lera, sand och grus) och hårbotten (sten, block och berggrund). Eftersom att andelen mjuk- och hårbotten korrelerade till 100 % användes enbart mjukbottnens täckningsgrad som variabel i modelleringen.

Utöver substratstorlek användes även filmerna till att analysera bottens vegetation. Täckningsgraden av vegetationen delades in enligt följande kategorier:

- Blåstång (*Fucus sp*)
- Gaffeltång (*Furcellaria sp*)
- Övriga storväxande alger
- Ålgräs (*Zostera marina*)
- Övriga rödalger
- Trådalger
- Löst liggande alger

Då flera av dessa variabler hade låg förekomst och var i vissa fall svåra att artbestämma på grund av låg videokvalitet delades även dessa in i två kategorier. Den ena utgjordes av de strukturbildande växterna där blåstång, gaffeltång och ålgräs ingick. I den andra kategorin ingick all övrig växtlighet, samt lösdrivande alger.

De sista förklaringsvariablerna som skattades utifrån videofilmerna var täckningsgrad av bottenlevande djur. De som var möjliga att klassa var:

- Blåmusslor (*Mytilus edulis*)
- Ishavsgråsugga (*Saduria entomon*)

2.3 BEHANDLING AV DATA

Då projektets data insamlades från stationer från ett flertal olika länder med olika utrustning och förutsättningar skiljde sig mängden tillgängliga data mellan stationerna. För att modellerna skulle kunna köras krävdes det att alla stationer hade data och att alla data hade samma format. Av den anledningen gjordes ett flertal olika åtgärder för att alla data skulle vara enhetliga.

I de fall där videosekvenser saknades interpolerades data över bottensubstrat och vegetation från geografiskt närliggande stationer med liknande djup. I vissa fall saknades inspelningar

helt från en station, då kopierades data från samma station ett annat år. I några enstaka fall saknades videosekvens från alla år, då togs stationen bort från modelleringen. Ett undantag var när avsaknaden var på någon av de djupaste stationerna (70 meter), då antogs det 100 % mjukbotten och ingen vegetation (för dåliga ljusförhållanden; Wulff, et al., 1990; Kratzer, et al., 2003). Detta antagande grundar sig i att det var fallet för alla andra stationer vid dessa djup.

Vissa länder mätte tryck istället för djup. Dessa data beräknades om till djupdata (Fotonoff & Millard, 1983). På liknande sätt beräknades saliniteten om vid vissa stationer från temperatur- och konduktivitetsmätningar (Fotonoff & Millard, 1983). Även uppmätta siktdjup saknades vid vissa stationer. I dessa fall kopierades närmast liggande stations värde taget vid samma provtagningsomgång.

Syremätningar saknades vid 121 av 435 stationer. Vid ytterligare 57 stationer gav syremätningarna orimliga värden. För att inte behöva ta bort de stationer där värden saknades och på grund av den höga osäkerheten vid de andra stationerna togs variabeln bort ur modelleringen helt och hållet.

Variabeln förekomst av ishavsgråsugga togs bort ur modelleringen då den enbart förekom djupare än där den djupaste piggvaren fångades. Den förkom enbart på 18 av 435 stationer, vilket ytterligare motiverade borttagningen.

2.4 STATISTISKA ANALYSER

2.4.1 Kollinearitetstest

Om variationen mellan två eller flera förklaringsvariabler är låg har variablerna en hög multikollinearitet. Hög multikollinearitet innebär att flera variabler kan representera samma påverkan på förekomsten. Det leder till att det är svårt att avgöra vilken av de samvarierande variablerna som var den egentliga orsaken till variationen i förekomsten (Graham, 2003). För att undersöka multikollineariteten beräknades ett VIF-värde (Variance Inflation Factor). Det gjordes för alla modellens potentiella förklaringsvariabler med det inbyggda skriptet "vif()" i mjukvaran R (R version 3.2.2, 2015). Teoretiskt beräknas VIF-värdet för varje förklaringsvariabel med formeln:

$$VIF = \frac{1}{(1 - R_i^2)}$$

R_i^2 är korrelationskoefficienten i kvadrat från en linjär regression där den variabel vars VIF-värde räknas ut är responsvariabel och alla övriga variabler är förklaringsvariabler (Zuur, et al., 2010). Ett högt VIF-värde ($VIF > 3$ enligt Zuur et al, 2010) för en variabel visar på att den har hög korrelation med en eller flera av de andra variablerna och bör därmed inte användas i modellen. Roten ur VIF-värdet ger en siffra på hur många gånger större variabelns koefficients standardfel är i förhållande till om den var helt okorrelerad med de andra variablerna.

För att få en tydlig överblick över hur de återstående förklaringsvariablerna samvarierade skapades en korrelationsgraf över dessa. Korrelationsgrafen visar alla förklaringsvariablerna plottade mot varandra två och två. Det ger en lättöverskådlig bild över om det finns mönster mellan variabler. För varje par av förklaringsvariabler visas även en korrelationskoefficient

som ger ett värde på hur mycket just dessa två variabler samvarierar med varandra. Korrelationskoefficienten har ett värde mellan -1 och 1 där 0 visar på ingen korrelation. Ett negativt värde visar på en negativ korrelation. I korrelationsgrafnen ingår även ett fördelningsdiagram för att tydligt kunna se hur värdena för respektive variabel är fördelade.

2.4.2 Statistisk modellering

För att skapa en habitatmodell för piggvaren användes tio olika modelleringstekniker som alla finns implementerade i BIOMOD2-paketet i R. BIOMOD2 är en uppdaterad version av R-paketet BIOMOD som står för BIOdiversity MODelling (Thullier, et al., 2015; Thuiller, et al., 2009). Syftet med BIOMOD är att på ett enkelt sätt samla ett flertal olika modelleringstekniker i samma paket så att användaren kan använda sig av flera modelltyper samtidigt. Ytterligare en fördel med BIOMOD2 är dess möjlighet att skapa ensemblemodeller som kombinerar de olika modellerna som ingår i modellpaketet för att minimera enskilda modellers svagheter. Ensemblemodellen kan viktas så att de modeller med bäst resultat väger tyngst när de olika modellerna slås ihop. En begränsning i BIOMOD2 är att det enbart är anpassat för förekomstmodellering och inte abundansmodellering. Vid abundansmodellering tas antalet piggvarar som fångats vid varje station med i beräkningarna, medan vid förekomstmodellering får varje station ett binärt värde för om piggvar har fångats där eller inte.

Modelleringen utfördes genom att varje modell först kalibreras för att sedan valideras. Kalibreringen gick ut på att modellerna avgjorde hur mycket varje förklaringsvariabel påverkade sannolikheten för förekomst av piggvar (Thuiller, et al., 2009). I valideringen testades alla värden i kontrollgruppen utifrån deras värden på förklaringsvariablerna. Det gav ett värde för alla kontrollgruppens datapunkter på sannolikheten att piggvar förekom där. Då förekomsten var känd i kontrollgruppen kunde modellen avgöra om den skulle klassa en datapunkt som förekomst eller icke-förekomst genom att sätta ett brytvärde där andelen rätt predikterade punkter är lika hög som andelen rätt predikterade icke-förekomster. Alla datapunkter som fick ett värde över brytvärdet klassades därmed som förekomst (Thuiller, et al., 2009). Sannolikheten att en modell klassat ett positivt värde rätt kallas för sensitivitet och sannolikheten för att ett negativt värde är rätt klassat kallas för specificitet (Thullier, et al., 2015).

För att validera modellerna användes två olika metoder; externvalidering och korsvalidering. Vid externvalidering delades data permanent upp i två dataset, ett för att kalibrera modellerna och ett för att validera modellerna. Data från fyra av fem stationer i varje transekt användes till att kalibrera modellen. Data från övriga stationer användes till att validera hur bra modellerna passar in på data som den inte hade kalibrerats på. Anledningen till att kalibrerings- och valideringsdata gjordes i varje transekt var för att båda dataseten skulle få samma geografiska spridning. De data som användes till kalibrering bestod av 356 datapunkter och till valideringen användes 79 värden.

Då mängden data var begränsad till 435 datapunkter gjordes även en korsvalidering. Korsvalidering används då det inte finns några oberoende data tillgängliga för validering eller då det inte finns tillräckligt många datapunkter för att få två tillräckligt stora grupper för en effektiv kalibrering och validering (Boyce, et al., 2002). Korsvalideringen åstadkoms genom att iterativt dela in hela datasetet i två slumpmässigt utvalda delar där ena användes till kalibrering och andra till validering. I varje iteration beräknades medelfelet i klassifikationen.

För att minimera det uppskattade felet i klassificeringen beräknades slutligen ett medelvärde från alla iterationer (Franklin, et al., 2000).

2.5 BESKRIVNING AV MODELLER

BIOMOD2 innefattar tio olika typer av de mest frekvent använda modellerna inom habitatmodellering. Alla modeller har olika egenskaper och är anpassade till olika användningsområden. Generellt kan modellerna delas in i fyra olika grupper; regressionsmodeller (GLM, GAM och MARS), klassificeringsmodeller (CTA och FDA), maskinlärande modeller (ANN, GBM, RF och MAXENT) och en så kallad envelope-teknik (SRE).

2.5.1 GLM - Generalized Linear Models

GLM är en mer flexibel form av klassisk multipel regression genom att den tillåter att den beroende variabeln har andra fördelningar än normalfördelning (Thuiller, et al., 2009). Det är en viktig egenskap för en modell vid habitatmodellering då den beroende variabeln sällan är normalfördelad. Vanligt är att den beroende variabeln är *förekomst av art* som kan ha värdena 1 (förekommande) eller 0 (ej förekommande), vilket inte ger en normalfördelning.

BIOMOD2 testar modellen stegvis genom att ta bort variabler beroende på hur de påverkar modellens förklaringsgrad. Med hjälp av Akaikes informationskriterium (AIC) avgörs vilka variabler som är överflödiga för modellen och därmed kan tas bort och minska multikollineariteten (Thuiller, et al., 2009).

GLM-modellen är byggd på parametriska funktioner, vilket innebär att dess kurva följer en förinställd form, i denna studie är den kvadratisk. Nackdelen med det kan vara att den har svårt att visualisera effekten av extrempunkter.

2.5.2 GAM - Generalized Additive Models

GAM-modeller är inom ekologin en frekvent förekommande modell för att förklara förhållandet mellan en art och miljövariabler (Thuiller, et al., 2009). Den är designad att utnyttja fördelarna med GLM, men utan att behöva använda sig av parametriska funktioner. Istället använder den sig av ickeparametriska "smoothing"-funktioner (Guisan, et al., 2007).

2.5.3 MARS – Multivariate Adaptive Regression Splines

MARS är en typ av regressionsmodell. Skillnaden mellan MARS och vanliga linjära modeller är att den kan ha olika koefficienter på olika nivåer för samma variabel (Reiss, et al., 2011). Det gör MARS till en användbar metod när modellens koefficienter misstänks ha olika optimala värden på olika intervall av den förklarande variabeln. De punkter där koefficienten byter värde bestäms automatiskt av modellen (Thuiller, et al., 2009).

2.5.4 CTA – Classification Tree Analysis

CTA-modeller bygger på en optimerande algoritm som repetitivt letar efter den optimala vägen i ett beslutsträd för att förklara en arts förekomst utifrån miljövariabler (Thuiller, et al., 2009). I varje nod i beslutsträdet tas ett beslut baserat på värdet hos en förklaringsvariabel (Vayssières, et al., 2000).

CTA kan vara effektiv för att anpassa en modell till ett urval av populationen. Risken med en för bra anpassad modell är dock att den blir överparametriserad och ger dåliga resultat när den används på andra data än den anpassats för (Vayssières, et al., 2000).

2.5.5 FDA – Flexible Discriminant Analysis

FDA är en klassificeringsmodell som är anpassad för förekomst/icke-förekomstdata. Det är en utveckling av mer kända LDA (Linear Discriminant Analysis), men till skillnad från LDA kräver inte FDA att förekomstdata är normalfördelad (Reiss, et al., 2011). FDA använder sig av MARS för regressionsdelen i modellen för att förbättra klassifikationsprocessen (Hastie, et al., 1994).

2.5.6 ANN – Artificial Neural Network

ANN är en maskinlärande iterativ icke-linjär modell som på ett flexibelt sätt generaliserar linjära regressionsfunktioner. ANN är så pass flexibel tack vare att den innefattar så många olika parametrar (Thuiller, et al., 2009). Nackdelen med det stora antalet parametrar är den förhöjda risken för överparametrisering. Det undviks effektivt genom att begränsa mängden iterationer vid anpassningen med hjälp av korsvalidering (Moisen & Frescino, 2002).

2.5.7 GBM – Generalised Boosted Models

GBM är en metod som bygger på resultaten från ett stort antal enklare modeller. Varje individuell modell består av klassifikations- eller regressionsträd. Genom att iterativt addera ”träd” till modellen och vikta om de data som tidigare gav ett dåligt värde, byggs den slutgiltiga modellen upp (Reiss, et al., 2011).

2.5.8 RF – Random Forest

RF är en maskinlärande ensemblemodell som generar ett stort antal klassifikationsträd (500 i denna studie) (Cutler, et al., 2007). Responsen i modellen predikteras sedan beroende på hur majoriteten av dessa träd har predikerat (Thuiller, et al., 2009). RF är lättare att använda och ger mer precisa prediktioner än traditionella klassifikationsträd. (Bucas, et al., 2013).

2.5.9 MAXENT – Maximum Entropy

MAXENT är en relativt ny modelleringsmetod som har använts i stor utsträckning inom ekologisk modellering sedan den släpptes (Merow, et al., 2013; Phillips, et al., 2006). Det är en modell som bara är baserad på om en viss art förekommer eller inte vid varje mätstation. Antalet förekomster vid varje station spelar alltså ingen roll för MAXENT. Målet med MAXENT är att hitta sannolikhetsfördelningar som uppfyller en uppsättning begränsningar härledda från förekomstdata. Varje begränsning förutsätter att väntevärdet för en variabel måste vara inom konfidensintervallet för dess empiriska medelvärde. MAXENT väljer den fördelning som maximerar entropin av alla fördelningar som uppfyller kraven (Guisan, et al., 2007).

2.5.10 SRE – Surface Range Envelope

SRE är en metod som kommer från BioClim, vilket var en av de första mjukvarorna för artdistributionsmodellering som fick genomslag bland ekologer (Booth, et al., 2013). Metoden går ut på att identifiera maximum- och minimumvärden för alla variabler där den undersökta arten förekommer. Enbart de stationer vars alla förklaringsvariabler ligger inom detta maximum-minimum-spann inkluderas i modellen. Modellen kan ställas in utifrån den percentilfördelning som önskas. På så sätt kan stationer vars förklaringsvariabler ligger nära gränsvärdena undvikas (Thuiller, et al., 2009).

2.5.11 Ensemblemodeller

En ensemblemodell är en kombination av flera modeller. I BIOMOD2 finns det sex olika metoder för att kombinera önskade modeller till en gemensam ensemblemodell (Thuiller, et

al., 2015). Den första metoden beräknar sannolikheten för förekomst för varje punkt i prediktionen genom att beräkna ett medelvärde från alla ingående modellernas sannolikhet. Två av metoderna utgår från konfidensintervallet runt medelvärdet. Den ena är baserad på sannolikheten att ett värde ligger högre än den övre gränsen i konfidensintervallet och den andra att värdet ligger under den nedre gränsen. En fjärde metod använder modellernas medianvärde istället för medelvärdet. Den femte metoden kallas *Comittee Average* (CA). Den avgör att en station ska klassas som förekomst om majoriteten av de ingående modellerna har klassat den som förekomst. Den sista metoden som används är ett viktat medelvärde. Hur en modell viktas beror på dess AUC-värde (se nästa kapitel).

2.6 MODELLUTVÄRDERING

För att avgöra hur bra en modell presterar producerades en ROC-kurva (Receiver Operating Characteristic Curve). En ROC-kurva är en graf som visar sambandet mellan modellens sanna förekomster (sensitiviteten) och falska positiva värden (1-specificiteten) längs ett spann av brytvärden (Thuiller, et al., 2009). Ett brytvärde är det värde på en prediktion där modellen avgränsar om sannolikheten är tillräckligt hög för att predikteras som förekomst eller icke-förekomst. Alla värden över brytvärdet är klassad som förekomst.

För att kunna jämföra olika modeller och få ett värde på hur väl modellerna presterar beräknades arean under ROC-kurvan (AUC). En perfekt modell som predikterar alla värden rätt har ett AUC-värde på 1. En dålig modell som är helt slumpartad har lika många falska positiva som sanna positiva värden och ett AUC-värde på 0.5. Ett index på hur AUC ska klassas är enligt Thuillier, et al., 2009 att värden $< 0,7$ är dåliga, 0,7-0,8 klassas som godkänt 0,8-0,9 är bra resultat och 0,9-1 är mycket bra. De modeller som hade ett AUC-värde $< 0,7$ uteslöts från ensemblemodellerna.

3 RESULTAT

3.1 VIDEOANALYSER

En jämförelse av videoanalyserna från två olika personer som analyserat samma filmer visade generellt små skillnader i skattningarna (Tabell 2). Den procentuella avvikelsen visade hur mycket varje skattning skiljer sig ifrån skattningarnas medelvärde i snitt.

Tabell 2. Jämförelse av videoanalyser. Två personer (Analys 1 och 2) hade oberoende av varandra analyserat och kategoriserat 72 filmer av havsbotten. För att bedöma svårigheten i kategoriseringen har de två analyserna jämförts från varje film. Analys 1 och 2 visar procentuell genomsnittlig förekomst av respektive variabel. Medelavvikelse visar hur mycket varje analys i snitt avviker från de två analysernas medelvärde.

	Mjukbotten	Strukturbildande växter	Övriga alger	Blåmussla
Analys 1 (%)	83,96	1,11	10,13	1,60
Analys 2 (%)	81,25	1,39	9,15	2,64
Medelavvikelse (%)	7,26	1,59	9,08	4,47

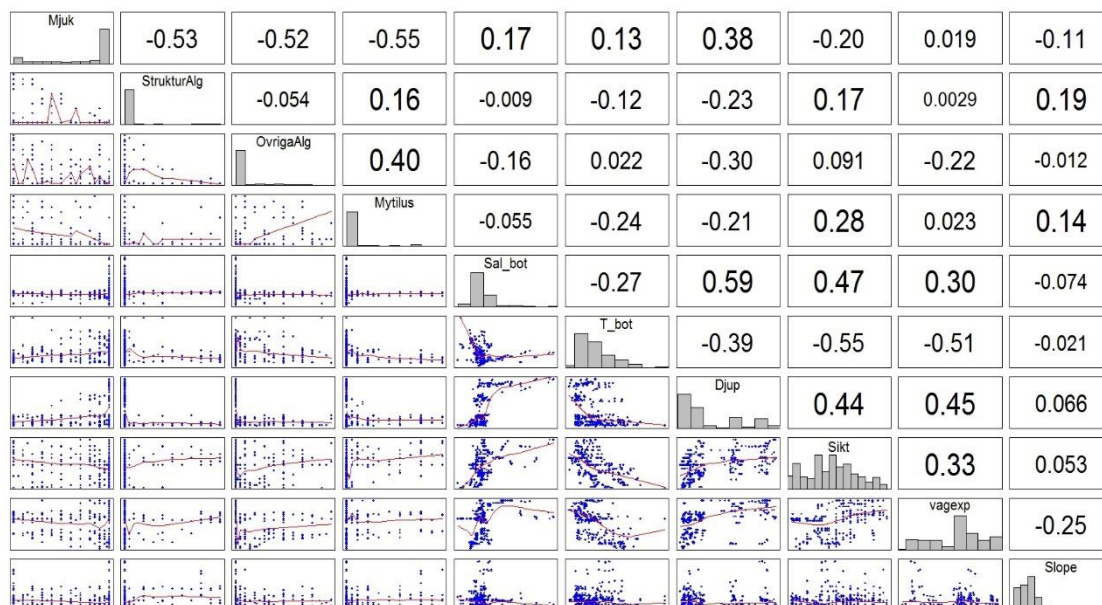
3.2 KOLLINEARITET

Det första VIF-testet (VIF_1) visade på att variabler med både botten- och ytvärden hade högst kollinearitet (Tabell 3). Av den anledningen togs ytvariablerna bort (VIF_2) vilket ledde till att kollineariteten för alla övriga variabler minskade, i vissa fall markant.

Tabell 3. VIF-test på alla förklaringsvariabler. De värden som saknas visar vilka variabler som uteslöts efter det första VIF-testet.

Förklaringsvariabel	Min-värde	Max-värde	VIF_1	VIF_2
Mjukbotten (%)	0	100	3,31	3,13
Strukturbildande växter (%)	0	100	2,15	2,05
Övriga alger (%)	0	100	2,17	1,96
Mytilus (%)	0	90	1,74	1,72
Bottensalinitet (psu)	4,06	18,83	2,15	1,78
Bottentemperatur (°C)	3,15	17,4	4,55	1,83
Ytsalinitet (psu)			3,32	-
Yttemperatur (°C)			4,49	-
Djup (m)	1,88	76,37	3,55	2,86
Siktdjup (m)	15	1,8	2,31	2,10
Vågexponering (m ² /s)	171928	697728	2,71	1,86
Bottenlutning (%)	0	2,7	1,45	1,31

Kollinearitetsgrafen visar på att flera av förklaringsvariablerna korrelerade till viss del med varandra (Figur 4). Tydligast korrelation hittades mellan mjukbotten - blåmusslor, botten-salinitet - djup och bottentemperatur - siktdjup. Det andra VIF-värdet ansågs ge tillräckligt låga värden ($\sim VIF < 3$) för att alla kvarvarande variabler skulle vara med i modelleringen (Zuur, et al., 2010).



Figur 4. Korrelationsgraf för alla stationer ($n=435$) med alla modellernas förklaringsvariabler. Histogrammet på diagonalen visar förklaringsvariabelns fördelning. Graferna till vänster visar två förklaringsvariabler plottade mot varandra. Siffrorna till höger visar R -värdet från korrelationerna. Den röda linjen är ett flytande medelvärde och har inget med R -värdet att göra.

3.3 MODELLRESULTAT

3.3.1 Förklaringsvariabler

Förklaringsvariablernas betydelse visade på att djupet följt av bottentemperatur och vågexponering var de tre viktigaste förklaringsvariablerna för piggvarens val av habitat (Tabell 4). Alla tio modellerna klassade djupet som en av de tre viktigaste variablerna. Responskurvorna visade på ett negativt samband mellan djupet och piggvärsförekomst (Figur 5 - 6). I responskurvorna från ensemblemodellerna sågs en tydlig minskning av förekomsten vid djup under 20 meter (Figur 6).

Även bottentemperaturen stack ut som en viktig förklaringsvariabel (Tabell 4). I flera av modellerna klassades den som den viktigaste. Här visade responskurvorna på en tydlig positiv korrelation med piggvärsförekomst. Störst sannolikhet för piggvärsförekomst var vid bottentemperaturer över 10 grader.

Vågexponeringen, som klassades som en av de tre viktigaste variablerna i sex av modellerna, visade en tydlig trend i responskurvorna. En ökad vågexponering gav en ökad sannolikhet för piggvärsförekomst ända till de mest vågexponerade stationerna där sannolikheten minskade igen.

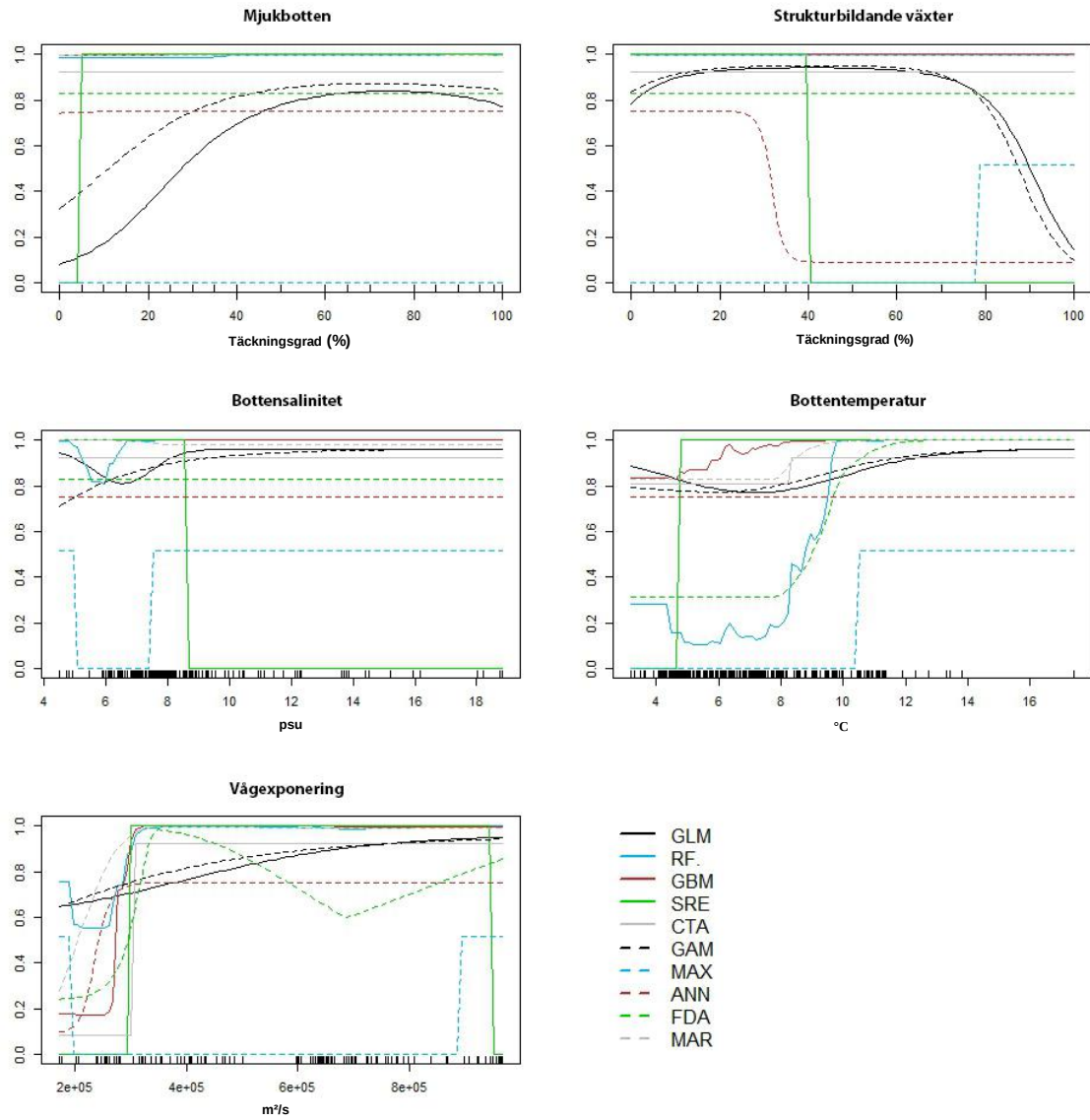
Ytterligare tre variabler klassades som en de tre viktigaste hos någon modell, det var förekomst av blåmusslor, bottensalinitet och bottenlutning. Den av dessa som visade tydligast trend i responskurvorna var bottenlutning. För bottenlutningen var korrelationen negativ. Sannolikheten för piggvärsförekomst minskade mellan lutningar på 0-1 % därefter höll den sig oförändrad på stationer med en lutning > 1 %. Bottensalinitetens responskurvor skiljde

sig mellan de olika modellerna. Generellt visade de på mycket små variationer i sannolikheten till piggyvarsförekomst. Blåmusslor hade en negativ inverkan på förekomsten av piggyvar. Alla modeller hade högst sannolikhet för piggyvarsförekomst vid 0 % täckningsgrad av blåmusslor.

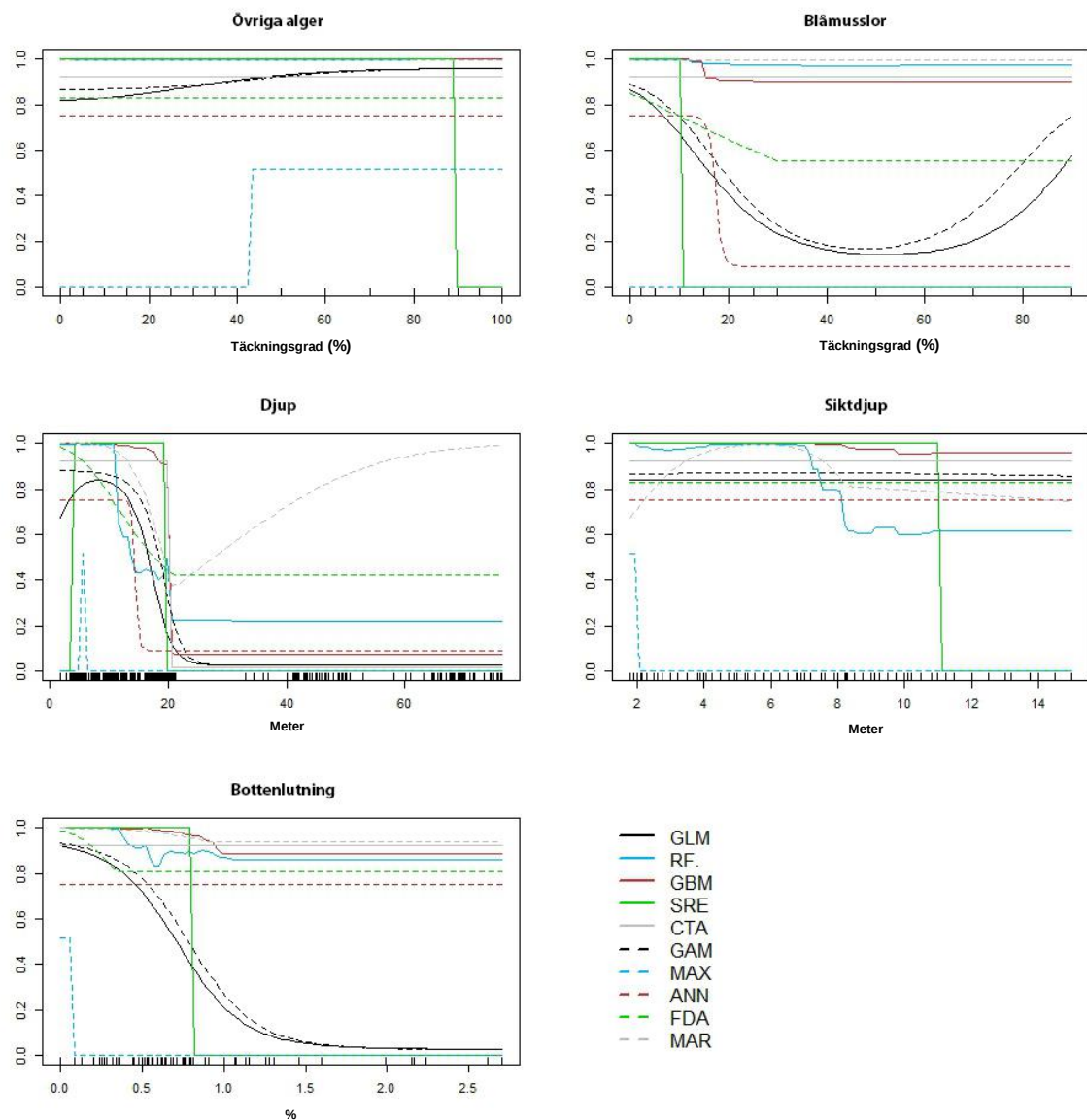
Strukturbildande växter och övriga alger visade på ingen eller låg korrelation med piggyvaren. För dessa variabler visade responskurvorna inte heller på några tydliga trender, med undantag från ett par modeller som visade på att strukturbildande växter hade en negativ korrelation för de stationer där de strukturbildande växterna upptog en yta på över 40 %.

Tabell 4. Förklaringsvariablernas betydelse för alla olika modelltyper. För varje modell har de tre viktigaste variablerna markerats med asterisker där tre asterisker visar på den viktigaste variabeln. Påverkan per modell visar hur stor inverkan varje förklaringsvariabel hade på modellerna i snitt. Det totala värdet för variablernas inverkan är olika för alla modeller vilket innebär att värdena bara är jämförbara inom varje modell. Snittpåverkan visar i snitt hur mycket varje variabel påverkar i varje modell. Rank är en rangordning av variablernas påverkan baserat på snittpåverkan.

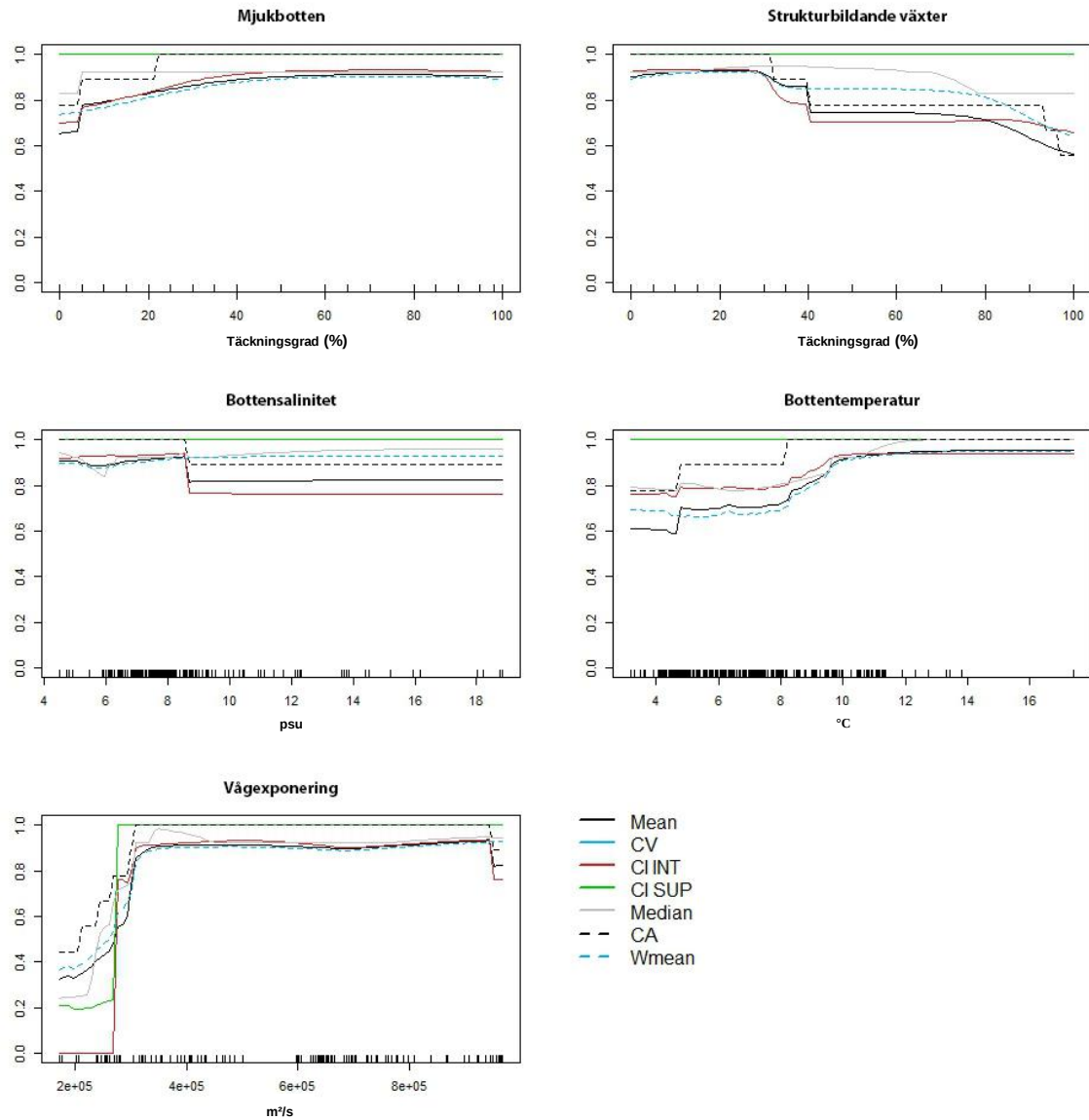
Variabel	GLM	RF	GBM	SRE	CTA	GAM	MAX- ENT	ANN	FDA	MARS	Snittpåverkan (%)	Rank
Mjukbotten	0,129	0,009	0,013	0,062	0	0,069	0,092	0,123	0	0	3	8
Strukturbildande växter	0,045	0	0	0,052	0	0,053	0,065	0,143	0	0	2,1	9
Övriga alger	0,085	0,015	0,001	0,008	0	0,058	0,166	0,005	0	0	1,8	10
Blåmusslor	0,121	0,005	0,05	0,175	0,101	0,122 *	0,224	0,253 **	0,007	0	6,8	4
Bottensalinitet	0,149 *	0,05	0,027	0,123	0,189	0,04	0,395 *	0,003	0	0,125	6,3	5
Bottentemp	0,044	0,173 ***	0,297 **	0,191 *	0,525 ***	0,075	0,457 **	0,006	0,736 ***	0,54 ***	22,3	2
Djup	0,808 ***	0,172 **	0,339 ***	0,423 ***	0,424 **	0,789 ***	0,539 ***	0,868 ***	0,091 *	0,296 *	31,3	1
Siktdjup	0	0,081	0,064	0,147	0,087	0	0,39	0	0	0,191	5,9	7
Vågexponering	0,117	0,093	0,237 *	0,203 **	0,283 *	0,077	0,385	0,206 *	0,251 **	0,355 **	15	3
Bottenlutning	0,187 **	0,047 *	0,05	0,089	0,108	0,183 **	0,293	0,002	0,021	0,038	6,2	6



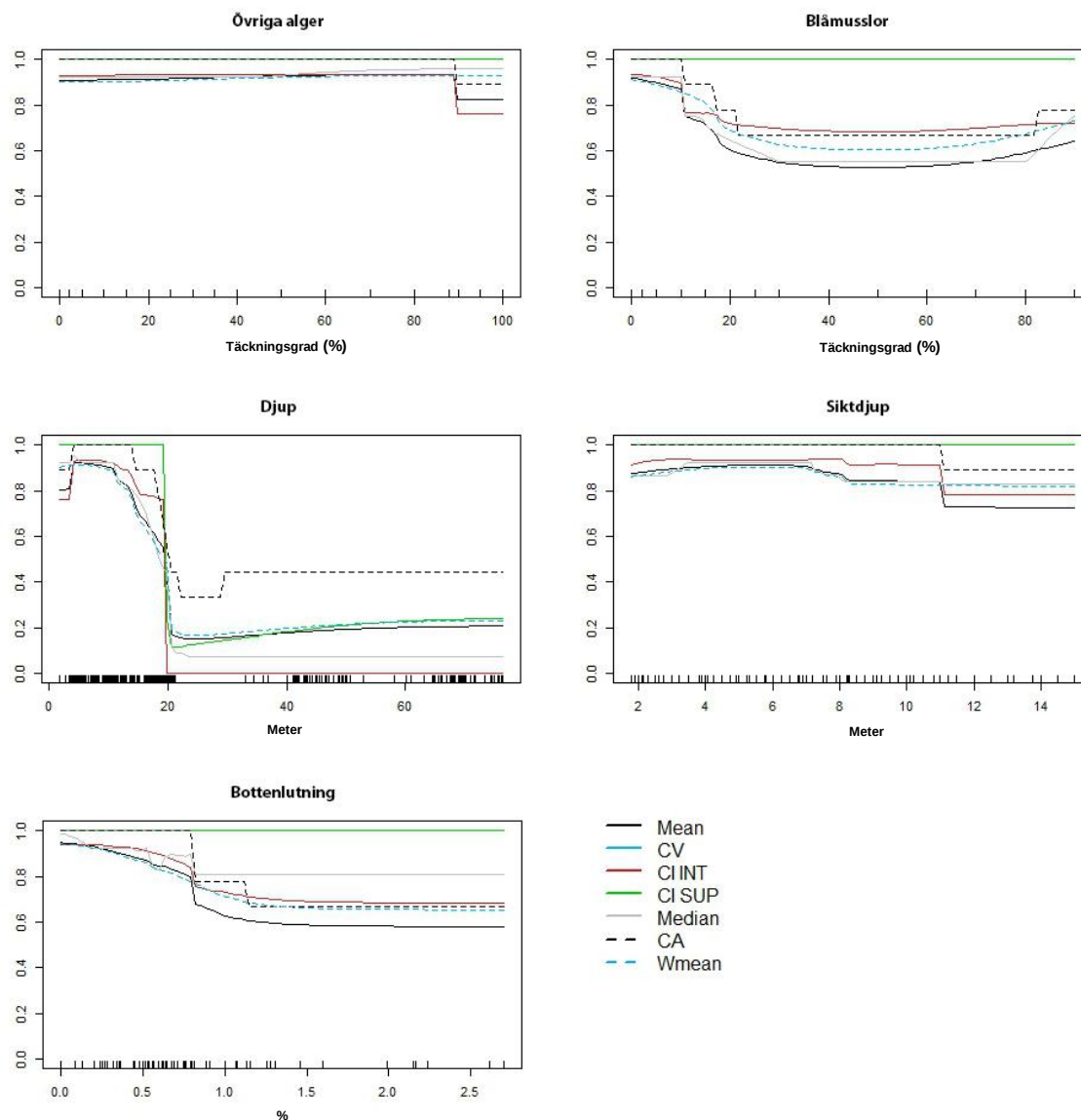
Figur 5a. De enskilda modellernas responskurvor för mjukbotten, strukturbildande växter, bottensalinitet, bottentemperatur och vågexponering. Y-axlarna visar sannolikheten för piggarförelkomst.



Figur 5b. De enskilda modellernas responskurvor för övriga alger, blåmusslor, djup, siktdjup och bottenlutning. Y-axlarna visar sannolikheten för piggyvarsförekomst.



Figur 6a. Ensemblemodellernas responskurvor för mjukbotten, strukturbildande växter, bottensalinitet, bottentemperatur och vågexponering. Y-axlarna visar sannolikheten för piggarförelkomst.



Figur 6b. Ensemblemodellernas responskurvor för övriga alger, blåmusslor, djup, siktdjup och bottenlutning. Y-axlarna visar sannolikheten för piggvarsförekomst..

3.3.2 Modellutvärdering

Modelleringen med korsvalidering gav mycket lika resultat i förhållande till modelleringen med separata data för validering och kalibrering. Av den anledningen visas enbart resultat från modellerna med separata kalibrerings- och valideringsdata.

Modellerna visade generellt på ett bra resultat där alla hade ett AUC-värde $> 0,7$ förutom MAXENT-modellen (Tabell 5). Den modell med det högst uppmätta AUC-värdet var RF som hade närmast perfekta modeller ($AUC = 0,997$) på kalibreringsdata. När den testades på valideringsdata fick den ett sämre resultat. Att en modell har stor skillnad på AUC-värdet mellan kalibrerings- och valideringsdata tyder på att den överparametriserar data (Thuiller, et al., 2009). Utöver RF stack CTA- och GBM-modellerna ut med AUC-värden över 0,9 på alla körningar.

En jämförelse hur de olika modelleringsmetoderna förhöll sig till varandra återfinns i Bilaga 1 och 2. GLM och GAM visar på en stor likhet då jämförelsen visar en tydlig diagonal trend i graferna. Jämförelsegrafén illustrerar också hur bra modellerna klassificerade data, hur de lyckades separera de gröna förekomstpunkterna från de röda icke-förekomstpunkterna. Grafén visar att de modeller med högst AUC-värde (RF, GBM och CTA) är de modeller som bäst klassificerade förekomst och icke-förekomstdata i två olika grupper.

MAXENT var den modell som hade lägst sensitivitet. Den predikterade bara 38,9% av alla stationer med förekomst av piggvar. GAM var den enda modellen som prickade in alla stationer med piggvarsförekomst. Den hade istället lägst specificitet, det vill säga att den predikterade in lägst antal rätt icke-förekomster.

Inför ensemblemodelleringen viktades de olika modellerna beroende på dess externa AUC-värde (Tabell 5). MAXENT-modellen hade för lågt AUC-värde ($< 0,7$) för att ingå i ensemblemodellen. Även SRE-modellen ströks då den metoden inte kan användas i en ensemblemodell.

Tabell 5. Utvärdering av de olika modellerna. De bästa modellerna med avseende på AUC har markerats med asterisker, där tre stycken markerar den bästa modellen. Viktningsvärdet är baserat på AUC och beskriver hur mycket respektive modell bidrar till ensemblemodellerna. Brytvärdet anger vid vilken sannolikhet i procent av piggvarsförekomst där det är lika stor andel rätt predikterade förekomster som rätt predikterade icke-förekomster.

Modell	AUC		Brytvärde	Sensitivitet	Specificitet	Viktning
	Kalibrering (n=356)	Validering (n=79)				
GLM	0,91	0,86	37,8	88,9	70,5	0,126
RF	0,99***	0,83	27,5	72,2	85,2	0,123
GBM	0,98**	0,91***	1,5	88,9	73,8	0,133
SRE	0,74	0,76	50,0	72,2	80,3	-
CTA	0,95*	0,90**	76,3	77,8	95,1	0,133
GAM	0,91	0,86	16,6	100	62,3	0,126
MAX- ENT	0,79	0,65	25,8	38,9	90,2	-
ANN	0,79	0,79	47,0	88,9	72,1	0,115
FDA	0,87	0,80	33,7	66,7	93,4	0,117
MARS	0,91	0,87*	53,6	72,2	91,8	0,127

Ensemblemodellerna fick alla mycket bra resultat med AUC-värden mellan 0,945 – 0,980 (Tabell 6). Alla ensemblemodeller presterade bättre än den bästa enskilda modellen. Den ensembleteknik som presterade bäst var när viktade medelvärden från de modellerna hade använts. Tabell 6 visar även att ökad sensitivitet gav en minskad specificitet och tvärt om. Den bästa modellen hade varken högst sensitivitet eller specificitet.

Tabell 6. Modellutvärdering för ensemblemodellerna. De bästa modellerna har markerats med asterisker, där tre stycken markerar den bästa modellen. Brytvärdet anger vid vilken sannolikhet i procent av piggvarsförekomst där det är lika stor andel rätt predikterade förekomster som rätt predikterade icke-förekomster.

Viktningsmetod	AUC	Brytvärde	Sensitivitet	Specificitet
Medel	0,97**	23,9	100	72,1
CI Inf	0,97**	36,8	66,7	93,4
CI Sup	0,97**	18,9	100	70,5
Median	0,97**	11,9	100	68,9
CA	0,95	27,8	100	70,5
Viktat medel	0,98***	27,1	91,4	73,8

4 DISKUSSION

Alla modeller utom MAXENT visade på en god förmåga att modellera pigghvarhabitat ($AUC > 0,7$). De enskilda modellerna hade generellt höga AUC-värden, där GBM och CTA stack ut med $AUC > 0,9$. GBM och CTA skiljde sig ganska mycket från varandra i specificitet och sensitivitet. CTA hade högre specificitet, det vill säga att den var bättre på att pricka in icke-förekomststationer rätt. GBM hade istället högre sensitivitet och prickar därmed in högre andel stationer med förekomst. Vilken metod av dessa som är viktigast beror därmed på om precisionen av förekomster eller icke-förekomster är viktigast.

De goda resultaten från modellerna tyder på att denna typ av modellering av pigghvarhabitat med fiske- och miljödata är användbar. Även om de flesta modeller gav ett bra resultat så har alla fortfarande svagheter. Ensemblemodellernas styrka att eliminera vissa av de enskilda modellernas svagheter sågs tydligt i Tabell 6. När de enskilda modellerna kombinerades med varandra fick de gemensamma ensemblemodellerna ett betydligt högre AUC-värde än de bästa enskilda modellerna. Fyra av ensemblemodellerna lyckades klassa alla förekomster rätt, men hade det svårare med icke-förekomster.

De bästa modellerna visade att piggvaren på våren framförallt föredrar grunda områden med höga temperaturer och hög vågexponering. Täckningsgrad av mjukbotten och vegetation hade däremot liten påverkan på piggvarens habitatval. Vid djup under 20 meter minskade sannolikheten till pigghvarförekomst markant (Figur 6a). Tidigare studier har kommit fram till att på våren i Östersjön lever piggvaren på djup ner till 30 meter, vilket stämmer bra med modellernas resultat (Voigt, 2002). Denna studie hade inga mätstationer mellan 20-40 meters djup vilket kan vara anledningen att sannolikheten för förekomst minskade markant under 20 meter och inte vid 30 meter som den tidigare studien föreslår. Även botten temperaturen följde tidigare studier. Piggvaren är ansedd att vara en förhållandevis sent invandrad varmvattensfisk i Östersjön som anpassar sitt habitat under årstiderna efter vattentemperatur (Voigt, 2002). Samtliga modeller gav en ökad sannolikhet för pigghvar med ökad temperatur. Endast habitatvalen för vuxen pigghvar i Östersjön på våren (april - maj) undersöktes och inga slutsatser kan därför dras om piggvarens habitatval övriga tider på året eller i andra delar av dess liv. På våren ansamlas piggvaren inför lek vilket kan innebära att den anträffas på eller i närheten av lämpliga uppväxtområden för juveniler.

Anledningen till vilken modell som presterade bäst är en komplex fråga som kan ha många svar. I tidigare studier har skillnader i modellernas prestation visat sig bero på mängden av datapunkter valda i modelleringen (Barbet-Massin, et al., 2012). I den studien presterade GBM och CTA bäst när antalet datapunkter var i storleksordningen 100-tal. Det stämmer bra överens med denna studie, vilket kan innebära att storleken på datasetet kan vara en av anledningarna till att just dessa modeller presterade bäst. En annan studie har kommit fram till att oförutsägbarheten i arters distribution ofta överskrider skillnader i modelleringsteknik (McPherson & Jetz, 2007). Det tyder på att osäkerheten i vilken metod som är bäst är ganska hög.

Analyser av havsbotten med hjälp av videosekvenser istället för dykning är en relativt billig metod (Dimming & Lindegarth, 2012). Det innebär en kostnadseffektiv metod för att följa den övervakning och uppföljning som EU:s art- och habitatdirektiv kräver. Fler fördelar med videoanalyserna är att de kan ske djupare och med mindre säkerhetsrisk än dykning. För att

metoden ska vara användbar inom habitatmodellering är det viktigt att analyserna speglar verkligheten och är repeterbara både inom och mellan personer. Tidigare studier har jämfört videoanalyser med dykanalyser genom att låta samma person skatta täckningsgraden av växtligheten på botten både efter dykning och via videoanalys (Sundblad, et al., 2012). Skillnaden mellan metoderna för att skatta vissa specifika arter var relativt stora, medan en mer generell skattning av vegetationen totalt gav en liten eller ingen skillnad mellan metoderna. Då denna studie hade en ganska bred indelning av vegetationen bedöms videoanalys vara en lämplig metod. Repeterbarheten i metoden testades genom att låta två oberoende personer skatta samma filmer (Tabell 2). De testerna visade på små skillnader mellan skattningarna, framförallt för substratstorlek.

Bildkvaliteten på filmsekvenserna skiljde sig kraftigt mellan ländernas filmklipp. Med dålig bildkvalitet finns det stora risker för felskattningar, speciellt när det gäller vegetationen. Alla länder hade även skilda strategier och kameraupphängningsanordningar vid filmningen. Detta innebär att filmsekvenserna var filmade i olika vinklar och på varierande avstånd till botten. Det kan ge en orättvis skattning mellan länder då täckningsgraden av vegetation kan uppfattas annorlunda i olika vinklar och avstånd.

Från videosekvenserna hämtades fyra av totalt tio variabler. Tre av dessa var de variabler med lägst genomsnittlig påverkan på modellerna (se Tabell 4). Den variabel från videoanalysen som var viktigast i modelleringen var förekomst av blåmusslor. Det var även den variabel som var svårast att skatta utifrån filmerna och den som hade störst skillnader mellan två oberoende analyser (Tabell 2). Det innebär en ökad osäkerhet i resultatet. Sannolikheten för piggvarsförekomst minskade när mängden blåmusslor ökade (se Figur 5a). En anledning till den negativa korrelationen med piggvarsförekomst kan vara blåmusslornas positiva korrelation med variabeln övriga alger. Tidigare studier har visat på en negativ korrelation mellan filamentösa alger och juvenil piggvarsförekomst (Florin, et al., 2009; Jokinen, et al., 2016). De filamentösa algerna ingår i övriga alger i denna studie. Därmed finns det en möjlighet att anledningen till att piggvaren inte trivs i närheten av blåmusslorna är på grund av att de lever vid de filamentösa algerna. Resultatet av de övriga tre variablerna från videoanalysen tyder på att bottensubstrat och vegetation hade liten påverkan på piggvarens val av habitat. Det talar emot tidigare studier som säger att piggvaren trivs på sand och grusbotten (Florin, 2005). Anledningen till skillnaden kan vara grupperingen av de olika substratstorlekarna. Eftersom att sandbottenarna som ofta var grunda, var i samma grupp som lerbottenarna som generellt var djupare, kan variabelns betydelse ha tappat värde. Genom att använda artspecifika förklaringsvariabler istället för breda artgrupper hade resultatet eventuellt sett annorlunda ut. Det skulle dock behövas bättre bildkvalitet för att den typen av indelning skulle vara aktuell. För att bottenanalyserna ska få högre precision, ökad repeterbarhet och därmed även en ökad användbarhet behövs en standardiserad metod för hur kamerariggen och filmningen ska ske. Det är även viktigt att använda en kamera som klarar av bottendjupens ansträngda ljusförhållanden.

För att få en bredare överblick över piggvarens potentiella habitat i hela Östersjön är nästa steg att skapa kontinuerliga kartor med prediktioner av piggvarens habitat. Med hjälp av kartprediktioner är det möjligt att få en bild av var de effektivaste ställena att lägga resurser på vid en eventuell bevarandeinsats. Dessa kartprediktioner kräver att det finns kontinuerliga data för varje förklaringsvariabel över hela det önskade området. Detta var inte aktuellt i

denna studie då variablerna från videoanalyserna inte var tillgängliga för hela Egentliga Östersjön. Däremot finns det möjlighet att skapa en sådan prediktion med flera av de variabler som visade sig ha störst påverkan på sannolikheten för piggvarsförekomst. Då dessa variabler var viktiga skulle kvaliteten på en eventuell kartprediktion vara betydande.

5 SLUTSATSER

Studien visade på att modellering av piggvarens habitat utifrån provfiske och miljödata är möjlig att genomföra på ett framgångsrikt sätt med ett flertal olika modelltyper. Alla studiens enskilda modeller utom MAXENT gav bra resultat ($AUC > 0,7$), där CTA och GBM var bäst med $AUC > 0,9$. Genom att kombinera modellerna med bra resultat till ensemblemodeller blev modellresultaten ännu bättre. Att alla ensemblemodeller presterade markant bättre än den bästa enskilda modellen tyder på att ensemblemodellering är en effektiv metod för att utnyttja de enskilda modellernas styrkor och minska effekten av deras svagheter. De bra resultaten indikerar på att det finns goda möjligheter att ta fram kartprediktioner från modellerna.

Modellerna säger att vuxen piggvar på våren i Östersjön föredrar grunda vatten med hög vågexponering och hög temperatur. De fem viktigaste variablerna var i turordning: djup, bottentemperatur, vågexponering, förekomst av blåmusslor och bottensalinitet.

Studien visade också att videoanalyser av havsbotten kan vara en bra metod för att kategorisera in havsbotten i potentiella förklaringsvariabler. Precisionen på klassificeringen är beroende av videosekvensernas bildkvalitet och filmteknik. Repeterbarheten av analyserna mellan två personer som har gemensamt kalibrerat sig var god.

6 REFERENSER

ArcMap, Version 10.2.2, 2014. *ArcMap Version 10.2.2.3552 [Computer program]*, Esri Inc. .

Arujo, M. B., 2007. Ensemble forecasting of species distributions. *Trends in Ecology & Evolution*, 22(1), pp. 42-47.

Barbet-Massin, M., Jiguet, F., Albert, C. H. & Willfried, T., 2012. Selecting pseudo-absences for species distribution models: how, where and how many?. *Methods in Ecology and Evolution*, 3(2), pp. 327-338.

Bergström, U., Bergström, L., Carlén, I. & Isaeus, M., 2011. *GIS-baserade metoder för att kartlägga fiskars livsmiljöer i grunda havsområden*, Stockholm: Naturvårdsverket.

Bergström, U. et. al., 2013. Evaluating eutrophication management scenarios in the Baltic Sea using species distribution modelling. *Journal of Applied Ecology*, 50(3), pp. 680-690.

Booth, T. H., Nix, H. A., Busby, J. R. & Hutchinson, M. F., 2013. BIOCLIM: the first species distribution modelling package, its early applications and relevance to most current MaxEnt studies. *Diversity and Distributions*, 20(1), pp. 1-9.

Boyce, M. S., Vernier, P. R., Niealsen, S. E. & Schmiegelow, F. K., 2002. Evaluating resource selection functions. *Ecological Modelling*, 157(2-3), pp. 281-300.

Bucas, M. et. al., 2013. Empirical modelling of benthic species distribution, abundance, and diversity in the Baltic Sea: evaluating the scope for predictive mapping using different modelling approaches. *ICES Journal of Marine Science*, 70(6), pp. 1233-1243.

Cutler, R. D. et. al., 2007. Random forest for classification in ecology. *Ecology*, 88(11), pp. 2783-2792.

Dimming, A. & Lindegarth, M., 2012. *Utveckling av metodbeskrivning vid övervakning, uppföljning och kartläggning av habitat och habitatbildande arter med undervattensvideo*, Uddevalla: Hav möter Land, Länsstyrelsen i Västra Götalands län.

Elith, J. et. al., 2006. Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29(2), pp. 129-151.

EMODnet, 2012. *EMODnet Seabed Habitats interactive map*. [Online]
Available at: <http://www.emodnet-seabedhabitats.eu/download>
[Använd 16 11 2015].

FishBase, 2015. www.fishbase.org, version (10/2015).. [Online]
Available at:
<http://www.fishbase.org/summary/SpeciesSummary.php?ID=1348&AT=turbot>
[Använd 28 01 2016].

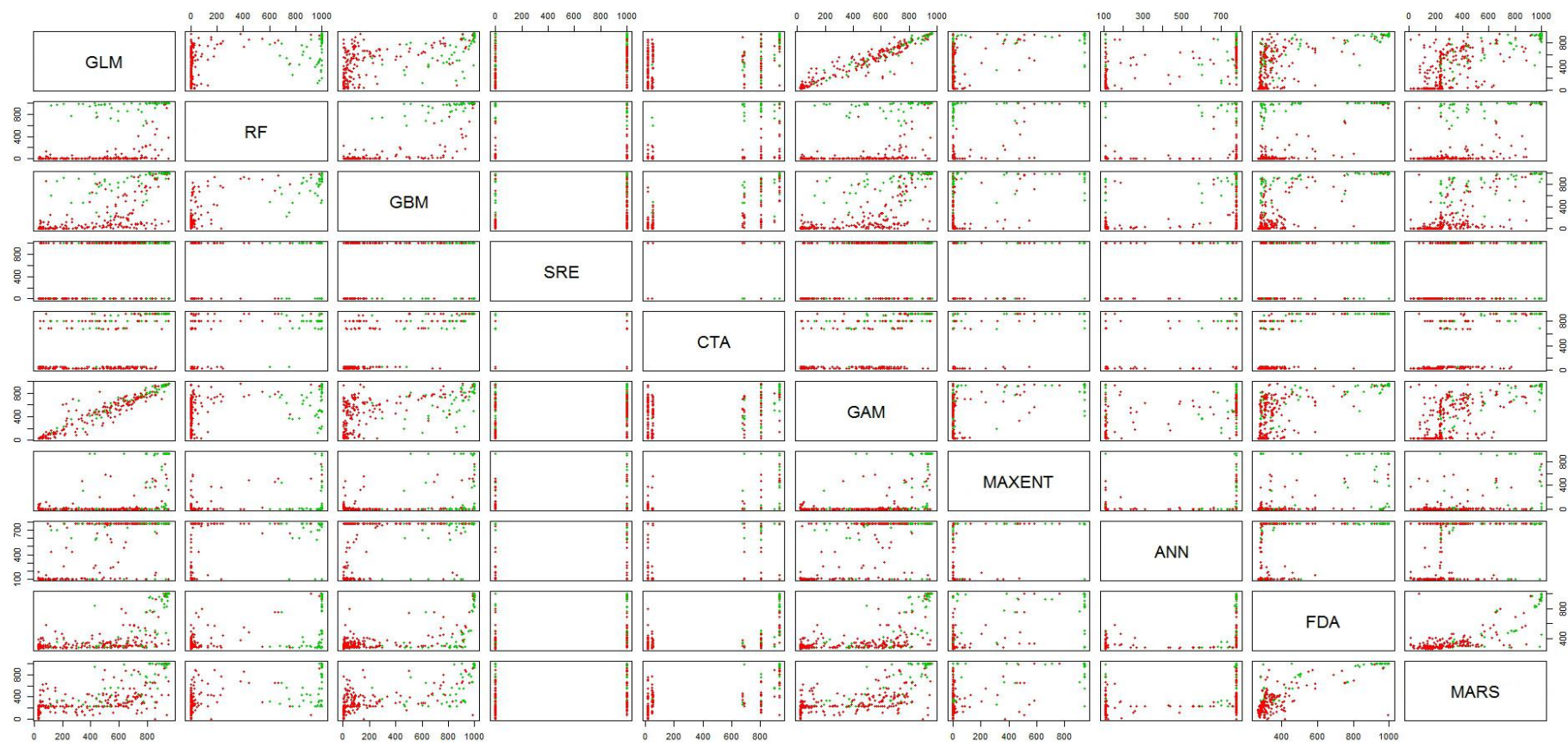
Florin, A.-B., 2005. *Flatfishes in the Baltic Sea – a review of biology and fishery with a focus on Swedish conditions*, Öregrund: Fiskeriverket.

Florin, A.-B. & Höglund, J., 2007. Absence of population structure of turbot (*Psetta maxima*) in the Baltic Sea. *Molecular Ecology*, 16(1), pp. 115-126.

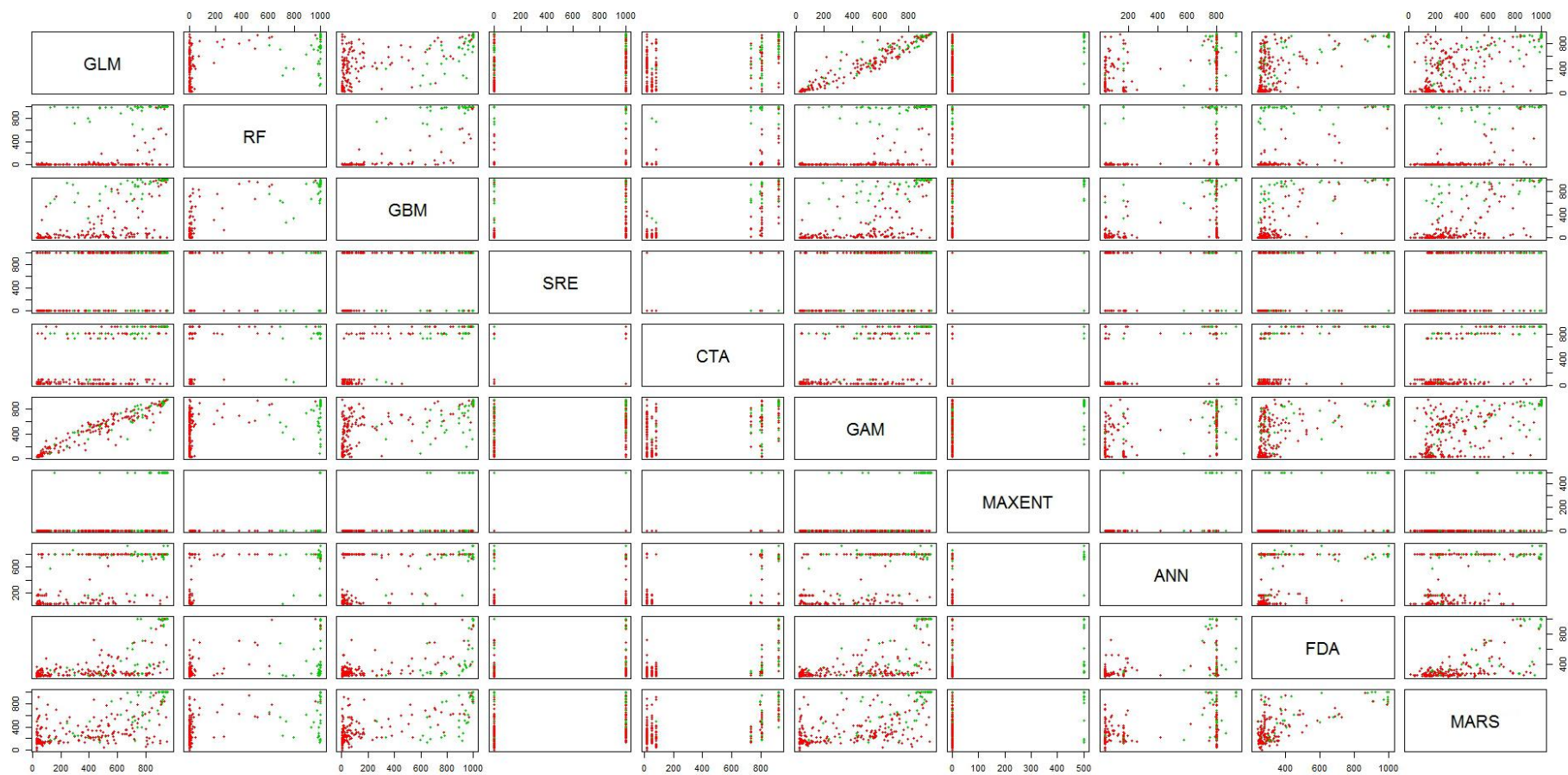
- Florin, A.-B., Sundblad, G. & Bergström, U., 2009. Characterisation of juvenile flatfish habitats in the Baltic Sea. *Estuarine, Coastal and Shelf Science*, 82(2), pp. 294-300.
- Fotonoff, N. P. & Millard, R. C., 1983. *Algorithms for Computations of Fundamental properties of seawater*, Paris: UNESCO.
- Franklin, J., McCollough, P. & Gray, C., 2000. Terrain Variables Used for Predictive Mapping of Vegetation Communities in Southern California. i: J. W. a. J. Gallant, red. *Terrain Analysis: Principles and Applications*. Wiley, New York: pp. 331-353.
- Graham, M. H., 2003. Confronting Multicollinearity in Ecological Multiple Regression. *Ecology*, 84(11), pp. 2809-2815.
- Guisan, A. et. al., 2007. What matters for predicting the occurrences of trees: Techniques, Data, or species' characteristics?. *Ecological Monographs*, 77(4), pp. 615-630.
- Guisan, A. & Zimmermann, N. E., 2000. Predictive habitat distribution models in ecology. *Ecological Modelling*, 135(2-3), pp. 147-186.
- Hastie, T., Tibshirani, R. & Buja, A., 1994. Flexible Discriminant Analysis by Optimal Scoring. *Journal of the American Statistical Association*, 89(428), pp. 1255-1270.
- HELCOM, 2013. *HELCOM Red List Species Information Sheets (SIS) Fish*, Köpenhamn: HELCOM.
- HELCOM, 2015. *Baltic Sea data and map service*. [Online]
Available at: <http://maps.helcom.fi/website/mapservice/index.html>
[Använd 16 11 2015].
- ICES, 2014. *Turbot in Subdivisions 22–32 (Baltic Sea)*, ICES.
- Jokinen, H. et. al., 2016. Juvenile flatfish in the northern Baltic Sea — long-term decline and potential links to habitat characteristics. *Journal of Sea Research*, 107(1), pp. 67-75.
- Karås, P. & Klingsheim, V., 1997. Effects of temperature and salinity on embryonic development of turbot (*Scophthalmus maximus* L.) from the North Sea, and comparisons with Baltic populations. *Helgoland Marine Research*, 51(2), pp. 241-247.
- Karlsson, R. & Hansbo, S., 1984. *Jordarternas indelning och benämning*, Stockholm: Svenska geotekniska föreningen (SGF).
- Kratzer, S., Håkansson, B. & Sahlin, C., 2003. Assessing Secchi and Photic Zone Depth in the Baltic Sea from Satellite Data. *AMBIO*, 32(8), pp. 577-585.
- McPherson, J. M. & Jetz, W., 2007. Effects of species' ecology on the accuracy of distribution models. *Ecography*, 30(1), pp. 135-151.
- Merow, C., Smith, M. J. & Silander, Jr, J. A., 2013. A practical guide to MaxEnt for modeling species' distributions: what it does, and why inputs and settings matter. *Ecography*, 36(10), pp. 1058-1069.
- Moisen, G. G. & Frescino, T. S., 2002. Comparing five modelling techniques for predicting forest characteristics. *Ecological Modelling*, 157(2-3), pp. 209-225.

- Nissing, A., Florin, A.-B., Thorsen, A. & Bergström, U., 2012. Egg production of turbot, *Scophthalmus maximus*, in the Baltic Sea. *Journal of Sea Research*, 84(1), p. 77–86.
- Ojaveer, H. et. al., 2014. *Integrating spatial processes into ecosystem models for sustainable utilization of fish resources*, BONUS.
- Phillips, S. J., Anderson, R. P. & Schapire, R. E., 2006. Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, 190(3-4), pp. 231-259.
- R version 3.2.2, 2015. u.o.: The R Foundation for Statistical Computing.
- Reiss, H. et. al., 2011. Species distribution modelling of marine benthos: a North Sea case study. *Marine Ecology Progress Series*, 442(71), pp. 71-86.
- Sundblad, G., 2010. *Spatial Modelling of Coastal Fish - Methods and Applications*, Uppsala: Uppsala Universitet.
- Sundblad, G., Isæus, M., Florén, K. & Wijckmark, N., 2012. *Utvärdering av undervattensvideo (dropvideo) i jämförelse med dykning som visuell metod för uppföljning av marina naturtyper: Jämförande metod och verifiering i Skåne metod och verifiering i Skåne*, Stockholm: AquaBiota Notes.
- Thuiller, W., Lafourcade, B. & Araujo, M., 2009. *Modelling Manual for BIOMOD*, Montpellier.
- Thuiller, W., Georges, D. & Engler, R., 2015. *Ensemble platform for species distribution modeling*.
- Vandamme, S. et. al., 2014. Regional environmental pressure influences population differentiation in turbot (*Scophthalmus maximus*). *Molecular Ecology*, 23(3), pp. 618-636.
- Vayssières, M. P., Plant, R. E. & Allen-Diaz, B. H., 2000. Classification trees: An alternative non-parametric approach for predicting species distributions. *Journal of Vegetation Science*, 11(5), pp. 679-694.
- Voigt, H.-R., 2002. Piggvaren i våra kustvatten. *Fiskeritidskrift för Finland*, 46(1), pp. 25-27.
- Wennberg, S. & Lindblad, C., 2006. *Sammanställning och Analys av Kustnära Undervattenmiljö (SAKU)*, Stockholm: Naturvårdsverket.
- Wulff, F., Stigebrandt, A. & Rahm, L., 1990. Nutrient Dynamics of the Baltic Sea. *Ambio*, 19(3), pp. 126-133.
- Zuur, A. F., Ieno, E. N. & Elphick, C. S., 2010. A protocol for data exploration to avoid common statistical problem's. *Methods in Ecology and Evolution*, 1(1), pp. 3-14.

7 BILAGOR



*Bilaga 1. Jämförelse av modelleringsmetoder från modelleringen med korsvalidering. Värden på axlarna är sannolikhet*1000. Grön punkt står för förekomst, röd står för icke-förekomst.*



Bilaga 2. Jämförelse av modelleringsmetoder från modelleringen med kalibrering och validering. Värden på axlarna är sannolikhet*1000. Grön punkt står för förekomst, röd står för icke-förekomster.