


Handelshögskolan
VID GÖTEBORGS UNIVERSITET

Bayesiansk klassificering av ursprung för fångster av sill

Kandidatuppsats i statistik
Henrik Fryk, 840924-4970
Institutionen för nationalekonomi med statistik
Göteborgs Universitet
Handledare: Robert Jonsson
Höstterminen 2012

Sammanfattning

I september 2008 fångades 311 juveniler (unga individer) av arten sill (Clupea harengus) in vid fyra olika områden på den svenska västkusten; Råssö, Hunnebo, Askerö- och Gullmarfjorden. Genetiska, kemiska och morfologiska data (längd och vikt) samlades in i syfte att bestämma vilka populationer de härstammade ifrån. Fångster av sill är så gott som alltid av blandat ursprung eftersom individer bara samlas i de populationer de stammar ifrån vid den tid på året då det är dags för parning (lek).

I den här uppsatsen klassificerar jag juvenilerna till tre regioner i vilka populationerna ingår, genom att jämföra de data som samlats in från juvenilerna med motsvarande data i en stor databas över populationer i Nordsjön, Skagerrak, Kattegat och Östersjön samt med ytterligare en population, Ris**ør,** som inte ingår i denna databas. Data om populationerna kommer från samples och de sanna populationsparametrarna (dvs. allelproportionerna) är därför okända, vilket komplicerar analysen.

Punktskattningar och sk sannolikhetsintervall för andelarna juveniler vid de olika fångstområdena som kommer från respektive region har räknats ut med Bayesianska metoder, implementerade i freewareprogrammen Geneclass2 (Piry et al, 2004) och Bayes (Masuda, 2002).

En utmaning har varit att bestämma andelen juveniler från Nordsjön som kommer från vårlekande populationer, eftersom de genetiska skillnaderna mellan dessa och de till antalet många fler höst- och vinterlekande populationerna är små. Separata analyser har gjorts för alla vårlekande juveniler (lekperiod bestämd mha sk otoliter, se avsnittet terminolgi) och antyder att det finns fler bidragande populationer än de som har undersökts här. Övriga resultat av de analyser som gjorts är att populationer från Nordsjön dominerar som ursprung vid alla fångstområden och att man kan konstruera en regel som innebär att alla individer med BMI (Body Mass Index) under ett visst värde klassificeras som vårlekare, åtminstone om ens sample består av juveniler och att de är infångade ungefär samtidigt. De vårlekande juveniler med lågt BMI verkar i högre grad komma från Nordsjöpopulationer än de med högt BMI. Däremot verkar de förstnämnda i lägre grad komma från Skagerrak än de sistnämnda.

Summary

In september 2008, 311 juvenile herring (Clupea harengus) were caught at four different sites on the Swedish west coast; Råssö, Hunnebo, Askerö- and Gullmarfjorden. Genetic, chemical and morphological (length and weight) data were gathered for the purpose of determining populations of origin. Herring catches are nearly always of mixed origin since individuals only gather in their populations of origin at the time of the year when they mate (spawn).

In this thesis, I determine the origin of the juveniles as one of three regions, each consisting of a number of populations. This is done by comparison of data gathered from the juveniles with the corresponding data in a large database of populations in the North Sea, Kattegat, Skagerrak and the Baltic Sea, as well as with data from an additional population, Risør, which is not included in the database. Data about the populations come from samples and the true population parameters (i.e. allele proportions) are therefore unknown, which complicates the analysis.

Point estimates and so-called probability intervals for the proportions of juveniles at the different sites coming from each of the regions have been calculated with Bayesian methods, implemented in the freewares Geneclass2 (Piry et al, 2004) and Bayes (Masuda, 2002).

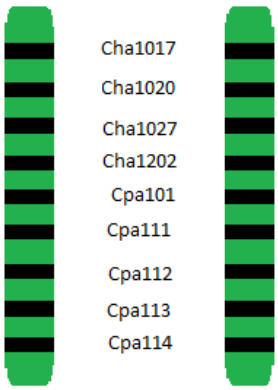
One of the challenges has been to determine the proportion of the juveniles from the North Sea that come from spring spawning populations, since the genetic differences between those and the more numerous autumn and winter spawning populations are small. Separate analyses have been performed for all spring spawning juveniles (spawning period determined using so-called otoliths, see the section on terminology) and suggest that there are more contributing populations than those examined. Other results of my analyses are that populations from the North Sea dominate as a source of the juveniles at all sites and that a rule can be constructed according to which all individuals with BMI (Body Mass Index) below a certain value are classified as spring spawners, at least if one´s sample consists of juveniles which have been caught at approximately the same time. The spring spawning juveniles with low BMI seem to come from the North Sea to a greater extent than those with low BMI. Also, the former seem to come from Skagerrak to a lower extent than the latter.

Innehållsförteckning

1. Terminologi

Populationsekolgi är som all vetenskap full av facktermer. Därför ges här en liten ordlista med vad man behöver veta för att kunna läsa uppsatsen. OBS, definitionen av allel nedan gäller specifikt för den typ av gener (sk mikrosatelliter) som har utnyttjats för analyserna i den här uppsatsen!

Term	**Förklaring**
Juvenil	Ung individ
Lek	Lek innebär att könsmogna individer samlas för att para sig med varandra
Lekperiod	Den period, höst, vinter eller vår, som en individ är kläckt och leker
Lektyp	Lektyp för en individ är antingen höst-, vår- eller vinterlekare, beroende på dess lekperiod.
Otolit	Hörselsten som kan ge information om när en individ är kläckt (samt därmed dess lekperiod och lektyp)
Kromosom	Det genetiska materialet är hos djur uppdelat på flera kromosomer, varje kromosom är en DNA-molekyl tätt paketerad tillsammans med proteiner, se figur 1.
Gen	Ett avsnitt av en kromosom som utgör en ärvbar enhet
Locus (pl. loci)	Genens position på kromosomen
Allel	Variant av en gen. Olika alleler har olika antal repetitioner av en kort DNA-sekvens.
Genotyp	En individs hela uppsättning av alleler vid en eller flera loci
Assignment methods	Några likartade metoder för klassificering av en individs eller en grupp av individers ursprung baserat på dess alleluppsättning (genotyp) för ett antal loci
Baseline	Uppsättningen populationer som antas utgöra ursprung för de individer eller grupper man vill klassificera



Figur 1: Mycket förenklad bild av två kromosomer med de 9 loci namngivna som utgör de aktuella genotyperna För enkelhetens skull är alla loci markerade som om de förekom på ett och samma par kromosomer.

2. Syfte

Förvaltning av arter som sill (Clupea harengus), som förflyttar sig över stora områden under året, kräver kunskap om rekryteringen av fiskebestånden.
Uppsatsen går ut på att använda statistiska metoder för att bestämma ursprungspopulation eller -region för fyra fångster med sammanlagt 311 juveniler från den svenska västkusten samt att undersöka om morfologiska data som BMI (Body Mass Index) kan användas för att bestämma lektyp för en individ.

3. Bakgrund

För forsknings- och naturvårdssyften, och även av rent ekonomiska skäl är det viktigt att kunna klassificera individer eller grupper av individer till sina populationer. Exempel på områden där sådan kunskap är viktig är vid åtal om tjuvjakt, förvaltning av fiskebestånd, bevaring av sällsynta genpooler och populationer etc. Förvaltning av bestånd är den huvudsakliga anledningen till att man vill ta reda på urprunget för fångster av sill. Arten förekommer i distinkta populationer som leker vid en viss plats och period under året. Individer rör sig över stora områden under resten av året men återvänder till den plats där de själva kläcktes när det är dags leka. På så vis blir lekpopulationerna reproduktivt isolerade från varandra och skillnader i allelproportioner uppkommer dem emellan, vilket kan utnyttjas när man vill bestämma ursprung för individer eller grupper av individer. Den typ av gener som kallas mikrosatelliter har visat sig speciellt lämpliga för detta ändamål eftersom de har hög mutationshastighet, vilket innebär att skillnader i allelproportioner mellan populationer uppkommer på relativt kort tid.

Informationen om vilka populationer som är ursprung för sill på svenska västkusten är ofullständig. De juveniler som skall klassificeras i den här uppsatsen har fått 9 loci undersökta, se tabell 19 i appendix. De fångades vid fyra områden på den svenska västkusten och skall med assignmentmetoder klassificeras till de populationer som ingår i HERGEN-databasen (HERGEN är ett stort EU-projekt som samlat information om populationer som leker i Nordsjön, Östersjön, Skagerrak och Kattegatt).
I HERGEN-databasen finns information på individnivå om datum för fångst, longitud och latitud för fångstplats, kön, vilket laboratorie de analyserats av, vikt, längd, alleler vid 9 loci (samma loci som undersökts hos juvenilerna i fångsterna), uppskattad lekperiod för några individer utifrån otoliternas (hörselstenarnas) utseende mm. De individer som ingår i HERGEN-databasen utgör inte i något fall en hel population, utan utgör samples från dessa.
Totalt 19 populationer, 18 från HERGEN-databasen samt ytterligare en (Risør), undersöktes som potentiella populationer för de infångade juvenilerna, se figur 2. Dessa utgör vad som kallas baseline.
De fyra områden där juvenilerna fångades in finns angivna i tabell 1. Områdenas namn används i fortsättningen som namn på själva fångsterna.

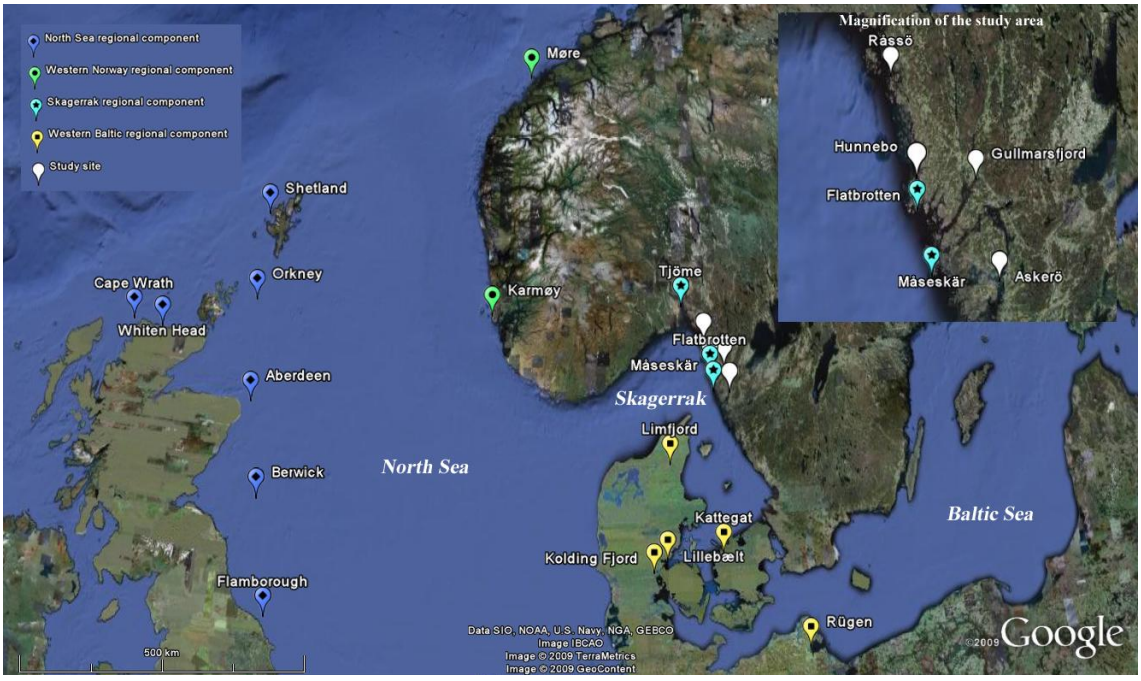
Fångstområde	**Askerö**	**Hunnebo**	**Råssö**	**Gullmaren**
Antal individer	89	100	100	23
Individers beteckning	ASx, x=001,..,089	HUx, x=001,..,100	RAx, x=001,...,100	GUx, x=001,...,023

Tabell 1: Fångstområden där juvenilerna fångades in

Populationerna är namngivna efter de områden där de leker. Grupper av populationer bildar de tre regioner som anges i figur 2.

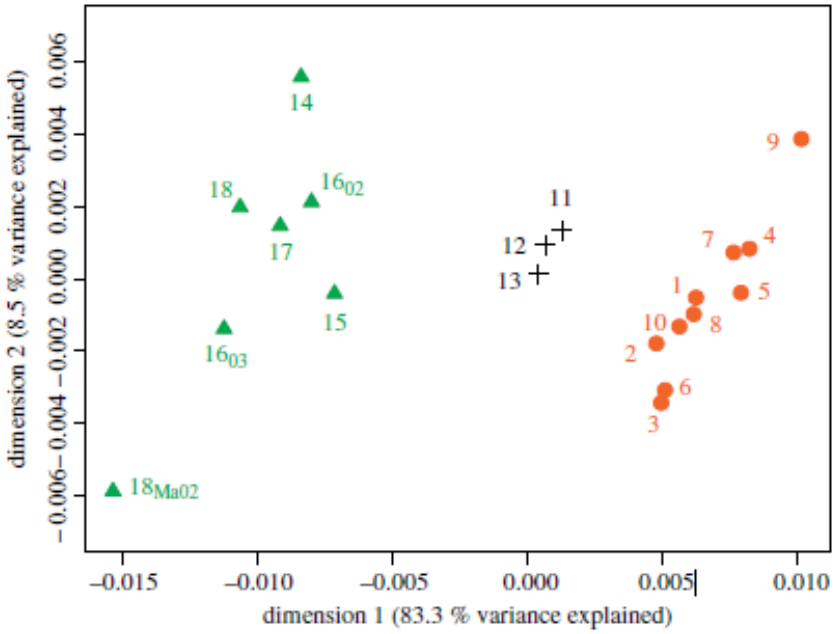
Region	Population, beteckning	Population, nummer
Nordsjön (NS)	CW	1
	WH	2
	SH	3
	OR	4
	AB	5
	BB	6
	FL	7
	EC	8
	MO	9
	KO	10
Skagerrak (Skg)	RO	11
	TJ	12
	MA	13
	FB	14
Kattegat och västra Östersjön (KaWBa)	LF	15
	KA	16
	KF	17
	LB	18
	RU	19

Tabell 2: De olika lekpopulationerna som juvenilerna skall klassificeras till. Populationerna är namngivna efter det områden där de leker. Förkortningar: CW=Cape Wrath, WH=Whiten Head, SH=Shetland, OR=Orkney, AB=Aberdeen, BB=Berwick Bank, FL=Flamborough, EC=English Channel, KO=Karmøy, MO=Møre, RO=Risør, TJ=Tjöme, MA=Måseskär, FB=Flatbrotten, LF=Limfjord, KA=Kattegat, KF=Kolding Fjord, LB=Lilla Bält, RU=Rügen.
Brun, blå respektive grå färg innebär att populationen är höst-, vinter respektive vårlekande



Figur 2. Karta över lekpopulationerna samt över fångstområdena där juvenilerna fångades (infällda bilden längst upp till höger).

I en artikel av Ruzzante et al (2006) visade det sig att de 18 lekpopulationerna i HERGEN-databasen kan delas in i tre regioner med små genetiska skillnader mellan populationerna inom en region men relativt stora genetiska skillnader mellan regionerna, se figur 3. Egna analyser visar att Risør är mest genetiskt lik populationerna i region Skagerrak.



Figur 3: Multidimensional scaling plot (MDS) av de 18 populationerna i HERGEN-databasen, från Ruzzante et al (2006). Denna MDS är en visualisering i två dimensioner av de parvisa genetiska skillnaderna mellan populationerna. Skillnaderna speglas av avstånden i dimension 1 och dimension 2. Dock motsvarar ett givet avstånd på x-axeln (dimension 1) en större genetisk skillnad än samma avstånd på y-axeln (dimension 2).
Gröna trianglar=Nordsjön, plus=Skagerrak, orangea cirklar=Kattegat och Östersjön. Risør är inte med i figuren men ingår i region Skagerrak.

De flesta populationerna samplades vid två olika tillfällen, år 2002 respektive 2003. Endast två av dem, Kolding Fjord och Rügen (nummer 16 respektive 18 i figur 3), var signifikant genetiskt olika under de båda åren.
För att ta fram figur 3 ovan har man utgått från en matris över s.k. F_{st}-värden (uträknade enligt definition av Weir & Cockerham 1984), framtagna för varje möjligt par av de 18 populationerna. Dessa värden är mått på skillnaden i allelproportioner för de 9 analyserade loci. En stor skillnad mellan två populationer i proportionerna av alleler vid flera loci ger ett högt Fst-värde.
En MDS som den i figur 3 är ett sätt att visualisera mått på skillnader/olikheter. En stressfunktion har definierats som mäter överrensstämmelsen mellan geometriskt avstånd i figuren och F_{st}-värden i matrisen. Ett datorprogram tar sedan fram den figur som minimerar stressfunktionen.

4 Teori

4.1 Bestämning av lektyp

4.1.1 Otolitanalyser

Populationerna har olika lekperioder; vinter, vår och höst (se figur 2). De lekande individerna sägs ha någon av tre lektyper beroende på deras populations lekperiod; vinter-, vår- eller höstlekare. Lektypen för en individ kan bestämmas genom analys av dess otolit, vilket gjordes för ett tjugotal av individerna från varje fångst.

I Clausen et al (2007) undersöks tillförlitligheten i sådana analyser genom att klassificeringen av sill på basis av otoliter jämförs med tid på året då sillen fångats i lekande tillstånd. Andelen korrekta klassificeringar presenteras i tabell 3 i artikeln och tabell 3 nedan innehåller samma information. Se tabellens förklarande text för förklaring av beteckningar.

	Lektyp (lt)		
	s	a	w
$P(s_d\|lt)$	0.97	0.01	0
$P(a_d\|lt)$	0.02	0.92	0.32
$P(w_d\|lt)$	0.01	0.07	0.68

Tabell 3:De årstider då individer observerats leka betecknas s, a och w vilket motsvarar vår, höst respektive vinter. Klassificering som vår-, höst- och vinterlekare betecknas s_d, a_d respektive w_d. Lektyp förkortas lt.

Det finns ingen tabell i Clausen et al (2007) som innehåller sannolikheterna för respektive lektyp givet klassificering som vår-, höst- respektive vinterlekare, men dessa sannolikheter kan enkelt räknas ut med hjälp av de data som finns angivna, vilket har gjorts i tabell 4. Dessa sannolikheter kallas prediktiva värden, till skillnad från värdena i tabell 3 som kallas sensitivitetsvärden (Altman, 1991).

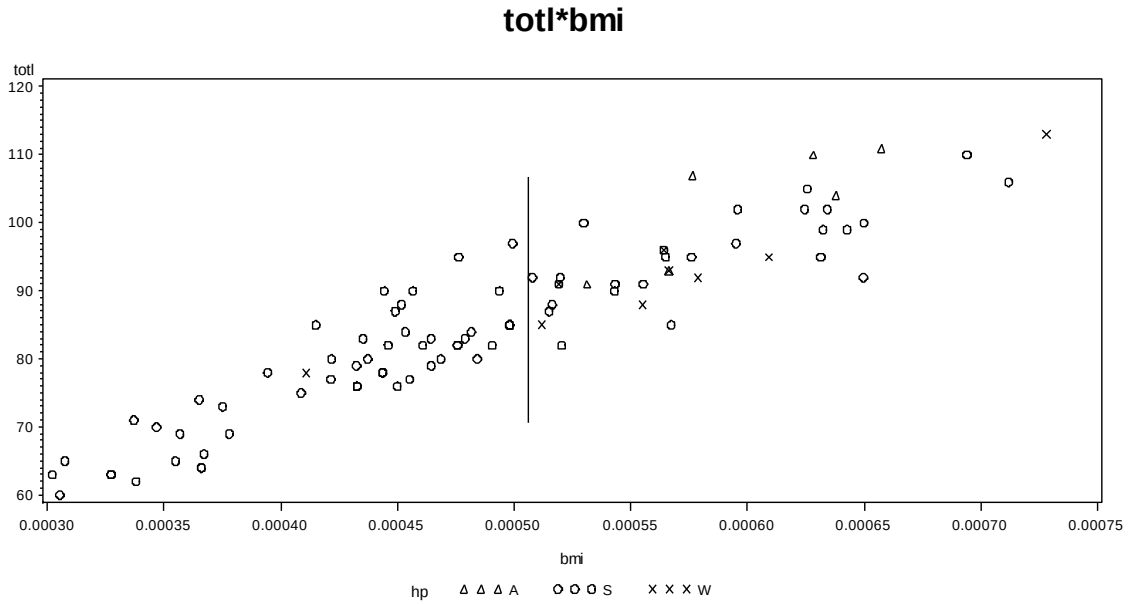
	Klassificering		
	s_d	a_d	w_d
P(s|klassificering)	0.99	0.02	0.04
P(a|klassificering)	0.01	0.90	0.27
P(w|klassificering)	0	0.08	0.70

Tabell 4: Beteckningar som i tabell 2. Summan i sista kolumnen blir större än 1 på grund av att värdena är avrundade.

Om man låter höst- och vinterlekare utgöra en gemensam kategori blir andelen korrekta klassificeringar högre:
$P(a \cup w|\ a_d \cup w_d) = 0.98$, $P(a \cup w|\ s_d) = 0.01$ och $P(s|\ a_d \cup w_d) = 0.02$.
Man bör förståss komma ihåg att de exakta siffrorna ovan är specifika för Clausen et als studie men man kan i alla fall dra slutsatsen att om analyserna av otoliter i vår studie håller likvärdig kvalitet är kvalificeringarna som vår- respektive höst-/vinterlekare mycket pålitliga.

4.1.2 BMI

Eftersom bara en del av individerna från Askerö, Hunnebo och Råssö fått sina otoliter analyserade undersökte jag om det går att bestämma lekperiod för en individ utifrån morfologiska data som BMI. Detta verkade troligt eftersom alla juveniler är fångade ungefär vid samma datum (några 15e och några 16e september, år 2008) och det därför borde finnas en, med tanke på att de är juveniler, stor ålderskillnad mellan individer av olika lektyp, se figur 4.



Figur 4: Sambandet mellan BMI och längd (totl) för olika lektyper.

Som framgår av figur 4 är spannet av BMI-värden för vårlekare stort. Det som är tydligast är att alla utom en individ med BMI mindre än 0.0005 fått diagnosen vårlekare. Därför har klassificeringen (till regioner) av juveniler som är vårlekare enligt BMI jämförts med klassificeringen av juveniler som är vårlekare enligt otolitdata, se avsnitt 6.4. Om det finns överrensstämmelse innebär det att man skulle kunna använda BMI istället för otolitdata, för att göra klassificeringar av lektyp.

4.2 Genetiska analyser

När man klassificerar individer till populationer kan man antingen analysera hela fångsten samtidigt och skatta andelerna från respektive ursprung. Man kan också analysera enstaka individer i taget. Det förra tillvägagångsättet kallas ”Mixed stock analysis” och det senare ”Individual assignments” och de har båda sina för- och nackdelar. Jag kommer i fortsättningen referera till dem som MSA respektive IA.
Eftersom det är sammansättningen av fångsterna som är av intresse snararen än vilken individ som har vilket ursprung är MSA mest användbart men även IA är intressant eftersom det tillåter att man kan identifiera vilka individer som skall analyseras vidare och kan utnyttja eller jämföra med annan information än genetiska data (t.ex. otolitdata).
Varje individ i lekpopulationerna och de juveniler som skall klassificeras har fått 9 loci analyserade med två alleler per locus. Att varje locus har två alleler beror på att det hos djur finns två kopior av varje, en från varje förälder, se figur 1.

Man kan göra ett antagande om att sannolikheten att ett locus på en kromosom innehåller en viss allel är oberoende av vilken allel som finns vid samma locus på den andra kromosomen, och av vilka alleler som finns vid övriga loci (biologer kallar dessa oberoenden för Hardy-Weinberg Equilibrium respektive Linkage Equilibrium). Om allelfrekvenserna i populationerna hade varit kända hade man därför kunnat räkna ut likelihooden för att en specifik genotyp skall uppstå i en viss population som produkten av nio multinomialsannolikheter med n=2 och $p_1 \ldots p_k$ motsvarande frekvenserna av de k alleler som förekommer vid locuset i den aktuella populationen. Men eftersom det är i praktiken omöjligt att bestämma de exakta allelfrekvenserna i en population har Bayesianska metoder utvecklats av bland andra Rannala & Mountain (1997) där frekvenserna själva får en sannolikhetsfördelning baserad på samples från populationerna.
Vid både de IA och MSA-beräkningar som presenteras nedan används Dirichletfördelningen som en aprioriförddelning för allelfrekvenserna i de olika populationerna och vid MSA dessutom för populationernas andelar av det analyserade samplet.

Dirichletfördelningen är en kontinuerlig multivariat fördelning som betecknas Dirichlet($\nu_1, \ldots, \nu_k$). Den har följande frekvensfördelning:

$$\mathrm{P}(\mathrm{x}_1, \ldots, \mathrm{x}_k) = \frac{\Gamma(\sum_{i=1}^{k} \nu_i)}{\prod_{i=1}^{k} \Gamma(\nu_i)} \prod_{i=1}^{k} x_i^{\nu_i - 1},$$

där $x_i \geq 0$, $\sum_{i=1}^{k} x_i = 1$ och $\nu_1, \ldots, \nu_k$ är parametrar.

4.2.1 Beteckningar

Nedan följer en sammanställning av de beteckningar som används i teori- och metoddelen.

Index

Population: i=1,…,I (I=19)
Individ i fångst: m=1,…,M
Locus: j=1,…,J (J=9)
Allel vid locus j: h=1,…,H_j

Antalet olika alleler i samplet med alla populationer som helhet vid respektive locus används som en skattning av antalet olika alleler i var och en av populationerna vid locuset. H_j är alltså samma alltså samma för alla populationer men kan variera mellan loci.

Matriser och vektorer
Matris med absoluta allelfrekvenser för alla M individer i en fångst:

$$\mathbf{X} = \left(\left((x_{mjh})_{m=1}^{M} \right)_{j=1}^{J=9} \right)_{h=1}^{H_j}$$

x_{mjh} är alltså antalet av allel h vid locus j hos den m:te individen i en fångst. x_{mjh} går från 0 till 2.

Matris med absoluta allelfrekvenser i samplen från alla populationer:

$$\mathbf{Y} = \left(\left((y_{ijh})_{i=1}^{I=19} \right)_{j=1}^{J=9} \right)_{h=1}^{H_j}$$

y_{ijh} är alltså antalet av allel h vid locus j i den i:te populationen. y_{ijh} går från 0 till två gånger samplestorleken för den i:te populationen.

Matris med de sanna allelproportionerna i alla populationer:

$$\mathbf{Q} = \left(\left((q_{ijh})_{i=1}^{I=19} \right)_{j=1}^{J=9} \right)_{h=1}^{H_j}$$

q_{ijh} är alltså proportionen av allel h vid locus j i den i:te populationen. q_{ijh} går från 0 till 1.

Matrisen med andelarna från respektive population: $\boldsymbol{p} = (p_i)_{i=1}^{I=19}$
Tillskriven populationstillhörighet för individerna i en fångst: $\boldsymbol{Z} = (\boldsymbol{z}_m)_{m=1}^{M}$
Populationstillhörighet för individ m: $\boldsymbol{z}_m = (z_{mi})_{i=1}^{I=19}$ där z_{mi} är 1 för den tillskrivna populationen och 0 för övriga. Om t.ex. den åttonde individen i fångsten tillskrivs population 3 ser alltså matrisen ut såhär:
$\mathbf{z}_8$=[0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0]

4.2.2 Individual assignments (IA)

Överblick

Metoden för att göra IA som beskrivs här bygger på en artikel av Rannala&Mountain (1997).
Målet med IA är att, givet insamlad data om allelfrekvenserna i de olika potentiella ursprungspopulationerna samt den observerade genotypen för en enskild individ, skatta sannolikheten för individens ursprung i var och en av populationerna. Om man gör vissa antaganden kan den sannolikheten räknas ut med följande ekvation:

$$P(m_i|\mathbf{X_m}, \mathbf{Y}) = \frac{P(\mathbf{X}_\mathrm{m}|m_i,\boldsymbol{Y})*P(m_i)}{\sum_k P(\boldsymbol{X}_m\,|m_{k,}\boldsymbol{Y})*P(m_k)}\,, \qquad (1)$$

(se ekvation (1) i Baudouin et al., 2004).
där m_i är händelsen ”individ m kommer från population i”, $\mathbf{X}_m$ är matrisen som innehåller individens genotyp och **Y** är matrisen med allelfrekvenser i samplen från de olika populationerna som utgör baseline (se avsnitt 4.2.1).

Apriorisannnolikheten, P(m_i),

Apriorisannolikheten i ekvation (1), P(m_i), är sannolikheten att en individ kommer från population i innan man tagit hänsyn till dess genotyp. Man kan alltså utnyttja annan information än den genetiska när sådan finns tillgänglig.

Likelihoodfunktionen, P($\boldsymbol{X}_m|m_i,\boldsymbol{Y}$)

Eftersom de olika loci antas vara oberoende av varandra blir Likelihoodfunktionen för hela den observerade genotypen, $\mathbf{X}_m$, produkten av sannolikheten för observerad genotyp vid varje locus:

$$\mathrm{P}(\mathbf{X}_\mathrm{m}|\mathrm{m_i}, \mathbf{Y}) = \prod_{\mathrm{j}=1}^{\mathrm{J}} \mathrm{P}(\mathbf{x}_\mathrm{mj}|\mathrm{m_i}, \mathbf{y}_\mathrm{ij}), \qquad (2)$$

Dock är de exakta allelfrekvenserna i populationerna okända och måste skattas mha $\mathbf{Y}_i$ (i=1,..,I). Vid ett visst locus beror likelihooden av $\mathbf{y}_{ij}$ enligt följande:

$$\mathrm{P}\left(\mathbf{x}_\mathrm{mj}\middle|\mathrm{m_i}, \mathbf{y}_\mathrm{ij}\right) = \int_0^1 \mathrm{P}(\mathbf{x}_\mathrm{mj}|\mathrm{m_i}, \mathbf{q}_\mathrm{ij})\, \mathrm{P}(\mathbf{q}_\mathrm{ij}|\mathbf{y}_\mathrm{ij})\mathrm{d}\mathbf{q}_\mathrm{ij}\,, \qquad (3)$$

Aposteriorifördelningen för allelfrekvenserna, P($\mathbf{q}_{ij}$|$\mathbf{y}_{ij}$), är en Dirichletfördelning uppdaterad med de observerade allelfrekvenserna i de samples som kommer från populationerna.

$\mathrm{P}\left(\mathbf{x}_\mathrm{mj}\middle|\mathrm{m_i}, \mathbf{q}_\mathrm{ij}\right)$ är en multinomialfördelning:

$$\mathrm{P}\left(\mathbf{x}_\mathrm{mj}\middle|\mathrm{m_i}, \mathbf{q}_\mathrm{ij}\right) = \frac{\mathrm{x_{mj}!}}{\mathrm{x_{mj1}!\, x_{mj2}!\ldots x_{mjH_j}!}} \prod_{\mathrm{h}=1}^{\mathrm{H_j}} \mathbf{q}_\mathrm{ijh}^{\mathrm{x_{mjh}}} \qquad (4)$$

x_{mj} är summan av antalet alleler för individ m vid locus j ($x_{mj} = \sum_{h=1}^{H_j} x_{mjh}$), inte att förväxla med $\mathbf{x}_{mj}$ som är matrisen över allelfrekvenserna i populationen i fråga. x_{mj}=2 för alla organismer med två kopior av varje kromosom, som sill, och därmed två alleler vid varje locus. Om man har starka skäl att tro att flera individer kommer från samma ursprung blir summan istället 2 gånger antalet individer och vinsterna i fråga om korrekta assignments kan bli enorma (det motsvarar ju en genotyp som är n gånger så stor, om man assignar n individer med samma ursprung samtidigt).

4.2.3 Mixed stock analysis (MSA)

De MSA-beräkningar jag har utfört bygger på metoder framtagna av Pella & Masuda (2001). Målet med beräkningarna är att skatta andelen av en fångst som kommer från varje potentiellt ursprung, dvs **p**, utifrån data om uppsättningen genotyper i fångst (**X**) och allelfrekvenser i baseline (**Y**).

Under vissa antagananden får **p** och **Q** därför en gemensam fördelning givet **X** och **Y**:

$$P(\mathbf{p},\mathbf{Q}|\mathbf{X},\mathbf{Y}) = \frac{P(\mathbf{X}|\mathbf{p},\mathbf{Q})*P(\mathbf{p})*P(\mathbf{Q}|\mathbf{Y})}{P(\mathbf{X})}, \quad (5)$$

Från aposteriorifördelningen kan man räkna ut punktskattningar för parametrarna, som median och medelvärde, och ett symmetriskt intervall som täcker 95% av fördelningen (i fortsättningen kallat sannolikhetsintervall).
P(**X**) är marginalfördelningen för **X:**

$$P(\mathbf{X}) = \int_{\mathbf{Q},\mathbf{p}} P(\mathbf{X}|\mathbf{p},\mathbf{Q}) * P(\mathbf{p}) * P(\mathbf{Q}|\mathbf{Y})\, d\mathbf{Q} d\mathbf{p} \quad (6)$$

P(**p**) är en Dirichletfördelning och P($\mathbf{q}_{ij}|\mathbf{y}_{ij}$) är som vid IA en Dirichletfördelning uppdaterad med de observerade allelfrekvenserna i de samples som kommer från populationerna.

P(Q|Y)
Eftersom allelfrekvenserna i olika populationer och vid olika loci antas vara oberoende blir P(**Q**|**Y**) produkten av sannolikheten vid varje locus i varje population:

$$P(\mathbf{Q}|\mathbf{Y}) = \prod_{i=1}^{I} P(\mathbf{Q}_i|\mathbf{Y}_i) = \prod_{i=1}^{I}\prod_{j=1}^{J} P(\mathbf{q}_{ij}|\mathbf{y}_{ij}), \quad (7)$$

(modifierad från ekvation (5) i Pella & Masuda, 2001).

P(X|p,Q)

Likelihooden för att en fångst skall innehålla en viss uppsättning genotyper beror på **p** och **Q** och är produkten av likelihooden för varje enskild genotyp. Varje enskild genotyps likelihood är summan av dess likelihood i var och en av populationerna gånger respektive populations andel av fångst. Om p_i är andelen av population i i fångst och f($\mathbf{X}_m|\mathbf{Q}_i$) betecknar den relativa frekvensen av genotyp $\mathbf{X}_m$ i population i är likelihooden för hela fångst med M individer:

$$P(\mathbf{X}|\mathbf{p},\mathbf{Q}) = \prod_{m=1}^{M}\left(\sum_{i=1}^{I} p_i\, f(\mathbf{X}_m|\mathbf{Q}_i)\right) \quad (8)$$

5. Metoder

5.1 Individual assignments (IA)

För att göra IA användes programmet Geneclass2 (Piry et al, 2004). Utdata från programmet fås i form av ”scores” och estimerade likelihoods (egentligen –log L) för varje kombination av population och individ som skall assignas. Estimerad likelihood för den m:te individens genotyp i population i betecknas $L_{m,i}$ och är helt enkelt ett kortare beteckning för första termen i täljaren i ekvation (1), $P(\mathbf{X}_m|m_i, \mathbf{Y})$.
Score för m:te individens genotyp i population *i* är ett specialfall av ekvation (1) när apriorisannolikheterna är lika för alla populationer, vilket i och för sig kan diskuteras.

$$score_{m,i} = \frac{L_{m,i}}{\sum_{J=1}^{k} L_{m,j}}, \qquad (9)$$

(se http://www1.montpellier.inra.fr/URLB/GeneClass2/Help/index.htm).

När det saknas information som leder till olika apriorisannolikheter för olika populationer (som t.ex. lektyp) eller när skillnaderna i likelihoodvärdena för de olika populationerna blir så stora att apriorisannolikheterna blir obetydliga vid uträkning av aposteriorisannolikheten, kan man med fördel göra assignments enbart utifrån scores. Det senare gäller inte individerna som skall assignas i den här uppsatsen, eftersom de genetiska skillnaderna mellan populationerna är relativt små.

I de fall det finns information som leder till olika apriorisannolikheter för de olika populationerna kan användaren av programmet använda likelihoodvärdena för att räkna ut aposteriorisannolikheter enligt ekvation (1) i t.ex. Excel.

Programmet kan även utföra beräkningar för att utesluta populationer som ursprung eller för att testa pålitligheten i assignments. Det ofta stora antalet alleler vid varje locus och det faktum att varje likelihood är en produkt av 9 sannolikheter (en för varje locus) gör att alla likelihoodvärden är mycket låga. Därför använder programmet en metod med simuleringar för att utföra dessa beräkningar. Det går till så att ett mycket stort antal (användaren väljer själv mellan 100, 1000, 10000, 100000 och 1000000) genotyper simuleras, utifrån observerade allelproportioner i samplet. Därefter räknas scores ut för varje simulerad individ och för den observerade individen varefter de rangordnas. Andelen genotyper med lägre scores motsvarar p-värdet vid ett statistiskt test av hypotesen att en individ kommer från en viss population. Vilket det kritiska värdet ska vara för att en population skall räknas som utesluten som ursprung bestäms av användaren själv.
Det finns tre olika algoritmer för dessa simuleringar. Den jag använt vid mina beräkningar är den som föreslogs av Paetkau et al. (2004). Den rekomenderas av skaparna av geneclass 2 och medför minst risk för att felaktigt utesluta en population som ursprung för en individ.

Algoritmen kännetecknas av följande:

- Två individer från den samplade populationen dras slumpmässigt. En könscell simuleras från var och en av de två individerna på så sätt att en av de två allelerna vid varje locus dras slumpmässigt. Slutligen förenas könscellen från vardera individ och bildar tillsammans genotypen för den simulerade individen.
- Flera sample av samma storlek som det verkliga samplet simuleras tills det totala antalet simulerade individer är uppnått. Ett likelihoodvärde räknas ut för varje simulerad individ baserat på allelfrekvenserna för dess sample minus individen själv.

Som framgår av ekvation (1) beror aposteriorisannolikheten för en individs ursprung i en viss population på vilka andra populationer som är potentiella ursprung för individen. Om populationen med högst

aposteriorisannolikhet, eller rent av alla potentiella ursprungspopulationer, utesluts som ursprung för en individ är det en indikation på att den sanna populationen inte har samplats.

Populationerna slogs samman map år, med undantag för Kolding Fjord och Rügen som uppvisade signifikanta genetiska skillnader mellan de båda åren de samplades.
Eftersom populationerna inom respektive region uppvisar små skillnader sinsemellan är risken stor att en individ får högre likelihood i en annan population än dess verkliga ursprungspopulatin, vilket också bekräftats vid testassignments av individer från HERGEN-databasen vars ursprung var känt. För enstaka alleler kan dock skillnader finnas i frekvens mellan olika populationer som blir betydelsefulla vid MSA eller IA och det kan därför vara olämpligt att slå samman alla populationer i en region. Istället har andelen från en region räknats ut som summan av de ingående populationernas andelar

5.2 Mixed stock analysis (MSA)

Programmet Bayes (Masuda, 2002) användes vid alla MSA. Bayes kan inte utföra de komplicerade beräkningar som krävs för att räkna ut punktskattningar och intervall för **p** ur fördelningen i ekvation (5). Istället simuleras samples av samma storlek som fångsten med en Markovkedje-algoritm (MCMC) för att skatta fördelningen. Vid varje cykel i en serie av simuleringar tilldelas varje individ ett ursprung med en multinomialsannolikhet som bestäms av **p**, **Q** och individens genotyp ($\mathbf{X}_m$):

$$w_{mi} = \frac{p_i f(\mathbf{X}_m|\mathbf{Q}_i)}{\sum_{k=1}^{I} p_k f(\mathbf{X}_m|\mathbf{Q}_k)} \qquad (10)$$

(modifierad från ekvation (8) i Pella & Masuda, 2001).
Värden för **p** och $\mathbf{Q}_i$ dras i sin tur ur fördelningar som bestäms av sammansättningen av populationstillhörigheter (**Z)** i föregående sample samt av **X** och **Y** enligt följande ekvationer (se längst upp på sid 10 för en beskrivning av Dirichletfördelningen):

$$P\left(\mathbf{p}^{(k+1)}\middle|\mathbf{X},\mathbf{Y},\mathbf{Z}^{(k)}\right) = P\left(\mathbf{p}^{(k+1)}\middle|\mathbf{Z}^{(k)}\right) = \text{Dirichlet}\left(\frac{1}{I} + \sum_{m=1}^{M} z_{m1}^{(k)}, \ldots, \frac{1}{I} + \sum_{m=1}^{M} z_{mI}^{(k)}\right), \qquad (11)$$

respektive

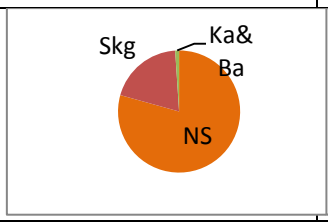
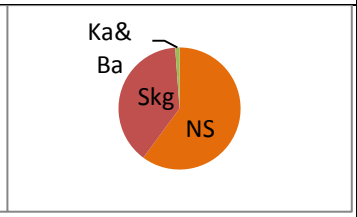
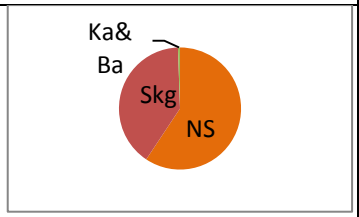
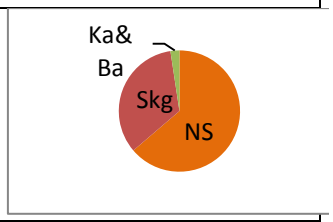
$$P\left(\mathbf{Q}_i{}^{(k+1)}\middle|\mathbf{X},\mathbf{Y},\mathbf{Z}^{(k)}\right) = \prod_{j=1}^{J} \text{Dirichlet}\left(\beta_{j1} + y_{ij1} + \sum_{m=1}^{M} z_{mi}^{(k)} x_{mj1}, \ldots, \beta_{jH_j} + y_{ijH_j} + \sum_{m=1}^{M} z_{mi}^{(k)} x_{mjH_j}\right), \qquad (12)$$

(modifierade från ekvation (9) och (10) i Pella & Masuda, 2001). I ekvation (11) och (12) är 1/I och $\boldsymbol{\beta}_j$ parametrar i aprioriförderlningarna för **p** respektive $\mathbf{Q}_i$.

Användaren rekommenderas i programmets manual (Masuda, 2002) att köra flera olika ”kedjor” med olika startvärden för p_i (i=1,...,I). Med kedja menas en serie av MCMC-simuleringar och utdata från de olika kedjorna vägs sedan samman för att skatta **p**. Så många cykler körs som krävs för att värden från dragningarna av **p** från dess aposteriorifördelning skall stabiliseras och **p** kunna skattas med önskvärd precision.

6. Resultat

Av de individer som fått lektyp bestämd var majoriteten vårlekare, vid varje fångstområde. Samtidigt visade MSA vid varje fångstområde att Nordsjön verkade vara den region som bidragit med flest juveniler. Sannolikhetsintervallen var dock breda, till stor del pga den relativt låga genetiska differentieringen även regioner emellan.
Råssö var det enda fångstområdet där den nedre gränsen för sannolikhetsintervallet för andelen juveniler från Nordsjön var större än 50%, se tabell 5. Samtidigt var området unikt på så sätt att alla juveniler som fått sina otoliter analyserade klassificerats som vårlekare. Om alla potentiella populationer har samplats borde därför majoriteten av dessa vara från de vårlekande nordsjöpopulationerna utanför Norges kust.
För en individ, RA074, hamnar alla populationer under det kritiska värdet vid geneclass uteslutningsberäkningar. Det är därför troligt att åtminstone en individ kommer från en osamplad population.

Fångs-tområde	Råssö	Askerö	Hunnebo	Gullmaren
				
Punkt- och intervall-skattning	NS 0.7932 (0.6453, 0.9145) Skg 0.1963 (0.0774, 0.3437) KaWBa 0.0103 (0, 0.0561)	NS 0.6011 (0.3604 ,0.8250) Skg 0.3874 (0.1629 ,0.6290) KaWBa 0.0115 (0,0.0627)	NS 0.5933 (0.3915 , 0.7885) Skg 0.4013 (0.2068 , 0.6040) KaWBa 0.0054 (0, 0.0326)	NS 0.6373 (0.2631, 0.9968) Skg 0.3388 (0, 0.7161) KaWBa 0.0238 (0, 0.1348)
Andel vårlekare	24/24	17/24	21/23	17/23

Tabell 5. Resultaten av MSA vid de fyra fångsplatserna

6.1 Vårlekare

För att undersöka andelen vårlekande juveniler som kom från Nordsjön gjordes MSA med bara dessa individer. Dels gjordes en MSA med bara vårlekande populationer i baseline, dels en där baseline var komplett. Såvida inte ett stort antal felklassificeringar av lektyp har gjorts och alla potentiella vårlekande populationer finns med i baseline borde resultaten av dessa MSA bli ungefär desamma, men som tabell 6, 8, 10 och 12 visar, var skillnaderna stora. Andelen juveniler från Nordsjön ökar markant när de höstlekande populationerna tas med, samtidigt som andelen från Skagerrak minskar kraftigt, medan andelen från Kattegat och Östersjön är nästan oförändrad. Förklaringen skulle kunna vara att flera juveniler från Skagerrak av en slump genetiskt liknar juveniler från de höstlekande populationerna, men de vårlekande Nordsjöpopulationerna liknar de höstlekande mer än vad de i Skagerrak gör. När man tittar på resultaten från IA ser man att de individer som får högst score i en höstlekande population oftast har en population från Skagerrak först på fjärde eller femte plats, vid Gullmaren och Råssö men även vid övriga fångstområden, se tabell 7, 9, 11 och 13.
För två individer, RA003 och RA023, utesluts alla vårlekande populationer, till skillnad från flera av de höstlekande.

Råssö

Baseline	Vårlekande pop.		Komplett	
	NS SS	0.3773 (0.0806, 0.6914)	Hela NS	0.7336 (0.4217, 0.9671)
	Skg	0.5635 (0.2416, 0.8792)	Skg	0.2347 (0.0113, 0.5474)
	KaWBa	0.0592 (0.0001, 0.239)	KaWBa	0.0317 (0, 0.1709)

Tabell 6. Resultatet av MSA när bara vårlekande populationer finns med respektive när baseline är komplett. NS SS=North Sea spring spawners, dvs. vårlekande Nordsjöpopulationer, se tabell 2.

	rank	score	rank	score	rank	score	rank	score	rank	score
Assigned sample	1	%	2	%	3	%	4	%	5	%
/RA001	KF02	36.702	LB	9.863	MA	7.937	OR	7.503	KA	7.29
/RA002	SH	27.095	BB	16.923	MO	10.906	RO	7.212	KO	6.973
/RA003	AB	37.515	BB	15.468	EC	15.406	CW	15.323	FL	7.294
/RA004	KO	43.717	MO	42.856	RO	5.392	FL	3.066	FB	2.668
/RA005	SH	38.978	RO	29.722	EC	5.871	KO	4.444	AB	3.285
/RA006	LB	30.569	KA	11.977	FB	9.541	BB	9.066	RO	8.81
/RA007	WH	25.937	EC	25.028	KO	13.711	FL	13.371	OR	6.097
/RA008	KO	32.094	AB	24.492	SH	11.354	CW	9.673	MA	4.735
/RA009	CW	17.71	AB	12.404	LF	10.307	FB	8.059	RO	6.917
/RA010	CW	50.302	TJ	9.824	WH	8.982	MA	8.969	SH	7.523
/RA011	SH	16.52	AB	15.636	OR	13.883	BB	11.87	MA	10.599
/RA012	MO	94.993	KO	1.411	RO	0.894	CW	0.627	BB	0.505
/RA013	WH	45.161	MO	12.022	FL	11.579	CW	11.565	KO	6.356
/RA014	SH	16.826	RO	13.904	EC	13.679	BB	8.03	AB	7.899
/RA015	TJ	20.272	RO	16.03	RU	12.247	CW	12.125	MA	9.8
/RA016	BB	52.804	CW	5.993	AB	5.266	KO	4.791	FL	4.714
/RA017	CW	82.172	WH	7.122	SH	6.042	OR	1.346	TJ	1.161
/RA018	LF	39.726	SH	26.827	FB	16.209	WH	3.272	EC	2.96
/RA019	LB	25.13	RU	11.954	TJ	11.373	OR	10.061	FL	7.871
/RA020	BB	21.915	OR	20.504	AB	9.737	KA	8.227	CW	6.712
/RA021	RO	84.736	TJ	8.696	FL	3.619	KO	1.391	FB	0.983
/RA022	MO	18.746	FL	16.073	FB	15.431	KO	11.565	WH	8.105
/RA023	MA	70.25	MO	16.513	FB	4.904	EC	2.719	WH	1.518
/RA024	KF03	13.96	WH	13.113	CW	11.655	RU-Ma	11.369	SH	6.35

Tabell 7. De fem populationer som får högst score för var och en av de juveniler som klassificerats som vårlekare. Populationerna är färgkodade enligt vilken region de tillhör: Grön för Kattegat och Östersjön, gulbrun respektive grå för höst- respektive vårlekande Nordsjöpopulationer och rosa Skagerrak. Populationer som hamnat under det kritiska värdet vid geneclass uteslutingnsberäkningar är inramade.

Askerö

Baseline	Vårlekande pop.		Komplett	
	NS SS	0.043 (0, 0.2845)	Hela NS	0.1943 (0.0004, 0.5966)
	Skg	0.8885 (0.6138, 0.9989)	Skg	0.7625 (0.3634, 0.9938)
	KaWBa	0.0685 (0.0002, 0.282)	KaWBa	0.0432 (0, 0.2361)

Tabell 8. Resultatet av MSA när bara vårlekande populationer finns med respektive när baseline är komplett. NS SS=North Sea spring spawners, dvs. vårlekande Nordsjöpopulationer, se figur 2.

	rank	score	rank	score	rank	score	rank	score	rank	score
Assigned sample	1	%	2	%	3	%	4	%	5	%
/AS002	RO	72.41	MO	18.081	SH	3.7	AB	1.519	KO	0.956
/AS003	AB	49.792	FL	15.718	EC	6.775	TJ	6.422	MA	5.89
/AS005	WH	44.654	OR	29.21	MO	12.165	RO	4.874	TJ	3.053
/AS007	TJ	65.245	LB	10.448	KF03	9.021	LF	6.002	KF02	2.18
/AS008	CW	52.656	RO	20.331	TJ	13.36	MO	1.932	OR	1.829
/AS009	BB	38.952	OR	10.359	TJ	10.3	FL	6.202	CW	6.141
/AS010	KA	15.524	SH	14.56	EC	10.47	RU03	8.288	WH	7.435
/AS011	RO	18.082	KA	17.977	RU03	16.501	LB	7.967	KO	5.723
/AS012	TJ	28.226	EC	17.294	BB	13.992	OR	12.484	SH	9.287
/AS015	LB	42.951	KO	14.518	KF03	10.189	WH	8.93	CW	7.628
/AS017	RO	65.662	BB	17.692	FL	7.854	MO	3.748	KO	1.727
/AS018	KO	19.58	SH	15.712	CW	15.387	MA	13.55	OR	10.944
/AS019	CW	17.497	MA	14.573	FL	10.69	SH	8.456	TJ	7.706
/AS020	CW	15.385	FL	12.476	OR	10.758	MA	10.381	AB	9.644
/AS022	RO	91.154	MO	6.729	KF03	0.625	KO	0.261	MA	0.189
/AS023	RO	36.665	CW	23.327	MO	12.341	FL	8.127	MA	6.978

Tabell 9. Färgkodning som i tabell 7

Hunnebo

Baseline	Vårlekande pop.		Komplett	
	NS SS	0.0837 (0, 0.4375)	**Hela NS**	0.1551 (0.0004, 0.5011)
	Skg	0.8796 (0.5183, 0.9993)	**Skg**	0.8213 (0.4703, 0.9965)
	KaWBa	0.0367 (0.0001, 0.1663)	**KaWBa**	0.0236 (0, 0.1348)

Tabell 10. Resultatet av MSA när bara vårlekande populationer finns med respektive när baseline är komplett. NS SS=North Sea spring spawners, dvs. vårlekande Nordsjöpopulationer, se figur 2.

	rank	score	rank	score	rank	score	rank	score	rank	score
Assigned sample	1	%	2	%	3	%	4	%	5	%
/HU001	TJ	49.059	RO	25.616	KO	15.552	FL	2.107	MO	1.911
/HU002	MO	34.667	KO	33.068	FL	15.768	TJ	6.56	RO	6.537
/HU003	MO	18.327	TJ	13.215	OR	10.31	RO	8.334	FL	6.471
/HU004	RO	33.687	MA	26.858	KF03	13.604	WH	5.87	CW	5.278
/HU006	EC	30.856	MO	19.822	CW	10.774	FL	7.291	AB	6.62
/HU007	CW	76.211	TJ	11.094	RO	9.938	OR	0.49	AB	0.417
/HU008	TJ	31.456	LB	13.26	KO	7.985	AB	6.73	LF	6.447
/HU009	CW	39.035	RO	34.399	TJ	13.516	KO	3.005	MO	2.89
/HU010	RO	27.957	FL	23.103	MO	12.385	CW	10.129	OR	5.165
/HU011	WH	22.394	EC	13.854	FL	12.512	AB	10.225	CW	8.754
/HU012	OR	22.394	KO	12.615	FL	11.413	SH	9.241	BB	8.283
/HU013	KF03	50.409	SH	23.732	BB	7.409	KA	5.061	OR	2.566
/HU014	KO	23.519	RO	11.456	CW	10.75	AB	6.945	RU03	6.52
/HU016	CW	27.768	AB	19.171	EC	9.214	BB	8.518	SH	6.766
/HU017	SH	16.827	RO	15.175	AB	15.096	EC	11.377	KO	10.497
/HU018	LF	71.139	MA	13.834	FB	3.715	TJ	2.296	BB	2.048
/HU019	TJ	23.581	FL	23.06	AB	13.963	SH	12.546	FB	5.612
/HU020	MO	75.596	RO	19.628	KO	0.916	OR	0.837	EC	0.797
/HU021	SH	60.539	KO	5.459	WH	5.362	OR	4.585	FB	4.181
/HU022	MO	19.09	EC	13.562	WH	11.138	TJ	8.724	MA	7.911
/HU024	AB	21.109	FL	18.158	TJ	10.747	KO	7.168	WH	6.871

Tabell 11. Färgkodning som i tabell 7

Gullmaren

Baseline	Vårlekande pop.		Komplett	
	NS SS	0.3345 (0, 0.8281)	**Hela NS**	0.6457 (0.1845, 0.9982)
	Skg	0.6141 (0.1142, 0.9959)	**Skg**	0.3187 (0, 0.7891)
	KaWBa	0.0514 (0.0001, 0.2353)	**KaWBa**	0.0357 (0, 0.203)

Tabell 12. Resultatet av MSA när bara vårlekande populationer finns med respektive när baseline är komplett. NS SS=North Sea spring spawners, dvs. vårlekande Nordsjöpopulationer, se figur 2.

	rank	score	rank	score	rank	score	rank	score	rank	score
Assigned sample	1	%	2	%	3	%	4	%	5	%
/GU001	TJ	23.392	RO	16.66	BB	15.967	OR	13.475	FL	13.342
/GU002	BB	37.266	SH	13.53	CW	10.132	FL	7.971	AB	5.69
/GU005	LB	40.119	BB	9.415	LF	6.894	FB	6.271	KF02	6.051
/GU006	BB	23.524	SH	17.097	WH	16.08	CW	7.996	OR	5.934
/GU007	FB	22.425	KO	12.252	RO	12.021	OR	8.683	EC	7.488
/GU010	FL	18.665	OR	17.647	WH	12.325	KO	11.541	CW	7.951
/GU011	FL	51.861	AB	18.784	BB	6.191	OR	4.681	MO	3.998
/GU012	BB	38.895	SH	26.331	OR	10.249	WH	8.484	EC	5.979
/GU015	CW	38.708	WH	24.158	BB	13.873	OR	5.825	FB	4.765
/GU016	KO	22.575	BB	19.353	FL	15.99	OR	9.764	EC	8.418
/GU017	FB	33.279	LB	18.98	KO	11.012	TJ	9.985	AB	6.668
/GU018	CW	13.839	OR	12.926	KO	11.469	WH	9.542	FL	9.056
/GU019	RU03	26.331	KF02	21.66	RU02	9.66	LF	7.719	OR	6.955
/GU020	FL	37.791	CW	15.839	SH	15.688	RO	10.662	WH	7.044
/GU021	RU03	28.218	KF02	18.234	RO	15.857	CW	6.442	RU02	5.291
/GU022	MO	70.606	RO	24.46	FL	2.807	AB	0.972	KO	0.324
/GU023	WH	15.017	EC	14.065	SH	13.77	CW	12.973	RO	10.107

Tabell 13. Färgkodning som i tabell 7

6.2 Höstlekare

	rank	score	rank	score	rank	score	rank	score	rank	score
Assigned sample	1	%	2	%	3	%	4	%	5	%
/HU005	OR	36.402	CW	32.141	TJ	22.178	AB	1.912	FB	1.599
/AS004	KO	72.775	MA	15.647	OR	1.861	SH	1.808	EC	1.333
/AS014	SH	32.487	CW	21.949	MO	14.765	EC	7.468	BB	7.123
/AS021	BB	82.648	SH	4.616	FB	4.596	RO	3.354	EC	0.973
/AS024	CW	39.184	OR	31.922	MO	8.994	MA	5.37	FL	4.976
/GU008	RO	33.796	TJ	20.465	EC	10.136	FB	6.818	MA	4.219

Tabell 14. Resultatet av IA för höstlekare från olika fångstplatser. Färgkodning som i tabell 7

6.3 Vinterlekare

	rank	score	rank	score	rank	score	rank	score	rank	score
Assigned sample	1	%	2	%	3	%	4	%	5	%
/HU015	WH	18.238	BB	13.824	KO	9.165	SH	8.739	CW	8.565
/AS001	AB	28.748	EC	11.545	BB	11.2	TJ	8.003	RO	7.554
/AS013	KF02	34.804	WH	11.884	AB	11.02	EC	6.361	FB	5.989
/AS016	WH	27.766	RO	19.843	MO	18.478	FB	17.078	LF	7.836
/GU003	OR	41.843	FL	12.817	EC	11.071	CW	9.66	WH	5.854
/GU004	SH	20.206	CW	18.154	WH	12.332	KO	11.652	AB	8.14
/GU009	KA	58.351	RO	8.352	EC	7.634	LF	7.268	FB	4.106
/GU013	CW	13.964	WH	12.243	BB	11.578	SH	10.868	FL	8.877
/GU014	MO	17.643	BB	14	KF03	12.731	CW	9.2	SH	7.649

Tabell 15. Resultatet av IA för vinterlekare från olika fångstplatser. Färgkodning som i tabell 7

6.4 BMI jämfört med otolitdata

MSA gjordes också för individer med BMI<0.0005 (men utan otolitdata) för att jämföra med de som klassificerats som vårlekare vid otolitanalyser. Nedan finns resultaten av MSA för dessa grupper av juveniler vid varje fångstområde.

Hunnebo				
	Metod för klassificering			
Baseline	**Otolit**		**BMI**	
Vårlekande populationer	NS SS	0.0837 (0,0.4375)	NS SS	0.4246 (0.0902, 0.7143)
	Skg	0.8796 (0.5183,0.9993)	Skg	0.5287 (0.2378, 0.8711)
	KaWBa	0.0367 (0.0001,0.1663)	KaWBa	0.0467 (0.0001, 0.1953)
Komplett	Hela NS	0.1551 (0.0004, 0.5011)	Hela NS	0.7343 (0.4425, 0.9606)
	Skg	0.8213 (0.4703, 0.9965)	Skg	0.2287 (0.003, 0.5288)
	KaWBa	0.0236 (0, 0.1348)	KaWBa	0.0370 (0, 0.1784)

Tabell 16. Resultaten av MSA utförda på individer som klassificerats som vårlekare vi fångstplatsen Hunnebo. Dels analyserades individer som var vårlekare enligt otolitdata, dels alla som var individer enligt uträknat BMI.

Askerö				
	Metod för klassificering			
Baseline	**Otolit**		**BMI**	
Vårlekande populationer	NS SS	0.043 (0, 0.2845)	NS SS	0.1343 (0, 0.5713)
	Skg	0.8885 (0.6138, 0.9989)	Skg	0.8223 (0.3799, 0.9988)
	KaWBa	0.0685 (0.0002, 0.282)	KaWBa	0.0433 (0.0001, 0.1945)
Komplett	Hela NS	0.1943 (0.0004, 0.5966)	Hela NS	0.3083 (0.0031, 0.7292)
	Skg	0.7625 (0.3634, 0.9938)	Skg	0.6639 (0.2407, 0.9851)
	KaWBa	0.0432 (0, 0.2361)	KaWBa	0.0278 (0, 0.1564)

Tabell 17. Resultaten av MSA utförda på individer som klassificerats som vårlekare vi fångstplatsen Askerö. Dels analyserades individer som var vårlekare enligt otolitdata, dels alla som var individer enligt uträknat BMI.

Råssö				
	Metod för klassificering			
Baseline	**Otolit**		**BMI**	
Vårlekande populationer	NS SS	0.3773 (0.0806, 0.6914)	NS SS	0.5307 (0.3332, 0.7223)
	Skg	0.5635 (0.2416 ,0.8792)	Skg	0.4473 (0.2529, 0.6464)
	KaWBa	0.0592 (0.0001, 0.239)	KaWBa	0.0219 (0, 0.0948)
Komplett	Hela NS	0.7336 (0.4217, 0.9671)	Hela NS	0.7564 (0.5543, 0.9239)
	Skg	0.2347 (0.0113, 0.5474)	Skg	0.2261 (0.0617, 0.4276)
	KaWBa	0.0317 (0, 0.1709)	KaWBa	0.0175 (0, 0.087)

Tabell 18. Resultaten av MSA utförda på individer som klassificerats som vårlekare vi fångstplatsen Råssö. Dels analyserades individer som var vårlekare enligt otolitdata, dels alla som var individer enligt uträknat BMI.

Största skillnaderna mellan individer som klassificerats mha BMI respektive otolitanalyser är i andelarna från Nordsjön och Skagerrak. Skillnaderna i andelar från Kattegat och Östersjön är små, dock med smalare sannolikhetsinterval för de som klassificerats mha BMI (vilka också är många fler till antalet).
Både när baseline bara innehåller vårlekare och när den är komplett är andelarna från Skagerrak större bland juveniler som klassificerats mha otolitanalyser och vice versa när det gäller andelarna från Nordsjön.

För individerna som klassificerats mha BMI ökar andelarna från Norsjön när även de höstlekande populationerna tas med i baseline. Det kan bero på att BMI inte är någon speciellt bra indikator på lektyp men eftersom andelarna från Nordsjön ökar också bland juveniler som klassificerats mha otolitanalyser är en annan möjlig förklaring återigen osamplade populationer.
Om BMI verkligen är en bra indikator på lektyp, varför är då resultaten av MSA så olika de som gjorts för juveniler som klassificerats mha otoliter? Kankse har juveniler från vårlekande Nordsjöpopulationer lägre BMI än de från Skagerrak. Ett sätt att undersöka denna fråga är att dela in de vårlekande (enligt otolitdata) juvenilerna i en grupp med lågt BMI och en annan med högt BMI och göra nya MSA för vardera av grupperna. MSA för juveniler som klassificerats enbart mha BMI borde då likna gruppen med lågt BMI men ha lägre andelar från Skagerrak och större andelar från Nordsjön än gruppen med högt BMI. På grund av de stora skillnaderna i resultat av MSA mellan de olika fångstområdena skulle en sådan analys behöva göras för varje fångstområde, men tyvärr finns det inte tillräckligt med juveniler med otolitdata för att dela upp dem ytterligare med avseende på BMI.

7. Diskussion

Något som gör det svårt att skatta ursprung för arten sill är den låga genetiska differentieringen mellan populationer. Även när andelarna från bara tre regioner skattas blir sannolikhetsintervallen mycket breda. Vid IA med individer med känt urprung (inte redovisade) blev andelen korrekt assignade individer bara runt 2/3. Därför har Geneclass 2 främst använts till att göra uteslutningsberäkningar.

Några slutsatser går ändå att dra från analyserna. Individerna från Kattegat och Östersjön utgör en klar minoritet vid alla fångstområden. Nordjsön verkar dominera som ursprung och det är troligt att det finns osamplade populationer här, antagligen vårlekande men genetiskt lika de höstlekande Nordsjöpopulationerna.
BMI är som man kunnat förvänta lägre hos de vårkläckta juvenilerna men varierar också mycket mellan olika vårlekande populationer och verkar vara lägst hos vårlekande juveniler från Nordsjön.

Det har tidigare gjorts ett mycket stort antal studier där man, med metoder liknande de som använts i den här uppsatsen, försökt klassificera ursprung för både sill och andra arter. Studier av sill är mest relevanta när det gäller jämförelser med mina resultat.
I artikeln ”Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring.” undersökte Ruzzante et al ett stort antal ansamlingar av sill i både Nordsjön och Skagerrak. Det som gör den studien lämplig för jämförelse med mina resultat är bland annat att samma programvara använts för MSA och att fångsterna klassificerats till samma tre regioner, med den enda skillnaden att Risør inte ingick som potentiell ursprungspopulation i region Skagerrak.
Fångsterna som analyserades i Ruzzante et al samlades in under år 2002 och 2003, dels på sommaren (juli) och dels på vintern (november och december) och bestod av individer av blandade åldrar. Separata MSA gjordes efter att ha delat upp fångsterna efter år, årstid och åldersgrupp och region där de fångades. Dock analyserades alla fångster från en viss region, år och årstid gemensamt. Mest intressanta för jämförelse med mina resultat är MSA som utförts på juveniler fångade i Skagerrak. Fyra sådana MSA utfördes, en för varje år och årstid. Under år 2002 är andelen från Nordsjön drygt 70% för både sommarfångsterna och vinterfångsterna, medan motsvarande siffra för 2003 är runt 90% (också för båda sommar- och vinterfångsterna). Även sannolikhetsintervallen visar större överrensstämmelse för de olika årstiderna under respektive år än för individer fångade samma årstid men olika år. Konfidensintervallen är breda, även om de är något smalare mina, förmodligen på grund av större samplestorlekar.
Under både sommaren och vintern 2003 är andelarna individer från de båda övriga regionerna mycket låga. Under år 2002 på sommaren och vintern är andelarna individer från Kattegatt och västra Östersjön (kallas region 3 i Ruzzante et al) runt 24% (sannolikhetsintervallet går från ca 17% till 31%) respektive 16% (sannolikhetsintervallet går från ca 4% till 27%), medan punktskattningarna för andelarna från Skagerrak är

några få procent med 0% under båda årstiderna och 0% som nedre gräns för båda sannolikhetsintervallen. Detta är nästan precis motsatt resultat jämfört med mina MSA.
Sammanfattningsvis visar jämförelsen med artikeln av Ruzzante et al ganska god överrenstämmelse när det gäller andelarna individer från Nordsjön samt under år 2003 även när det gäller andelarna från Kattegatt och västra Östersjön. De skillnader som finns i resultat kan bero på att juvenilerna som undersökts i den här uppsatsen fångades på hösten, att skillnader förekommer mellan olika år och att Ruzzante et al gjorde gemensamma analyser för ett stort antal fångster från olika delar av Skagerrak medan mina fångster alla kom från Sveriges västkust.

Det finns många tänkbara felkällor som påverkar korrektheten i skattningarna som gjorts negativt. Att beräkningsmodellernas antaganden, t.ex. oberoende mellan loci, inte håller är en sådan. Att vissa bidragande populationer inte blivit samplade påverkar förståss korrektheten i resultaten. Till och med antagandet att varje population leker vid en bestämd tid på året är en förenkling eftersom det finns belägg för att populationer har bytt lekperiod och att enstaka individer migrerat till en ny population. Samples från populationerna kan också vara orepresentativa, t.ex. på grund av stor genetisk variation mellan olika år även om arten förekommer i mycket stora populationer och variationen mellan olika år antas vara liten.
Sambandet mellan BMI och lektyp har vad jag vet inte undersökts tidigare på det sätt som presenteras i den här uppsatsen. På grund av det och de små stickprovsstorlekarna bör man inte dra för långtgående slutsatser bara utifrån en individs BMI, men resultaten som presenterats tyder ändå på att det finns ett interesant samband som är värt att undersöka ytterligare för framtida assignments.
En annan grupp av felkällor är experimentella fel, t.ex. vid bestämning av lektyp eller bestämning av individers genotyper. Genotypdata undersöktes därför extra hos de juveniler som uteslöts från samtliga vårlekande populationer även om de enligt otolitdata var vårlekare samt hos individen som uteslöts från samtliga populationer i baseline, för att se om fel genotypning av enstaka alleler kan ha varit orsaken till dessa juveniler utesluts från så många populationer.
Individen som utesluts från samtliga populationer, RA074, hade en allel som inte finns hos någon annan juvenil vid just det locuset. Däremot finns den hos fyra individer i baseline, två från Karmøy, en från Møre och en från Måseskär. Om uteslutningsberäkningarna i geneclass körs utan detta locus utesluts individen ändå från de flesta populationer, med undantag för Risør och några höstlekande populationer. Övriga av dess alleler finns också hos andra juveniler vid respektive locus.
De två juveniler som utesluts från alla vårlekande populationer trots att de är vårlekare enligt otolitdata, RA003 och RA023, har alleler som också finns hos andra juveniler vid respektive locus. Däremot har de alleler som inte finns vid respektive locus hos några av de andra juvenilerna som är vårlekare enligt otolitdata. RA003 har en sådan allel och RA023 har tre stycken.

8. Referenser

Pella J, Masuda M (2001) Bayesian methods for analysis of stock mixtures from genetic characters. Fish Bull 99: 151−167

Piry S, Alapetite A, Cornuet, J.-M., Paetkau D, Baudouin, L., Estoup, A. (2004) GeneClass2: A Software for Genetic Assignment and First-Generation Migrant Detection. Journal of Heredity 95:536-539.

Cornuet JM, Piry S, Luikart G, Estoup A, Solignac M (1999) New methods employing multilocus genotypes to select or exclude populations as origins of individuals. *Genetics* **153**, 1989-2000.

Rannala B, Mountain JL (1997) Detecting immigration by using multilocus genotypes. *Proceedings of the National Academy of Sciences USA* **94**, 9197-9201.

Paetkau D, Slade R, Burden M, Estoup A (2004) Direct, real-time estimation of migration rate using assignment methods: a simulation-based exploration of accuracy and power. *Molecular Ecology* **13**:55-65.

Clausen LAW, Bekkevold D, Hatfield EMC, Mosegaard H (2007) Application and validation of otolith microstructure as stock identifier in mixed Atlantic herring (Clupea harengus) stocks in the North Sea and western Baltic. ICES J Mar Sci 64:1–9

Ruzzante DE, Mariani S, Bekkevold D, André C and others (2006) Biocomplexity in a highly migratory pelagic marine fish, Atlantic herring. Proc R Soc Lond B Biol Sci 273: 1459−1464

L. Baudouin, S. Piry, and J. M. Cornuet
Analytical Bayesian Approach for Assigning Individuals to Populations
J Hered (2004) 95(3): 217-224

Manel S, Gaggiotti OE, Waples RS (2005) Assignment methods:
matching biological questions with appropriate techniques.
Trends Ecol Evol 20:136–142

Masuda M (2002) User's Manual for Bayes: Bayesian Stock
Mixture Analysis Program. National Marine Fisheries
Service, Alaska Fisheries Science Center, Auke Bay Laboratory,
Juneau, AK

Weir, B. S., och C. C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. Evolution 38: 1358–1370.

Altman, Douglas G. (1991) Practical Statistics for Medical Research, London: Chapman & Hall

http://www1.montpellier.inra.fr/URLB/GeneClass2/Help/index.htm

Dirichlet distribution. L.N. Bol'shev (originator), Encyclopedia of Mathematics. URL:
http://www.encyclopediaofmath.org/index.php?title=Dirichlet_distribution&oldid=14736

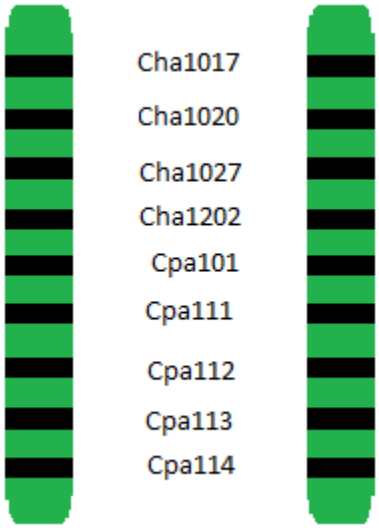
Jaworska, N. och Chupetlovska-Anastasova, A. (2009) A Review of Multidimensional Scaling (MDS) and its Utility in Various Psychological Domains.
Tutorials in Quantitative Methods for Psychology 5(1):1-10.

9. Appendix

	Cha1017		Cha1020		Cha1027		Cha1202		Cpa101		Cpa111		Cpa112		Cpa113		Cpa114			Längd (mm)	Vikt (g)	Lektyp enligt otolit
AS01	164	192	172	184	114	170	112	120	218	234	278	278	312	334	172	184	232	240		96	5,2	w_d
AS02	176	180	184	188	126	162	96	112	234	246	278	278	312	326	136	148	200	204		92	4,3	s_d
AS03	164	172	196	216	150	170	112	116	214	250	270	278	330	330	144	176	224	240		90	4	s_d
AS04	184	188	172	180	150	154	112	116	210	230	278	278	342	348	128	148	204	228		91	4,4	s_d
AS05	168	168	152	184	130	146	112	116	194	238	278	286	318	346	144	172	224	248		75	2,3	s_d
AS07	164	168	120	172	126	158	112	120	226	230	278	290	306	322	148	152	216	220		83	3,2	s_d
AS08	168	176	180	192	114	154	112	112	206	218	278	278	310	320	152	176	204	216		91	4,3	s_d
AS09	168	180	176	180	142	150	116	120	222	238	270	278	334	338	132	156	220	220		87	3,4	s_d
AS10	164	164	200	208	146	162	112	112	206	214	278	278	342	346	128	160	212	224		85	3,6	s_d
AS11	168	172	168	192	158	158	96	112	230	230	278	286	330	334	128	160	212	236		84	3,2	s_d
AS12	176	180	176	212	114	114	108	116	222	246	278	278	292	312	148	188	208	208		92	4,4	s_d
AS13	164	168	176	208	146	158	96	112	214	242	278	278	322	330	120	156	204	256		91	4,3	s_d
AS14	168	188	160	180	114	158	116	116	226	226	270	278	310	322	160	160	216	228		107	6,6	w_d
AS15	176	196	176	204	114	126	112	112	230	242	278	278	334	334	132	164	204	240		90	3,7	s_d
AS16	168	176	176	176	122	154	104	112	218	242	266	278	322	334	164	180	200	264		78	2,5	s_d
AS17	172	180	184	198	114	114	112	120	206	226	278	278	338	346	144	168	216	216		95	5,1	w_d
AS18	168	168	192	224	118	138	96	112	214	222	278	282	322	346	152	156	224	228		85	3,6	s_d
AS19	164	184	176	204	126	166	112	132	206	250	278	278	326	346	152	152	224	244		97	4,7	s_d
AS20	176	180	176	180	158	166	96	132	214	218	278	278	322	338	156	160	212	224		100	5,3	s_d
AS21	164	188	164	176	154	170	112	116	222	222	278	278	340	350	124	160	212	212		111	8,1	s_d
AS22	168	168	172	192	154	174	96	112	218	234	278	278	312	334	152	160	204	212		95	4,3	s_d
AS23	168	168	176	188	142	170	96	132	222	258	278	282	346	350	120	140	208	224		95	5,2	s_d
AS24	168	176	184	220	126	194	112	112	210	222	278	278	322	342	152	152	224	236		110	7,6	s_d
AS25	160	168	172	192	134	134	96	124	234	246	278	278	294	302	120	168	220	224		90	4,2	s_d

Tabell 19

Tabell 19 visar delar av datasetet från en av de fångster som skall klassificeras, den från Askerö. Genetiska data för samma loci finns i HERGEN-databasen. Kolumnen längst till vänster innehåller beteckningar för individerna i fångsten. Följande 18 kolumner innehåller genetiska data, med namn på aktuellt locus på översta raden (två kolumner per locus, en för varje allel vid locuset). Siffrorna i dessa kolumner är ett mått på längden av allelerna, men längden är i sig inte intressant för analyserna i den här uppsatsen. Vad som har betydelse är bara att alleler av olika längd har olika proportioner i de populationer som juvenilerna skall klassificeras till.



Figur 5: Identisk med figur 1, en bild av två kromosomer och de loci namngivna där allelerna förekommer.