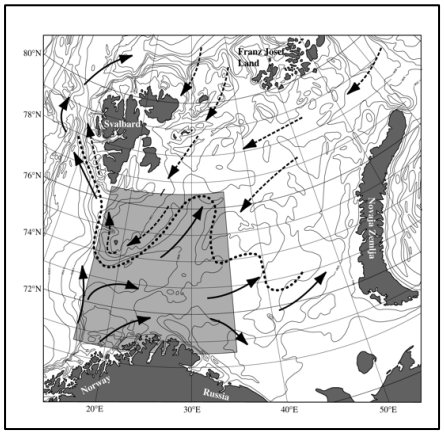
다변량 통계학 TERM PROJECT

바렌츠해의 생물지리정보

201611539 하성진

서 론

1. 분석 내용



바렌츠해는 북극해의 일부로 북서쪽으로는 스발바르 제도, 북동쪽은 젬랴프란츠요세프 제도, 동쪽은 노바야제믈랴 제도에 둘러싸여 있으며, 노르웨이와 러시아 국경에 맞닿아 있는 바다이다. 여기에는 현재 다양한 어종이 분포하고 있다. 이번 분석에선 바렌츠해의 회색을 색칠된 지역을 여러 개의 구역으로 나누어 각 구역의 환경적 특징과 어종의 종류와 개체수를 조사해보고, 조사된 데이터를 바탕으로 다변량 분석을 통해 구역별 특징 및 변수의 특징들을 알아보고자 한다.

2. 분석 데이터

(1)변수

Environmental variables
Latitude (N): sampling station's latitude
Longitude (E): sampling station's longitude
Depth (m): max station's depth
Temperature (C): water temperature at station

Species abundance data
List of species and their abbreviations in data set

Species abundance data
List of species and their abbreviations in data set

Abb.	Scientific name	Family	Common name
An de	*Anarhichas denticulatus*	Anarhichadidae	Jelly wolffish/Arctic catfish
An lu	*Anarhichas lupus*	Anarhichadidae	Wolffish/Atlantic cafish
An mi	*Anarhichas minor*	Anarhichadidae	Spotted wolffish/catfish
Le de	*Leptagonus decagonus*	Agonidae	Atlantic poacher
Cl ha	*Clupea harengus*	Clupeidae	Herring
Ar at	*Artediellus atlanticus*	Cottidae	Atlantic hookear sculpin
Tr spp	*Triglops murrayi*	Cottidae	Moustache/mailed sculpin
Tr spp	*Triglops pingelii*	Cottidae	Ribbed sculpin
Ca re	*Careproctus reinhardti*	Cyclopteridae	Longfin seasnail
Cy lu	*Cyclopterus lumpus*	Cyclopteridae	Lumpsucker
Bo sa	*Boreogadus saida*	Gadidae	Polar cod
Ga mo	*Gadus morhua*	Gadidae	Cod
Me ae	*Melanogrammus aeglefinus*	Gadidae	Haddock
Mi po	*Micromesistius poutassou*	Gadidae	Blue whiting
Tr es	*Trisopterus esmarkii*	Gadidae	Norway pout
Be gl	*Benthosema glaciale*	Myctophidae	Glacier lanternfish
Ma vi	*Mallotus villosus*	Osmeridae	Capelin
Pa bo	*Pandalus borealis*	Pandalidae	Shrimp
No rk	*Notolepis rissoi kroyeri*	Paralepididae	White barracudina
Hi pl	*Hippoglossoides platessoides*	Pleuronectidae	Long rough dab
Re hi	*Reinhardtius hippoglossoides*	Pleuronectidae	Greenland halibut
Ra ra	*Raja radiata*	Rajidae	Starry ray
Se ma	*Sebastes marinus*	Scorpaenidae	Golden redfish
Se me	*Sebastes mentella*	Scorpaenidae	Deepwater redfish
Le ma	*Leptoclinus maculatus*	Stichaeidae	Spotted snake blenny
Lu la	*Lumpenus lampraetaeformis*	Stichaeidae	Snake blenny
Ly es	*Lycodes esmarkii*	Zoarcidae	Esmark´s eelpout
Ly eu	*Lycodes eudipleurostictus*	Zoarcidae	Eelpout (ncn)
Ly pa	*Lycodes pallidus*	Zoarcidae	Pale eelpout
Ly re	*Lycodes reticulatus*	Zoarcidae	Arctic eelpout
Ly se	*Lycodes seminudus*	Zoarcidae	Eelpout (ncn)
Ly va	*Lycodes vahlii*	Zoarcidae	Vahl´s eelpout

변수는 크게 두종류로 나뉜다. 환경적 변수와 종의 다양성 변수이다. 환경적 변수는 각지역의 위도, 경도, 심도, 기온 등이 있으며, 종 다양성 변수로는 위와 같이 다양한 생물들이 해당한다. 이번 데이터의 경우 총 4개의 환경적 변수와 총 30개의 종 다양성 변수를 이용할 예정이다. (또한, 표에서 데이터 변수를 나타낼 수 있는 공간의 한계로, 변수의 이름을 간략하게 줄여 표현하여 사용할 예정이다)

(2)개체

ID No	Latitude	Longitude	Depth	Temperature	Re_hi	An_de	An_mi	Hi_pl	An_lu	Me_ae	Ra_ra	Mi_po	Ar_at	No_rk	Lu_la	Ma_vi	Bo_sa	Cy_lu	Cl_ha	Se_me	Le_de
400	72.28	25.17	273	3.35	0	0	0	87	0	294	0	0	3	0	0	22	0	0	1	181	0
401	72.65	24.10	335	3.35	0	5	0	40	0	63	1	2	0	0	0	7	0	2	0	256	0
402	73.02	25.43	438	1.85	8	0	0	82	0	30	3	5	0	8	0	0	0	0	0	1107	0
403	73.25	26.80	440	2.15	18	1	1	68	0	42	11	0	0	1	0	8	0	0	0	637	0
404	73.53	28.07	382	2.15	5	1	1	60	0	44	0	2	0	2	0	1	0	0	0	0	1
405	73.80	29.42	358	1.85	3	1	0	187	0	60	11	0	1	2	0	45	0	0	0	166	0
406	74.12	30.83	315	0.65	1	1	1	22	0	4	3	0	7	1	0	19	0	0	0	13	0
407	74.43	30.92	285	0.95	0	1	1	39	0	3	2	0	10	0	0	2	0	0	0	9	4
409	74.33	32.15	225	0.35	0	0	0	38	0	0	1	0	10	0	4	3	1	0	0	0	6
410	74.75	31.25	300	1.35	5	1	2	106	0	1	6	0	19	0	5	5	0	0	0	14	5
411	75.08	31.43	347	0.55	6	2	0	253	0	0	5	0	10	0	1	67	31	0	0	0	6
412	75.40	31.85	306	0.75	5	0	1	87	0	1	3	0	5	0	9	6	6	0	0	0	13
413	75.72	32.10	311	1.15	1	0	0	145	0	0	0	0	12	0	10	7	207	0	0	3	27
414	75.93	33.42	272	0.35	1	0	0	222	0	0	0	0	19	0	7	3	138	1	0	0	16
415	75.73	32.55	306	0.55	1	0	0	73	0	0	1	0	11	0	4	4	66	0	0	0	32
416	75.47	30.50	375	1.05	26	1	0	98	0	0	2	0	6	0	2	42	35	0	0	48	49
417	75.82	30.77	327	0.95	7	0	0	58	0	0	3	0	42	0	10	42	3647	1	0	0	0
418	75.62	29.10	317	0.95	7	4	1	161	0	0	4	0	116	0	1	80	881	1	0	0	20
419	75.62	28.05	256	0.35	0	0	0	128	0	0	0	0	13	0	0	398	68	0	0	0	8
420	75.18	27.63	278	0.55	0	0	0	130	0	1	3	0	8	0	0	119	1	0	0	27	4
428	75.22	29.05	342	0.85	25	1	0	99	0	1	1	0	13	0	2	49	168	0	0	59	36
429	75.12	30.33	380	1.25	13	0	0	131	0	0	9	0	4	0	0	12	5	0	0	73	23
430	74.80	30.00	384	1.05	7	0	1	75	0	0	4	0	0	0	0	43	3	0	0	62	3
431	74.48	29.78	373	1.15	9	1	0	78	0	6	3	0	1	1	0	84	14	0	0	185	0
432	74.17	29.48	372	1.65	10	3	1	110	0	12	7	0	4	0	0	16	0	0	0	232	1
439	74.18	28.28	394	1.85	4	5	0	50	0	5	5	0	0	0	0	36	0	0	0	82	0
440	74.53	28.50	395	1.35	6	3	0	83	0	1	5	0	5	2	0	32	0	0	0	184	1
441	74.85	28.77	362	0.85	5	3	0	67	0	0	7	0	17	0	0	78	6	0	0	30	10

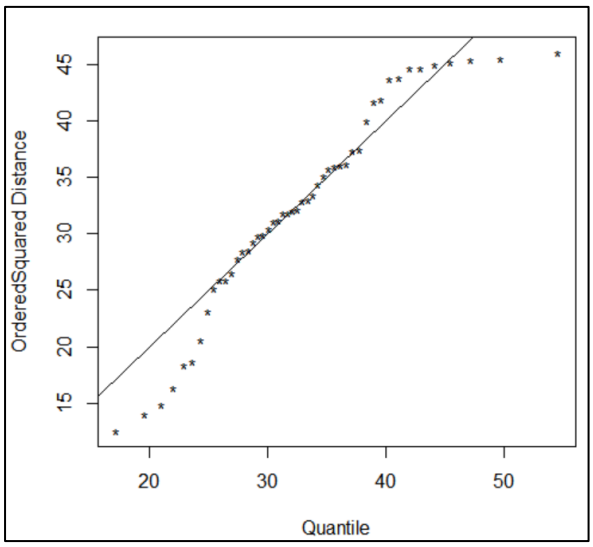
(공간의 한계로 441번 구역까지의 개체와 Le_de(Leptagonus decagonus)까지의 종다양성 변수 까지만 표현하였음, 나머지 데이터는 원데이터 참고바람)

여기서 개체는 바렌츠 해의 일정한 기준에 의해 나누어진 구역에 해당한다. 각 구역을 숫자로 명명하고 있으며, 이번 분석에선 400~465번에 해당하는 구역대를 집중적으로 분석해볼 예정이다.

본론

다변량 통계학에서의 많은 기법과 접근방법들은 다변량 정규분포를 필요로 한다. 따라서 정규성 검정을 위해 마할라노비스 거리를 활용해보고자 한다.

$$d_m(X,Y) = \left[(X-Y)^T S^{-1}(X-Y)\right]^{1/2}$$

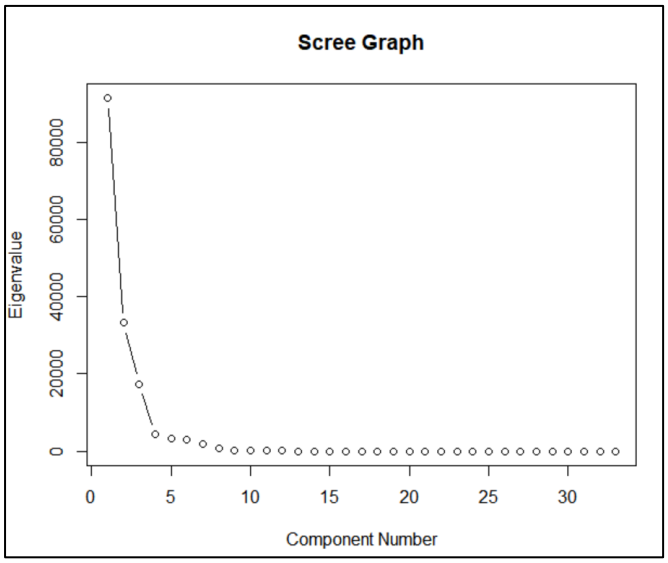



```
> rq=cor(cbind(q,m))[1,2]; rq
[1] 0.9633116
```

마할라노비스 거리를 이용한 피어슨 상관계수의 경우 거의 1에 가까운 값(0.9633116)으로 카이제곱그림의 직진성이 인정되며, 자료는 다변량 정규성을 만족한다고 할 수 있다.

1. 주성분 분석(PCA)

현재 34개의 변수가 존재하는 만큼, 저차원공간에서의 표현은 한계가 있다. 차원축소를 위해 주성분분석(PCA)를 진행해보고자 한다.

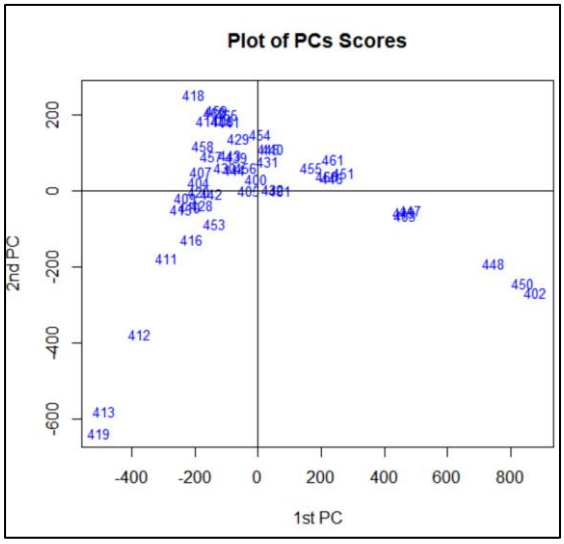


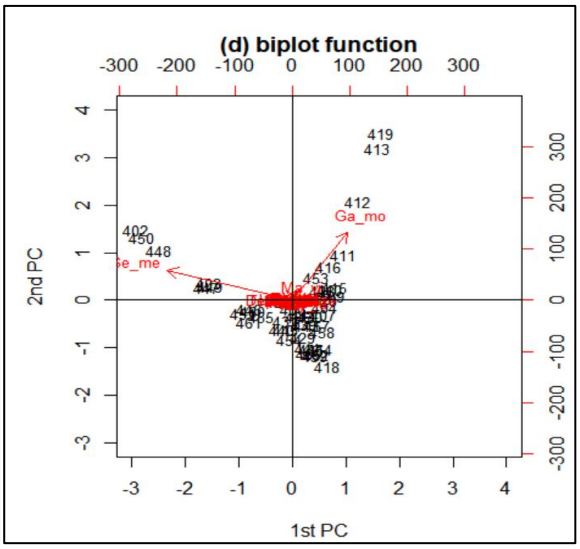
	제1주 성분	제2주 성분	제3주 성분	제4주 성분	제5주 성분	.....
Gof	58.69	21.32	11.16	2.82	2.07	....

현재 제1,2주성분으로 설명력이 70%를 넘어, 2개의주성분만으로도 충분히 설명이 가능하다.

	[,1]	[,2]
400	-5.663863	32.161888
401	72.574817	2.541008
402	881.958597	-267.327172
403	466.708217	-64.205692
404	-186.459963	25.155138
405	-26.388903	2.430118
406	-111.305203	185.882728
407	-179.036063	52.156088
409	-228.564843	-18.148022
410	-214.888283	-40.897142
411	-290.423823	-175.464532
412	-374.550783	-375.683592
413	-485.637523	-579.833262
414	-158.707663	184.938488
415	-241.399843	-47.657022
416	-208.973523	-126.622912
418	-202.653523	255.142638
419	-503.508823	-636.803612
420	-187.367223	-1.159032
428	-177.267523	-35.914712
429	-62.192603	140.583828
430	-104.858183	61.730258
431	32.124877	79.482218
432	49.612797	3.839128
439	-68.493823	89.480718
440	49.779657	112.615328

441	-90.435103	182.124618
442	-145.907523	-7.392312
443	-87.886743	96.613348
444	-73.236123	55.864868
445	34.219317	112.235208
446	237.637297	34.496868
447	483.055997	-48.609432
448	748.438097	-188.699222
449	462.795037	-54.675312
450	842.635857	-241.533402
451	274.369157	48.662928
453	-137.898603	-86.509672
454	9.651117	151.196558
455	171.208897	61.337578
456	-37.962283	63.767408
457	-146.805883	93.809958
458	-174.308723	121.345688
459	-131.342163	212.878938
460	219.601057	39.936298
461	241.186777	84.382838
462	-137.482763	208.705388
465	-95.949663	201.637988





제1 주성분에서 413,419,412 구역이 가장 높은 음의 상관관계를 보이며, 448,450,402구역이 가장 높은 양의 상관관계를 보인다. 제2 주성분에선, 419, 413, 412, 402 등이 가장 높은 음의 상관관계를 보이고 있다. 또한, Ga-mo 변수와 Se-me 변수는 제1,2주성분과 높은 상관관계를 보이고 있다.

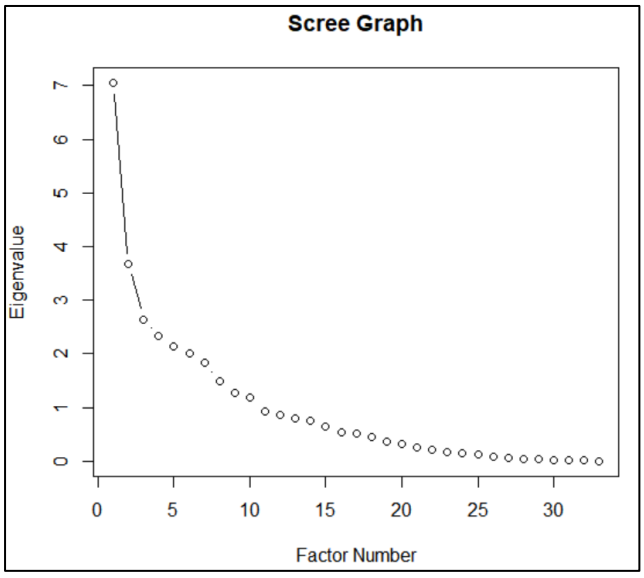
ID No	*Ga_mo*	*Se_me*
400	142	181
401	141	256
402	62	1107
403	45	637
404	230	0
405	172	166
406	44	13
407	196	9
409	277	0
410	295	14
411	439	0
412	672	0
413	904	3
414	53	0
415	308	0
416	369	0
417	0	48
418	5	0
419	903	0
420	227	27
428	270	59
429	66	73

(노란색 표시를 해둔 개체는 제1주성분에서 음의 상관관게를 가지는 412, 413, 419)
(주황색 표시를 해둔 개체는 제1주성분에서 양의 상관관게를 가지는 402)
위에 제시된 주성분 분석을 참고할 때, 제1 주성분의 경우, Ga-mo 계수와 Se-me 계수와 연관성이 높다고 할 수 있다. 특히 Ga-mo 계수와는 음의 상관관계를 가지며, Se-me 계수와는 양의 상관관계를 가진다고 할 수 있다.

ID No	total
400	859
401	531
402	1315
403	837
404	354
405	674
406	131
407	278
409	370
410	486
411	793
412	817
413	1192
414	490
415	672
416	718
418	552
419	1500
420	534
428	606
429	341
430	353
431	446
432	550
439	306
440	375
441	242
	550

(Total 변수는 개체별 기존의 종의 다양성 변수의 합을 표현한 것, 기존 변수에는 존재하지 않음)
(노란색 표시를 기준으로 총합이 높은 개체는 419, 413, 412, 411 등의 개체가 해당한다.)
제2 주성분의 경우, 종의 다양성 변수에 있어 개체수의 합에 영향을 받고 있다. 특히 합계가 클수록 음의 성질을 띄기 때문에 종의 다양성 변수와 음의 상관관계를 보인다고 할 수 있다.

(2) 인자분석



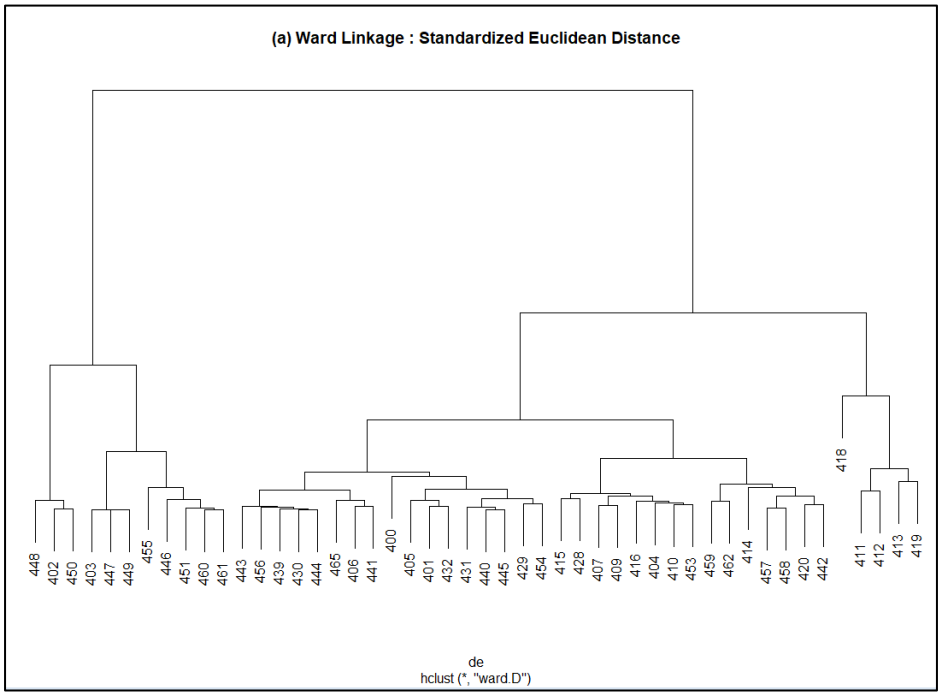
	제1주성분	제2주성분	제3주성분	제4주성분	제5주성분	제6주성분
Gof	21.348	11.131	8.011	7.067	6.472	6.082

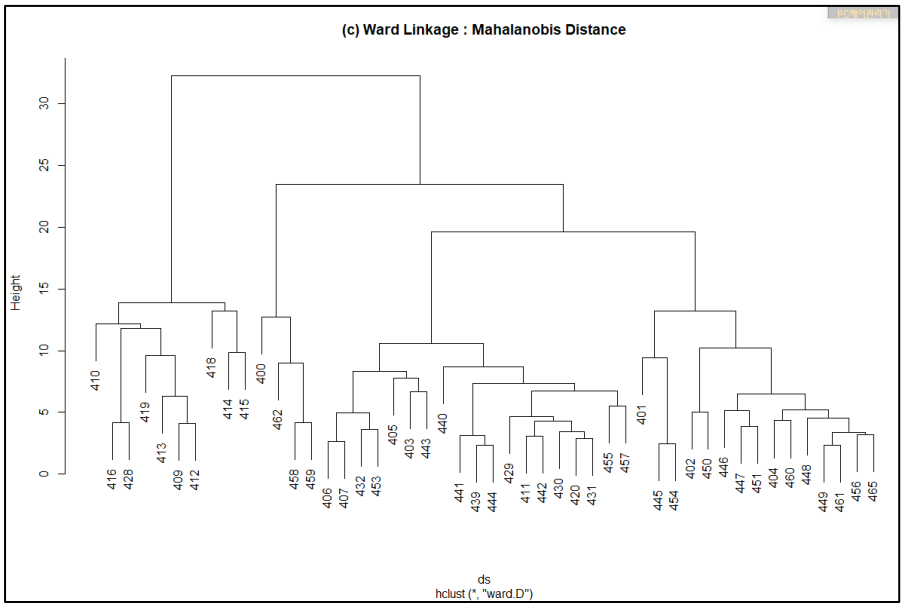
인자분석에 앞서 PCFA를 수행하여 각 주성분들의 설명력을 측정해본 결과, 설명력이 70%이상을 맞추기 위해선 6개이상의 공통인자를 이용해야만이 가능하다. 이는 결국 기본적인 PCFA만으로도, 총 7개이상의 인자에 대한 인자적재 값이 필요하며, 21개의 인자적재 그림이 필요하게된다. 따라서 인자분석에 있어 이번 자료는 비효율적이며, 적합하지 않은 자료랗 할 수 있다. 아쉽게도 이번 분석에서는 인자분석을 활용하지 못할 것으로 보인다.

(3)군집분석

a. 와드연결법

주성분 분석에서 얻은 데이터 분석을 바탕으로, 좀더 자세한 분석을 진행하기 위해 군집분석을 진행하려 한다. 먼저, 계층, 비계층 군집분석에 앞서 와드 연결법을 통해 군집을 분석해보았다. 결과는 아래와 같다





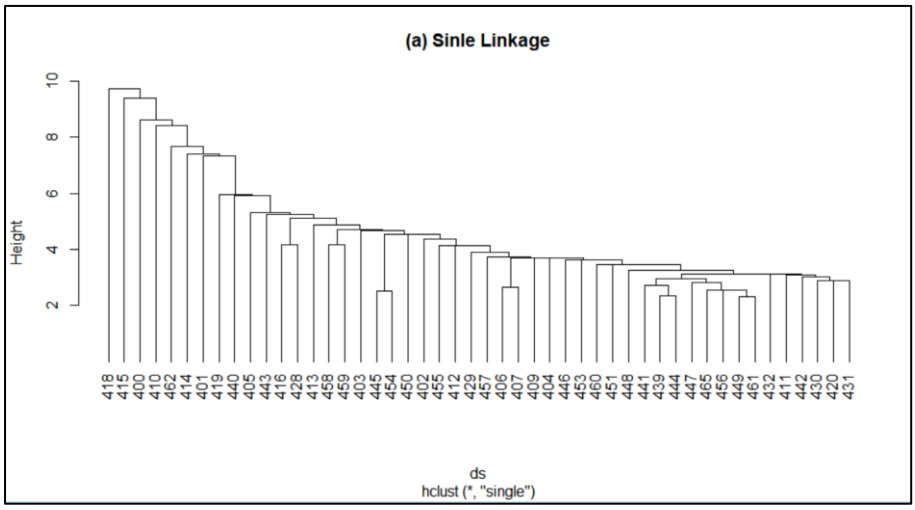
	그룹1	그룹2	그룹3
유클리드	402,403,446,447,448,449, 450,451,455,460,461	411,412,413,418,419	그외 나머지
마할라노비스	401,402,404,416,418, 447,449,450,454,460,	409.410,413,414,415, 416,419,428	그외 나머지
특징	Se-me계수가 크다	Ga-me 계수가 크다	Se-me 계수와 Ga-me계수가 작으며, 종의 종류 및 양이 적다

결과를 정리한 표를 참고하면, 현재 크게 그룹을 3가지로 나눌 수 있으며 각각 그룹은 주성분 분석과 비슷하게 Se-me 계수와 Ga-me 계수에 큰 영향을 받고 있다.

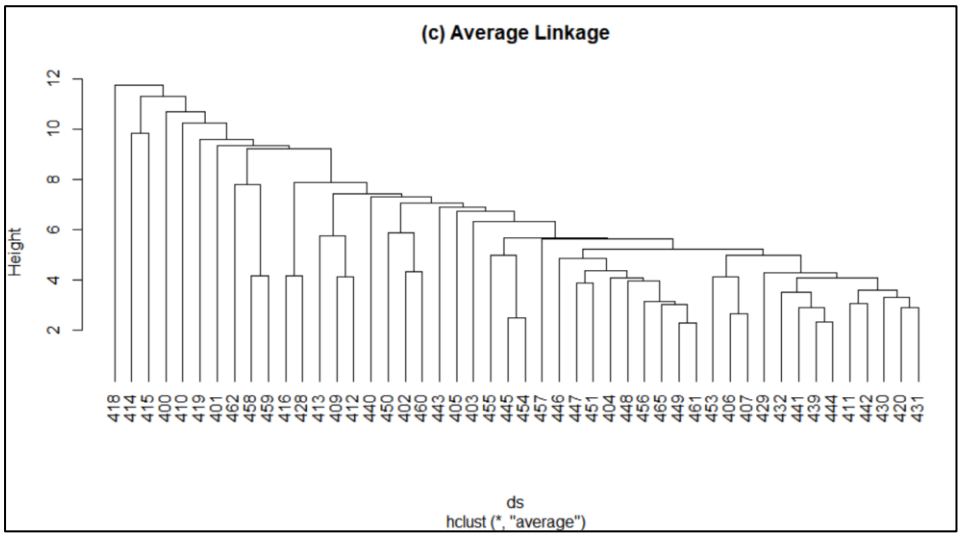
b. 계층군집 분석

와드연결법 분석을 바탕으로 계층 군집분석도 함께 실행해보려 한다.

(단일연결법)



(평균 연결법)



규모가 큰 자료 특성상, 개체들이 소규모의 군집을 이루고 있으며, 이는 유의미하다고 볼 수 없다. 따라서, 계층 군집분석이 아닌 비계층 군집분석을 활용하는 것이 적절할 것으로 보인다.

c. 비계층 군집분석

(k-평균법)

	company	cluster
403	403	1
405	405	1
406	406	1
407	407	1
409	409	1
411	411	1
420	420	1
429	429	1
430	430	1
431	431	1
432	432	1
439	439	1
440	440	1
441	441	1
442	442	1
443	443	1
444	444	1
445	445	1
446	446	1
447	447	1
453	453	1
454	454	1
455	455	1
456	456	1
457	457	1

	company	cluster
400	400	2
401	401	2
402	402	2
404	404	2
448	448	2
449	449	2
450	450	2
451	451	2
458	458	2
459	459	2
460	460	2
461	461	2
462	462	2
465	465	2

	company	cluster
412	412	3
413	413	3
414	414	3
415	415	3
419	419	3

	company	cluster
410	410	4
416	416	4
418	418	4
428	428	4

	그룹1	그룹2	그룹3	그룹4
k평균법	403,405,406,407,409 ,410,420 429-432 439-447 453-457	400,401,402,404, 448,449,450,451 458,459,460,461,462 465	412,413,414,415,419	410,416,418,428
특징	기온 변수, Latitude 변수, Tr-spp 변수가 낮다	기온(Temperature) 변수가 크다	Latitude 변수가 크다.	Tr-spp 변수가 크다

(k-대표개체법)

	company	cluster
400	400	1
401	401	1
402	402	1
403	403	1
404	404	1
439	439	1
440	440	1
445	445	1
446	446	1
447	447	1
448	448	1
449	449	1
450	450	1
451	451	1
454	454	1
455	455	1
456	456	1
460	460	1
461	461	1
465	465	1

	company	cluster
405	405	2
406	406	2
407	407	2
409	409	2
410	410	2
411	411	2
412	412	2
413	413	2
414	414	2
418	418	2
419	419	2
420	420	2
429	429	2
430	430	2
431	431	2
432	432	2
441	441	2
442	442	2
443	443	2
444	444	2
453	453	2
457	457	2

	company	cluster
415	415	3
416	416	3
428	428	3

	company	cluster
458	458	4
459	459	4
462	462	4

	그룹1	그룹2	그룹3	그룹4
k평균법	400-404외 기타 나머지	405-414, 418, 419, 420, 429-432, 441-444, 453, 457	415,416,428	458, 459, 462
특징	기온 변수가 높으며, Tr-spp 변수가 작으며, Depth변수가 높다	기온 변수가 낮다	Tr-spp 변수가 크다	Depth 변수가 낮다

비계층 분석의 경우 총 그룹을 4가지로 나뉘어 분석을 진행하였다. 이전 분석과는 달리 종의 다양성 변수의 영향 뿐만 아니라, 환경적 변수역시 영향을 주고 있다. 이로써, 각 구역의 환경적 변수가 종의 다양성과 연관되어 있음을 확인해볼 수 있다. 또한, 환경적 변수의 경우 기온, Depth, Latitude가 영향을 주고 있으며, 종 다양성 변수의 경우, 기존의 , Ga-mo 변수와 Se-me 변수 뿐만 아니라, Tr-spp 변수역시 영향을 주고 있음을 알 수 있다.

결론

1.자료 특성상 특정 구역을 세분화하였기 때문에 환경적 변수의 특성이 두드러지지 않는다.

특정 지역내에서의 세분화된 구역에서 환경적 변수(위도, 경도, 깊이, 기온) 모두 큰 차이를 보이지 않아 사실상 유의미한 환경적 변수에서의 차이를 포착해내기 어려웠으며, 주성분분석에 있어 눈에 띄는 변수의 환경적 변수들의 역할을 포착해내기 어려웠다. 이는 바렌츠 해내의 지역별 환경적변수의 차이가 크지 않음을 의미한다.

2.자료 특성상 특정 생물의 분포가 두드러지며, 이로 인한 영향력이 두드러진다

주성분 분석을 진행함에 있어 주성분분석에 있어 눈에 띄는 변수 2개를 제외하고는, 높은 영향력을 띄는 조의 다양성 변수를 포착해내기 어려웠다. 주성부분석에서 드러난 변수는 Se-me 변수와 Tr-spp 변수이다. 또한 추가적인 군집분석을 통해 Tr-spp 변수가 다음으로 높은 영향력을 띄는 것으로 확인된다. 이처럼 30여개의 다양한 종의 다양성 변수 중에서 오직 3개만이, 소수의 개체종을 중심으로 군집을 형성하고 주성분을 결정하는데 대부분의 역할을 해내고 있다.

3.환경적 변수의 영향력은 미비하나, 영향력은 존재한다.

비록 주성분 분석에 있어, 환경적 변수의 영향력을 포착해내기에는 어려웠다. 그럼에도 불구하고, 군집분석에 있어 기온, 위도, 깊이 등이 군집을 형성하는데 중요한 역할을 하고 있다. 이로써 조금이나마 환경적 변수들이 종의 다양성 변수와 상관관계가 존재함을 의미한다.

참고문헌

자료데이터 출처 fbbva「https://www.fbbva.es/microsite/multivariate-statistics/data.html」