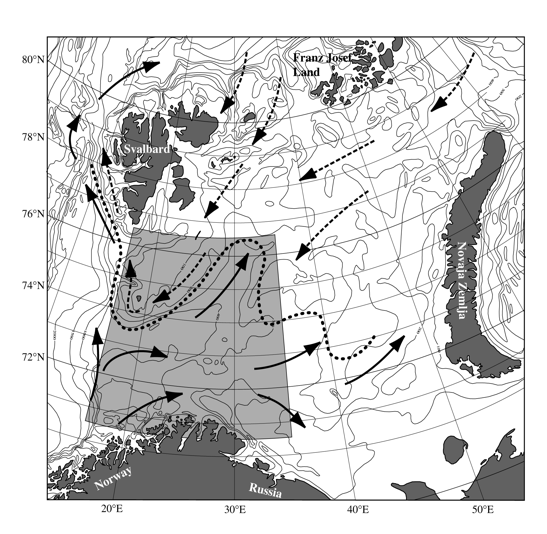
**다변량 통계학 TERM PROJECT**

**바렌츠해의 생물지리정보**

**201611539 하성진**

**서  론**

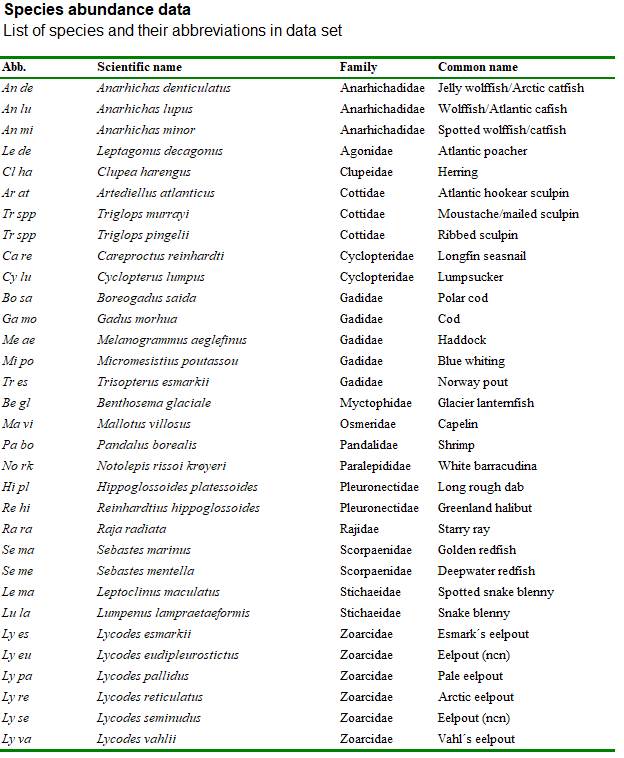
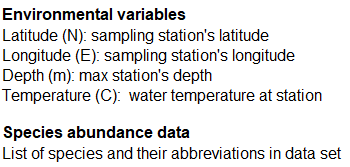
**1. 분석 내용**



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 여러개의 구역으로 나누어 각 구역의 환경적 특징과 어종의 종류와 개체수를 조사해보고, 조사된 데이터를 바탕으로 다변량 분석을 통해 구역별 특징 및 변수의 특징들을 알아보고자 한다.

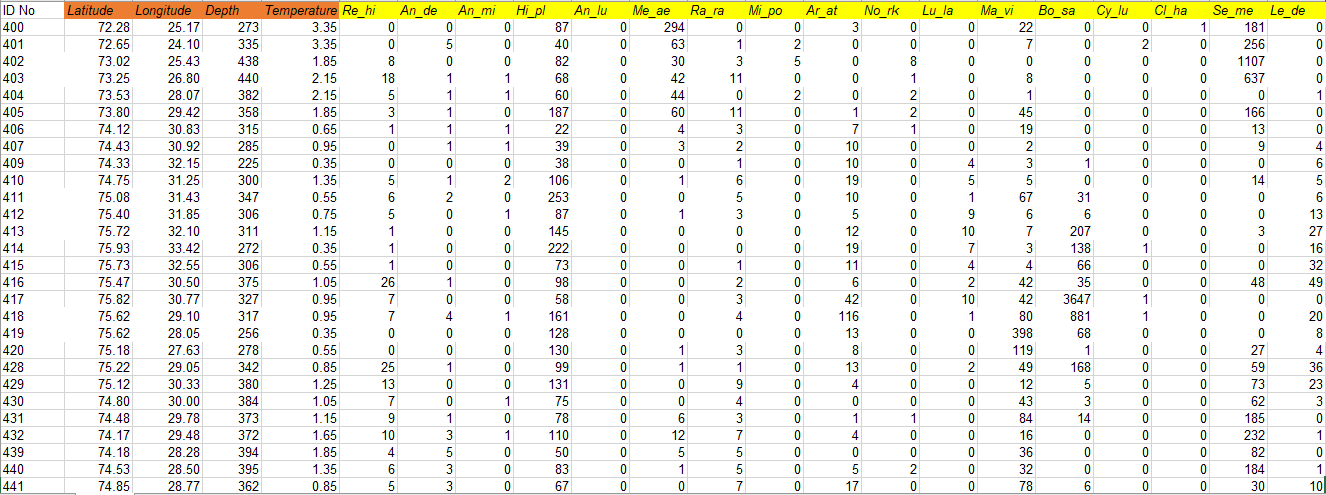
**2. 분석 데이터**

**(1)변수**



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 사용할예정이다)

**(2)개체**

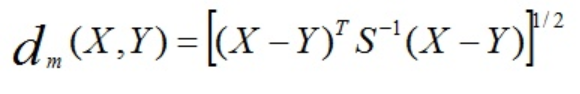


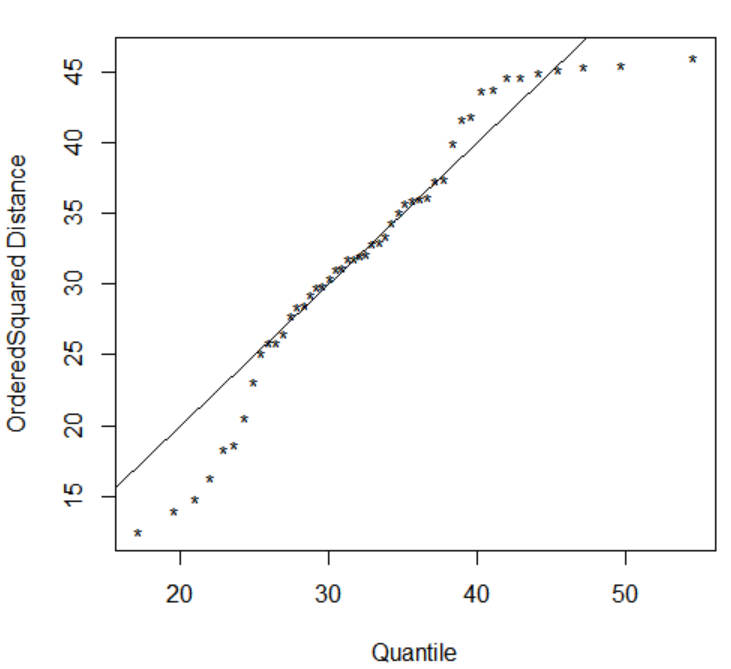
(공간의 한계로 441번 구역까지의 개체와 Le\_de(Leptagonus decagonus)까지의 종다양성 변수 까지만 표현하였음, 나머지 데이터는 원데이터 참고바람)

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.



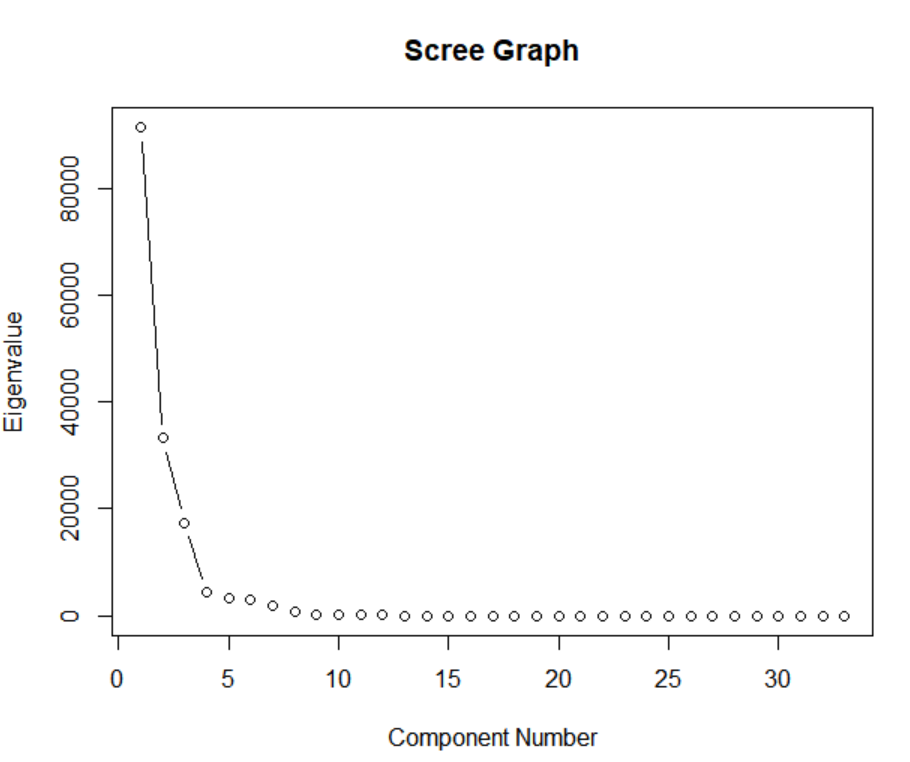




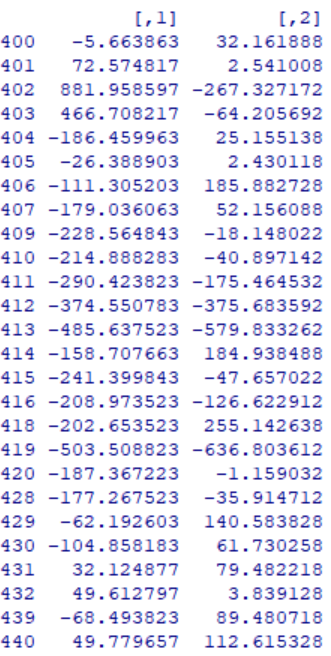
마할라노비스 거리를 이용한 피어슨 상관계수의 경우 거의 1에 가까운 값(0.9633116)으로 카이제곱그림의 직진성이 인정되며, 자료는 다변량 정규성을 만족한다고 할 수 있다.

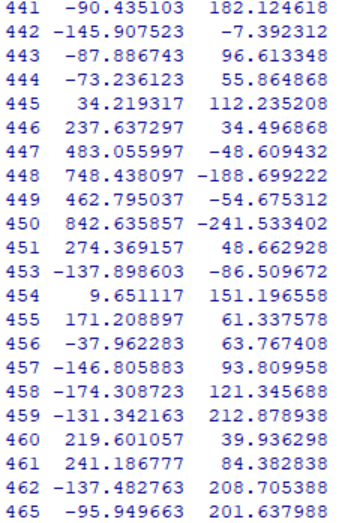
**1. 주성분 분석(PCA)**

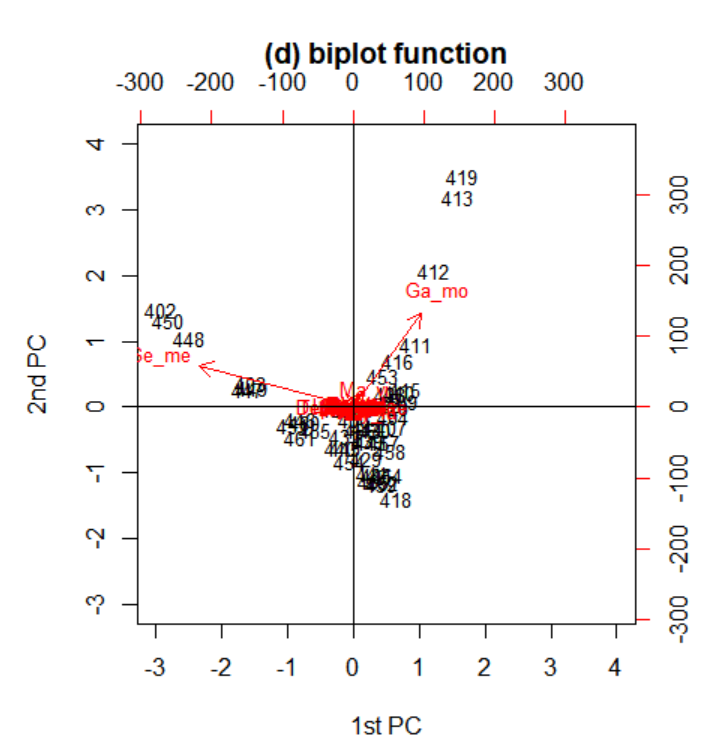
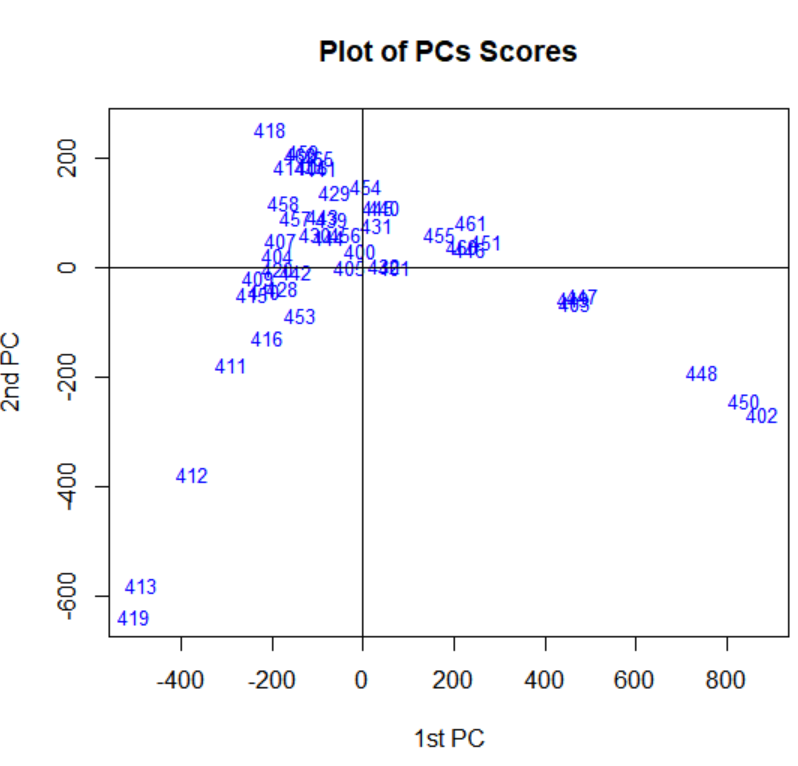
현재 34개의 변수가 존재하는 만큼, 저차원공간에서의 표현은 한계가 있다. 차원축소를 위해 주성분분석(PCA)를 진행해보고자 한다.



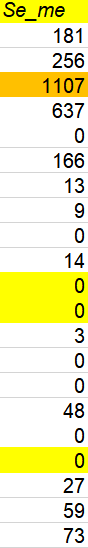
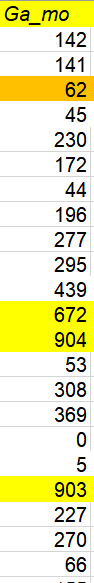
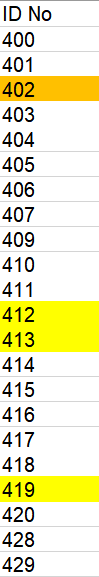
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 제1주성분 | 제2주성분 | 제3주성분 | 제4주성분 | 제5주성분 | ……. |
| Gof | 58.69 | 21.32 | 11.16 | 2.82 | 2.07 | …. |

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





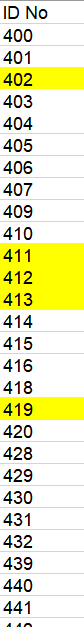
제1 주성분에서 413,419,412 구역이 가장 높은 음의 상관관계를 보이며, 448,450,402구역이 가장 높은 양의 상관관계를 보인다. 제2 주성분에선, 419, 413, 412, 402 등이 가장 높은 음의 상관관계를 보이고 있다. 또한, Ga-mo 변수와 Se-me 변수는 제1,2주성분과 높은 상관관계를 보이고 있다.



(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.

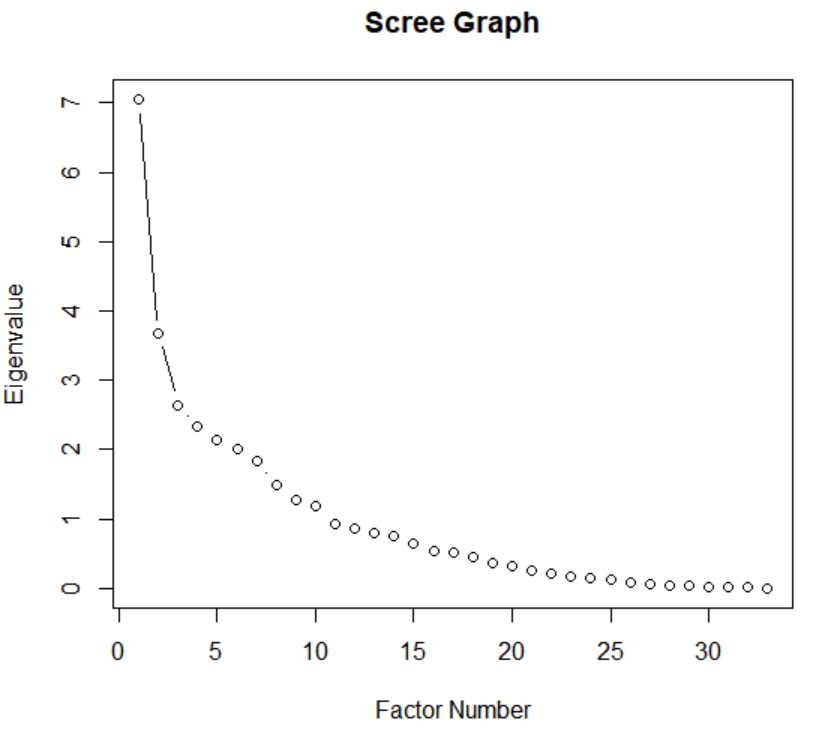
 

(Total 변수는 개체별 기존의 종의 다양성 변수의 합을 표현한 것, 기존 변수에는 존재하지 않음)

(노란색 표시를 기준으로 총합이 높은 개쳬는 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**



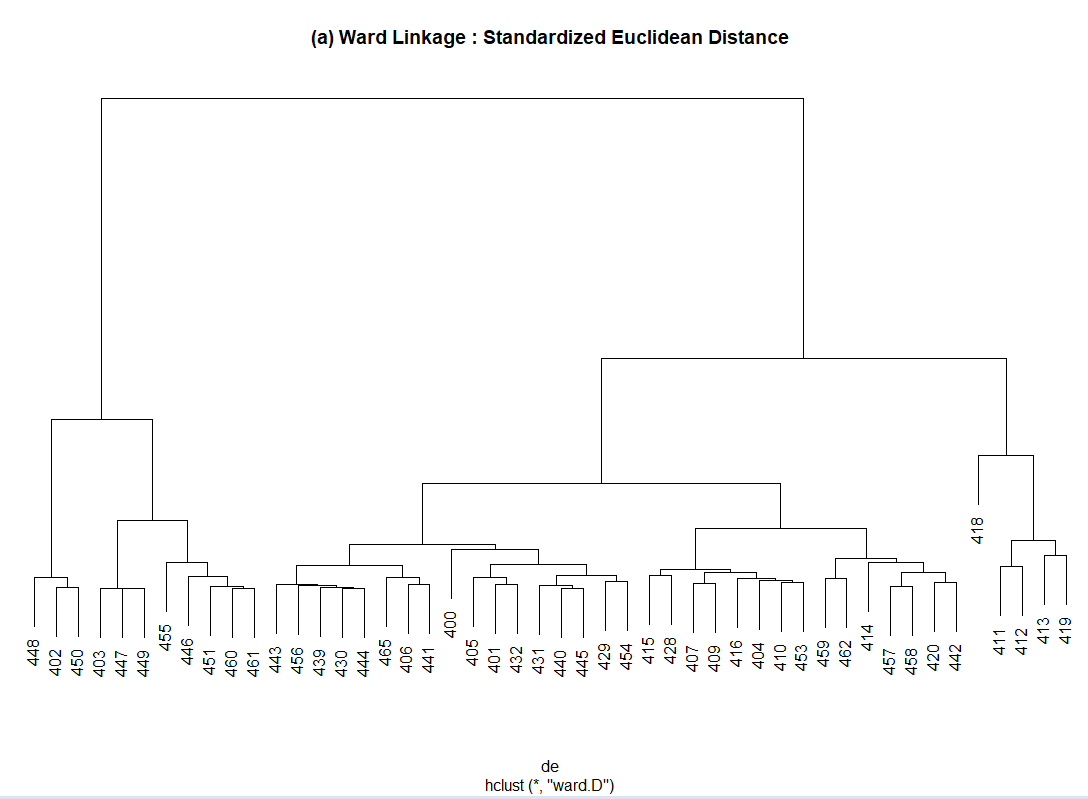
|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | 제1주성분 | 제2주성분 | 제3주성분 | 제4주성분 | 제5주성분 | 제6주성분 |
| Gof | 21.348 | 11.131 | 8.011 | 7.067 | 6.472 | 6.082 |

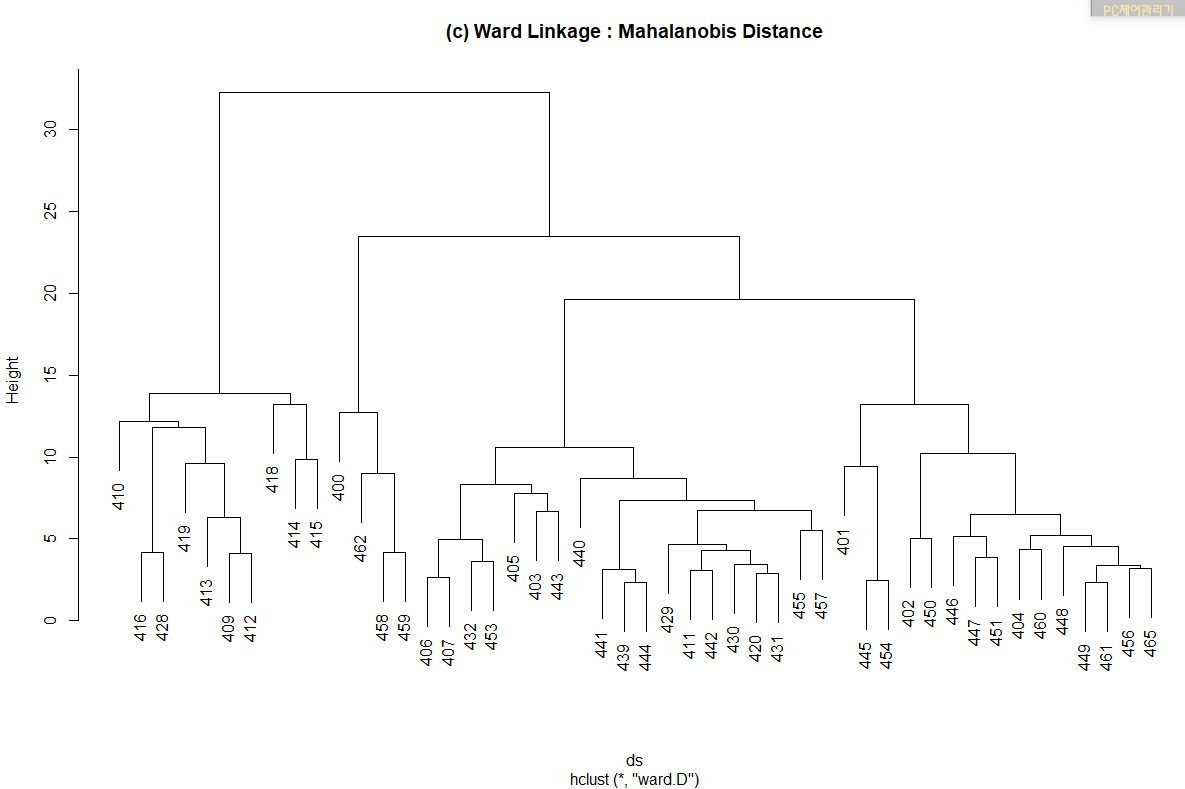
인자분석에 앞서 PCFA를 수행하여 각 주성분들의 설명력을 측정해본 결과, 설명력이 70%이상을 맞추기 위해선 6개이상의 공통인자를 이용해야만이 가능하다. 이는 결국 기본적인 PCFA만으로도, 총 7개이상의 인자에 대한 인자적재 값이 필요하며, 21개의 인자적재 그림이 필요하게된다. 따라서 인자분석에 있어 이번 자료는 비효율적이며, 적합하지 않은 자료랗 할 수 있다. 아쉽게도 이번 분석에서는 인자분석을 활용하지 못할 것으로 보인다.

**(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





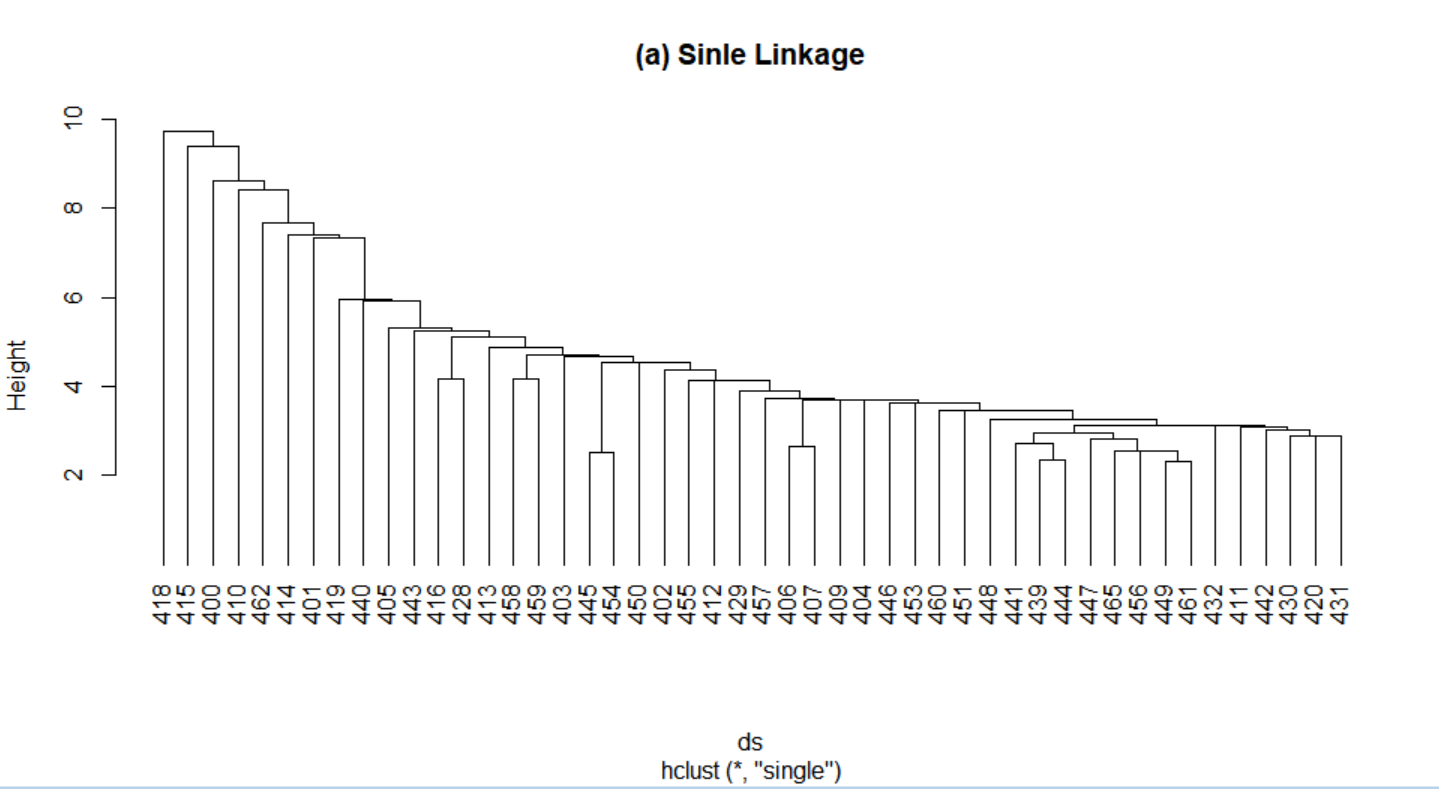
|  |  |  |  |
| --- | --- | --- | --- |
|  | 그룹1 | 그룹2 | 그룹3 |
| 유클리드 | 402,403,446,447,448,449,  450,451,455,460,461 | 411,412,413,418,419 | 그외 나머지 |
| 마할라노비스 | 401,402,404,416,418,  447,449,450,454,460, | 409.410,413,414,415,  416,419,428 | 그외 나머지 |
| 특징 | Se-me계수가 크다 | Ga-me 계수가 크다 | Se-me 계수와 Ga-me계수가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

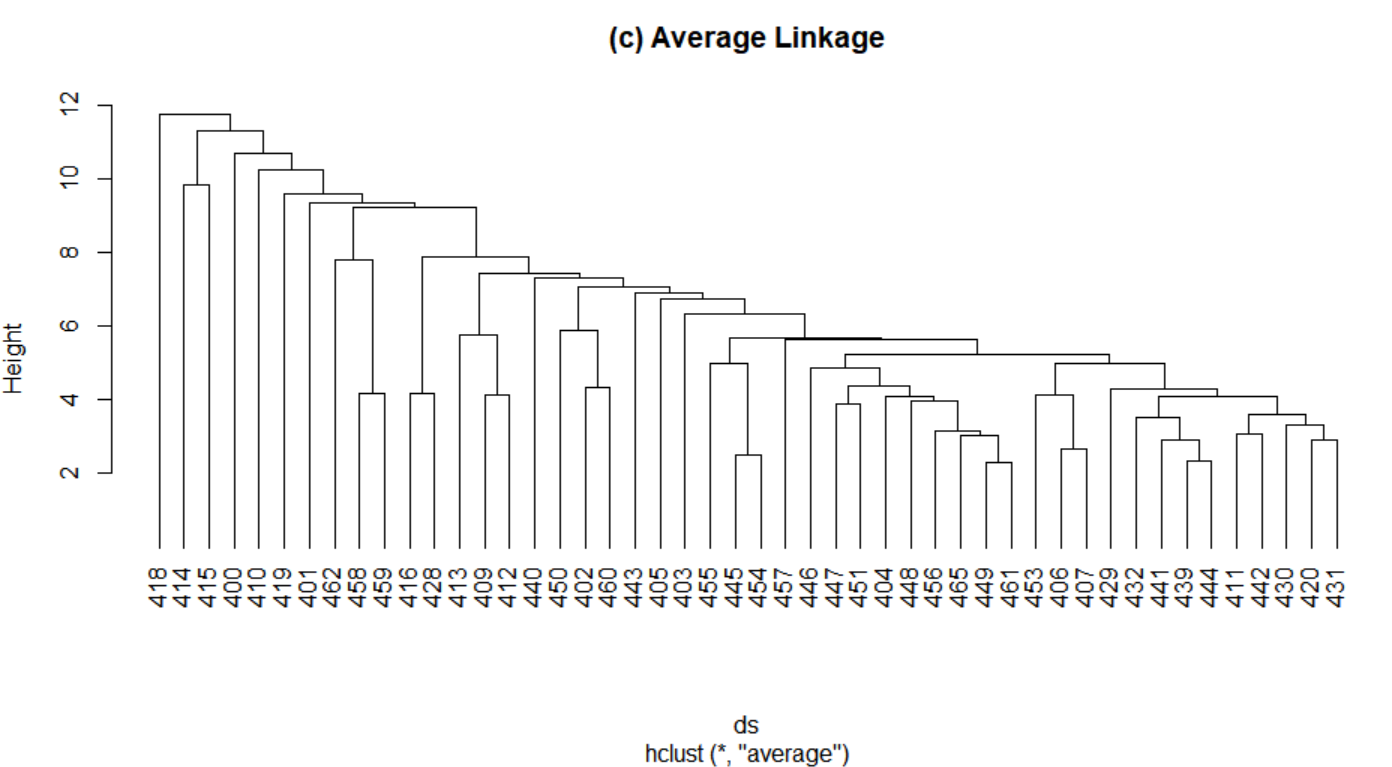
**b. 계층군집 분석**

와드연결법 분석을 바탕으로 계층 군집분석도 함께 실행해보려 한다.

(단일연결법)



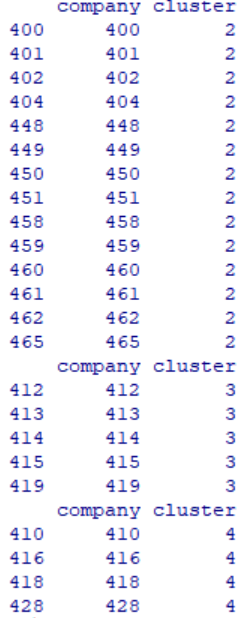
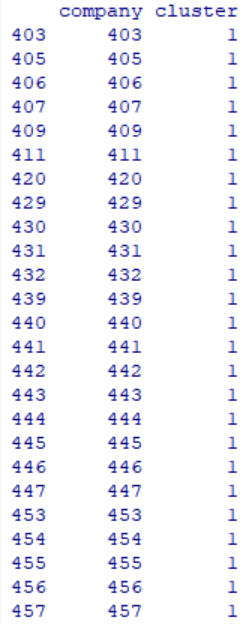
(평균 연결법)



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.

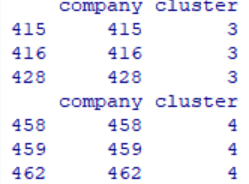
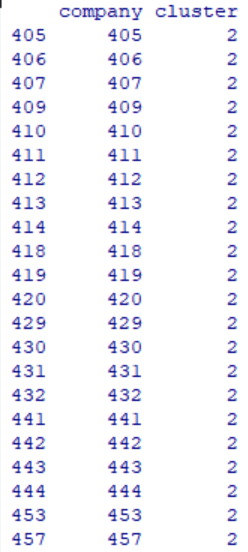
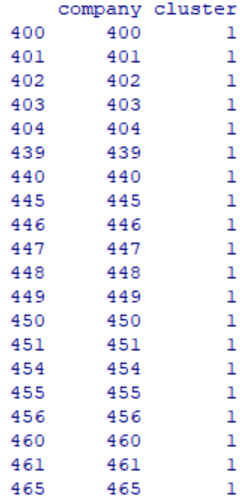
c. 비계층 군집분석

(k-평균법)



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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)



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | 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」