Baayen, R. H. (2008). *Analyzing Linguistic Data: A practical introduction to statistics using R.* Cambridge: Cambridge University Press.

金田　拓

【　続　クラスター分析　】　pp. 142~148

・　パプアとオセアニアの言語分類を試みる。phylogenyというデータ[1]を用いる
・　各言語は非常に異なっているため、単語を基にした分類は不可能。
　　⇒　文法的特徴の有無で系統づけることが可能なのではないか。

```
> phylogeny[1:5, 1:5]
```

・　分割クラスタリング(divisive clustering)を行う。

```
> phylogeny.dist = dist(phylogeny[ ,3:ncol(phylogeny)], method = "binary")
```

・　認識しやすいように、パプア系の言語は大文字にする

```
> plotnames = as.character(phylogeny$Language)
> plotnames[phylogeny$Family == "Papuan"] =
+   toupper(plotnames[phylogeny$Family == "Papuan"])
```

・　分割クラスタリングと視覚化を行う

```
> library(cluster)
> plot(diana(dist(phylogeny[ , 3:ncol(phylogeny)], method = "binary")), labels = plotnames,
+   cex = 0.8, main = " ", xlab = "aa", which.plot = 2)
```

・　パプア系言語とオセアニア系言語の 2 つに、クラスターがくっきりと分かれているのが見て取れる

・　ape パッケージの関数 nj ( )を使って、近隣結合法で系統樹形図を出力する。

```
> library(ape)
> phylogeny.dist.tr = nj(phylogeny.dist)
```

☆　視覚化する際に、各言語をはっきり分かるように区別したい　⇒　行番号ではなく名前を代わりに代入する。
・　行番号を families に数列として代入する

```
> families = as.character (phylogeny$Family)
[as.numeric(phylogeny.dist.tr$tip.label)])
```

[1] Dunn, M., Terrill, G., Reesink, R., Foley, A., & Levinson, S. C. (2005). Structural Phylogenetics and the Reconstruction of Ancient Language History. *Science, 309*, 2072-2075.

・　言語の名前を languages に代入

```
> languages = as.character (phylogeny$Language[as.numeric(phylogeny.dist.tr$tip.label)])
```

・　行の代わりに名前を使って置き換える

```
> phylogeny.dist.tr$tip.label = languages
```

・　プロットする

```
> plot(phylogeny.dist.tr, type = "u", font = as.numeric(as.factor(families)))
```

- type = "u"を指定すると、根のついていない樹形図になる
- 5.12 が基本的には同じ分類に基づいているのは明らか

・　単語を基に分類しようとしてもうまくいかないが、文法項目基準だとはっきりとカテゴリー分けが可能

```
> papuan = phylogeny[phylogeny$Family == "Papuan", ]
> papuan$Language = as.factor(as.character(papuan$Language))
> papuan.meta = papuan[ , 1:2]
> papuan.mat = papuan[ , 3:ncol(papuan)]
> papuan.meta$Geography = c("Bougainville", "Bismarck Archipelago", "Bougainville",
+   "Bismarck Archipelago", "Bismarck Archipelago", "Central Solomons",
+   "Bougainville", "Louisiade Archipelago", "Bougainville", "Bismarck Archipelago",
+   "Bismarck Archipelago", "Bismarck Archipelago", "Central Solomons",
+   "Central Solomons", "Central Solomons")
> papuan.dist = dist(papuan.mat, method = "binary")
> papuan.dist.tr = nj(papuan.dist)
> fonts = as.character(papuan.meta$Geography[as.numeric(papuan.dist.tr$tip.label)])
> papuan.dist.tr$tip.label =
+   as.character (papuan.meta$Language[as.numeric(papuan.dist.tr$tip.label)])
> plot(papuan.dist.tr, type = "u", font = as.numeric(as.factor(fonts)))
```

- どのようなデータにどのような手法を用いるのが確実・速いなどという公式は無い
- いろいろ試した結果、一つだけ載せることもよくある。クラスターを見せるのにベストな方法を選んでいるだけ。

【ブートストラップ法】

- クラスター分析の妥当性を検証するにはブートストラップ法を用いる。
- 行列からリサンプリング（重複含む）を繰り返して、得られたデータについて距離の行列を求め、対応根無し樹形図を node-joining 法で描く。
- もともとの系統樹と、ブートストラップ法でできた系統樹とを比較する。

- 温度計(thermometer)でツリーの違いを表現。温度が高ければ高いほど、サブツリーがサポートされている。
- 以下のブートストラップ分析はParadis[2](2006:117)のデータ・手法を踏襲。まず何回リサンプリングするかを決める

```
> B = 200
> btr = list()
> length(btr) = B
```

- 次にリサンプリングを行い、200 のブートストラップツリーを作る。

```
> for (i in 1:B){
+   trB = nj(dist(papuan.mat[ ,sample(ncol(papuan.mat), replace = TRUE)], method =
+   "binary")) trB$tip.label = as.character(papuan.meta$Language[as.numeric(trB$tip.label)])
+   btr[[i]] = trB }
```

- 実際に元のツリーと合っているかどうかは prop.clades()を用いる

```
> props = prop.clades(papuan.dist.tr, btr)/B
> props
```

- 元のツリーをプロット

```
> plot(papuan.dist.tr, type = "u", font = as.numeric(as.factor(fonts)))
```

- 温度計を付け足す

```
> nodelabels(thermo = props, piecol = c("black", "grey"))
```

*** black & grey では見づらい。こっちの方が温度計らしくて見やすい ***

```
> nodelabels(thermo = props, piecol = c("red", "white"))
```

➢ 中心に行くに従って温度が下がっている　＝　合意(consensus)が無い

[2] Paradis, E. (2006). *Analysis of Phylogenetics and Evolution with R*. Springer: New York.

- 合意樹(consensus tree)を描くという方法もある。その場合、ブートストラップツリーに含まれないものが仲間外れということになる。
- Figure 5.13 でいうと、真中に集まっている 9 つが多項的分類法(?) ape を使って求められる

```
> btr.consensus = consensus(btr, p = 0.5)
```

```
> x = btr.consensus$tip.label
> x
```

```
> x = data.frame(Language = x, Node = 1:length(x))
> x = merge(x, papuan.meta, by.x = "Language", by.y = "Language")
> head(x)
```

```
> x = x[order(x$Node), ]
> x$Geography = as.factor(x$Geography)
> plot(btr.consensus, type = "u", font = as.numeric(x$Geography))
```

- 結果、ツリー間で一定でないグループがある
- だが、125 ある文法項目のうち 80 しか用いていないので、より多くの言語、より多くの文法項目を用いれば、元々の樹形図の結果が支持されるかもしれないことは念頭に置く必要がある。