ID	KK300604
種別	解説
タイトル	データの視覚化（6） —Rによる樹形図の作成—
Title	Data Visualization (6): Making Dendrogram in R statics
著者	林　直樹
Author	HAYASHI Naoki
掲載号	30巻6号
発行日	2016年9月20日
開始ページ	378
終了ページ	390
著作権者	計量国語学会

計量国語学 30 巻 6 号（2016 年 9 月）pp.378-390.

解説

データの視覚化（6）
―Rによる樹形図の作成―

林　直樹（日本大学文理学部）

要旨

本稿では，統計ソフトRを使用したグラフの作成方法を述べる．グラフの作成にあたっては，初期的な出力をただ貼り付けるのではなく，種々の変更を加えるコマンドと，出力結果を交互に挙げていく．説明のために取り上げるグラフは，樹形図（デンドログラム）である．解説では，初期的な出力から諸情報を変更するための方法を述べた．また，出力されたグラフをWordに貼り付ける際の方法も解説した．最後に，Rでグラフを作成する際の注意点についてもふれた．

キーワード：R，クラスター分析，樹形図

1．はじめに

本稿では，統計ソフトRを使用してグラフを作成するため，コマンドと出力結果を示しながら種々の改良方法について説明していく．取り上げるグラフは，Rであれば簡単に作ることができる樹形図（デンドログラム）とする．Excelでは基本的に作成することができず，統計ソフトウェアのSPSSでは作図できるものの独自の変更は加えられないため，対象とした．

なお，本稿はRの出力をMicrosoft Wordといった文書作成ソフトに貼り付ける場合を想定して説明をする．著者の2016年現在の環境を述べると，OSはWindows7，使用ソフトのヴァージョンはWord2010（Microsoft Office2010）となっている．Rのヴァージョンは3.2.2で，Emacs上のESSから操作しているが，通常のRコンソールやRコマンダーを使用しても同じような動作が実現するはずである．

2．Rの概要

Rはオープンソースの統計解析システムである．さまざまな計算や統計解析を行うことができ，その上無料で入手できる点が最も大きな特徴といえる．Rの詳しい説明や，インストールの方法などは公式サイト（https://www.r-project.org/）や日本語Wiki（http://www.okadajp.org/RWiki/）で記述されている．概説書としては金明哲（2007）や船尾暢男（2008），山田剛史・杉澤武俊・村井潤一郎（2008）があり，日本語研究に関連させて解説を行っているものには，鑓水兼貴（2012）がある．

Rの操作上の特色としては，基本的にコマンドによって操作を行う点が挙げられる．そ

のため，Excel や SPSS といったクリックによって操作するタイプのソフトウェアに慣れたユーザーは導入をためらってしまうかもしれないが，近年は Web 上の解説が大変充実しており，困ったときは検索をすれば大抵のことは解決できる．また，RStudio といった，ユーザーが馴染みやすいソフトウェアも登場しているため，R をこれから導入する場合はさまざまな動作環境を試し，自身に合った形で使用していくことを推奨する．

上記の日本語 Wiki では，「R の強みの一つは，巧くデザインされた出版物並みのプロットを容易に作成できる点」と述べられており，実際にそのとおりではあるが，作成した図表を綺麗に出力するためにはそれなりに気をつけるべき点もあるため，本稿ではそれらの点を解説していく．

なお，R のライブラリーに，強力なグラフ作成機能が搭載されている ggplot2 があるが，通常の作図方法とは異なる面もあるため，本稿では説明を省く．

3．クラスター分析

本稿で解説のために用いる多変量解析手法は，クラスター分析である．クラスター分析は「分類対象の集合の各要素を個体間の類似度に基づき，いわゆる「似たもの同士」の部分集合に分類する」（佐藤義治，2009：87）手法とされる．Romesburg（1992）では，生物学・医学・海洋学といった諸分野での適用例が報告されている．

クラスター分析結果を示すためには，「グループの形成状態を樹形図で示す階層的クラスター分析の方法」と，「どの個体がどのグループに属するかを示す非階層的クラスター分析方法」がある（金明哲，2007）．本稿では，「最も頻繁に用いられる階層的クラスター分析」（Romesburg，1992）の結果を樹形図で示すことに焦点を絞り解説していく．樹形図を変更させることに特化した説明を行うが，ここで使用する関数はその他のグラフ作成時にも共通して使えるものもある．

4．クラスター分析のためのデータ

本稿では，田中ゆかり（2013）で使用されている「全国方言意識調査」の首都圏・関西圏データ（以下，方言意識調査データ）を用いて，首都圏・関西圏に属する都道府県が，方言意識からどのようにクラスタリングされるかを例に説明していく．

上記のデータは多数の項目があるが，本稿では結果をわかりやすく出力するために，質問項目「出身地方言好悪」「共通語好悪」「対家族への方言使用」「対地元友人への方言使用」「対非地元友人への方言使用」のみを用いる．また，これらの項目のうち，「出身地方言好悪」「共通語好悪」は「好き」と回答した比率，「対家族への方言使用」「対地元友人への方言使用」「対非地元友人への方言使用」は「よく使う」と回答した比率を都道府県別にまとめたものをデータとした．

以上のデータを csv 形式で「test」という名前で準備した場合，R に読み込ませるには，以下のようにコンソール上で入力する．

```
1  test <- read.csv("test.csv")
```

この結果，表 1 のようなデータが R に格納される．

表 1：方言意識調査データの読み込み結果

No	都道府県	サンプル数	出身地方言好き	共通語好き	対家族方言よく使う	対地元友人方言よく使う	対非地元友人方言よく使う
11	埼玉県	37	0.216	0.541	0.027	0.000	0.000
12	千葉県	57	0.456	0.754	0.088	0.105	0.123
13	東京都	102	0.500	0.853	0.324	0.314	0.275
14	神奈川県	77	0.506	0.662	0.182	0.182	0.169
25	滋賀県	18	0.556	0.278	0.778	0.778	0.611
26	京都府	27	0.778	0.370	0.630	0.556	0.407
27	大阪府	82	0.744	0.378	0.695	0.720	0.598
28	兵庫県	45	0.578	0.578	0.689	0.689	0.489
29	奈良県	13	0.615	0.154	0.615	0.615	0.538
30	和歌山県	13	0.615	0.231	0.769	0.846	0.462

このデータのうち，左から 1 番目・2 番目の列は都道府県を表すアイテムデータ，左から 3 番目の列はサンプル数を示す数値データとなっているため，左から 4 番目以降の要素を対象として，クラスター分析を試みていく．

5．クラスター分析実例・作図

まず，クラスター分析を行うため，各行列間の距離を算出していく．距離を算出する方法には多種あるが，ここでは，平方ユークリッド距離を用いる．

```
2  test_dist <- (dist(test[,4:8], method = "euclidean")^2)
```

dist 関数では，ユークリッド距離（“euclidean”）の他に，最大距離（“maximum”），マンハッタン距離（“manhattan”），キャンベラ距離（“canberra”），二項距離（“binary”），ミンコフスキー距離（“minkowski”）が計算できる．
次に，Ward 法を用いてまとめ上げる[1]．

```
3  test_clst <- hclust(test_dist, method= "ward")
```

以上が，樹形図を描くための基本的な準備である．あとは plot 関数で描画をすれば，樹形図が R Graphics 上に表示される．

```
4  plot(test_clst)
```

結果を示したのが以下である．

1　R のヴァージョンが 3.0.3 以降の場合，Ward 法のアルゴリズムを 2 種類選択できるようになった．ここでは，以前から格納されている“ward.D”を使用した．詳細は Rhelp（https://stat.ethz.ch/R-manual/R-devel/library/stats/html/hclust.html）参照．

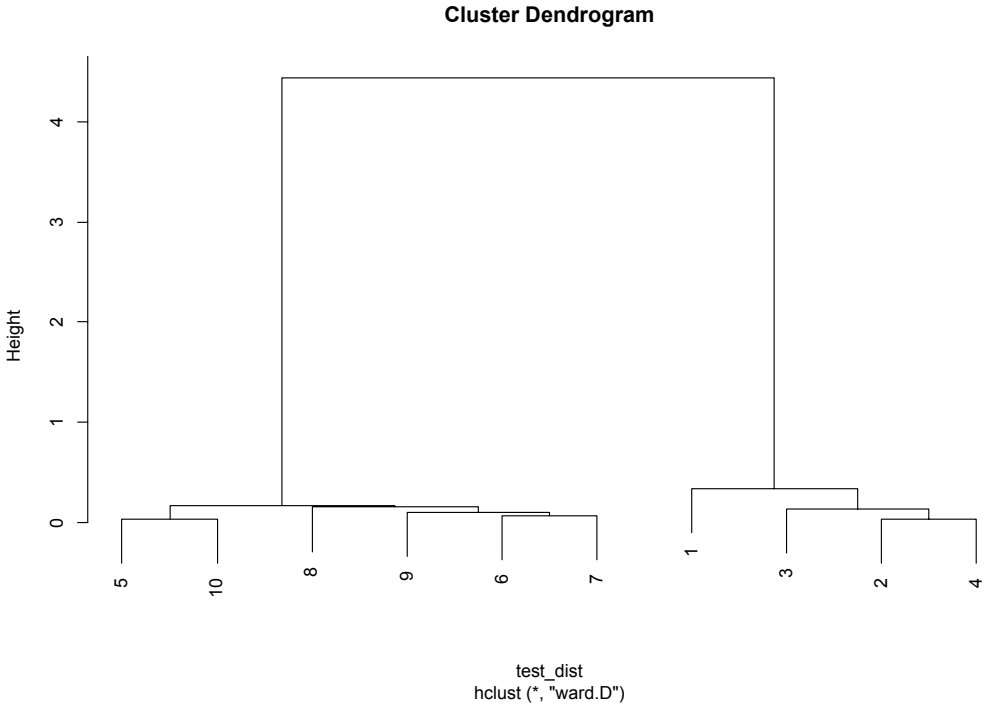


図 1：方言意識調査データによる首都圏・関西圏のクラスター分析結果

樹形図が出力されたものの，一見しただけで，

1）都道府県の名前が番号になっており，結果がわかりにくい
2）図上部の“Cluster Dendrogram”，下部の“test_dist”など，分析に直接関係のない情報が入り込んでいる
3）文字が小さく，余計な空白も多い

といった問題があることがわかる．以下では，これらの問題を解決してすっきりとわかりやすいグラフを作成するための方法を述べていく．

5.1　アイテムデータの反映

樹形図に都道府県名などのアイテムデータを反映させるためには，元データの行名にその情報を反映させる必要がある．例えば，表 1 では左から 2 列目に都道府県名が格納されているため，それを行名に反映させるためには，データ読み込み時に以下のように入力する．

```
5  test <- read.csv("test.csv", row.names = 2)
```

ただし，このコマンドでは元々の 2 列目の情報がそのまま行名に移行してしまい，データ

本体の行列から除かれてしまうため，後に都道府県名を使用して分析する場合は若干手間取ることになる．このような不便さを回避するため単純に行名を追加する場合には，以下のような入力を行う．

```
5′ rownames(test) <-c(" 埼玉県 "," 千葉県 "," 東京都 "," 神奈川県 ",
   " 滋賀県 "," 京都府 "," 大阪府 "," 兵庫県 "," 奈良県 "," 和歌山県 ")
```

コマンド 5 で入力したデータを用いて再度分析した結果，ならびに出力された樹形図は以下である．

```
6 test_dist <- (dist(test[,3:7], method = "euclidean")^2)
7 test_clst <- hclust(test_dist, method= "ward")
8 plot(test_clst)
```

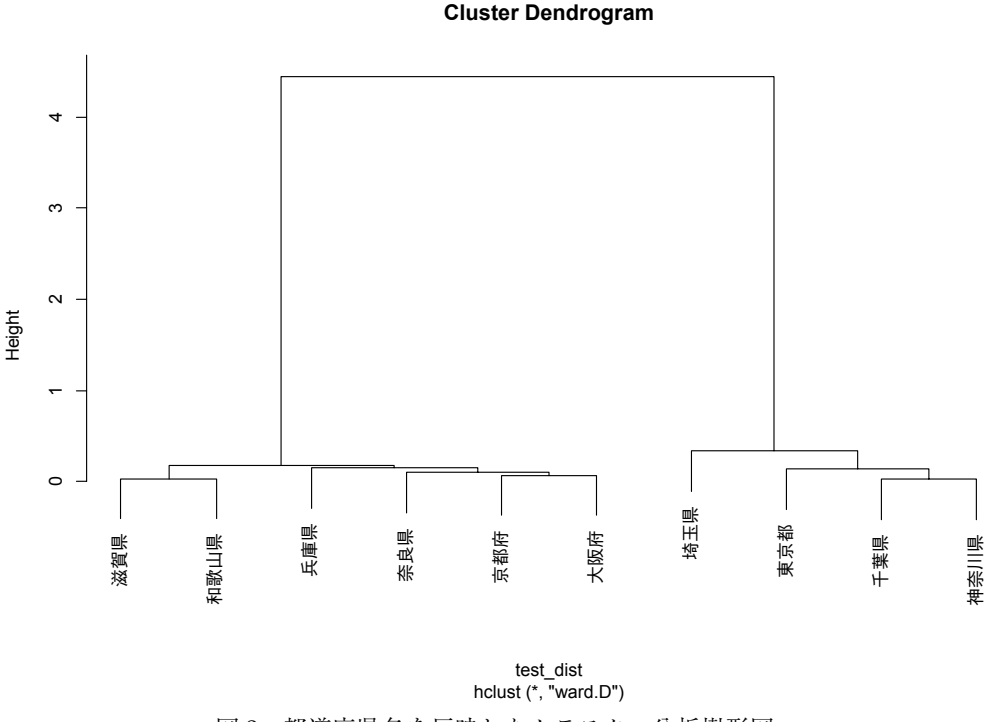


図 2：都道府県名を反映したクラスター分析樹形図

これで，首都圏と関西圏で大きなクラスターが構成されていることが一目瞭然となった．

5.2　タイトルの変更

次に，単純に樹形図を示す際には不要と思われる，ラベルの変更について述べていく．まず，樹形図で示されている各情報を以下に示す．

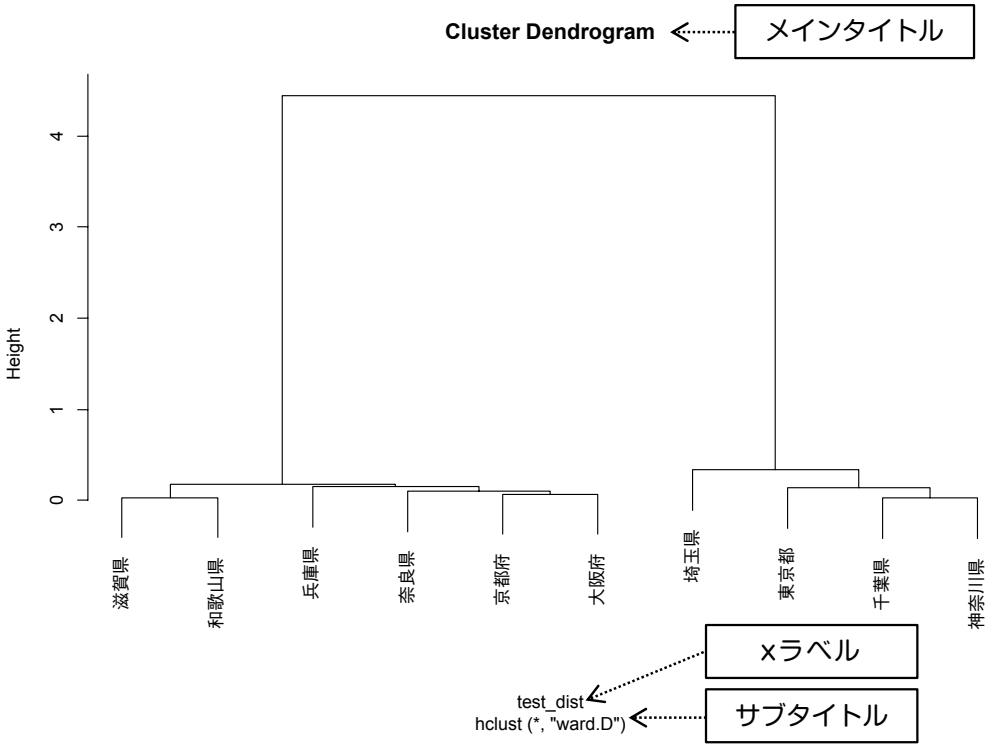


図 3：樹形図の各情報

これらのタイトル・ラベルを取り除くには，plot 関数[2]を使い，以下のように入力する．

```
 9  plot(test_clst,
10   main = "",                    ## メインタイトルの非表示
11   sub = "",                     ## サブタイトルの非表示
12   xlab = ""                     ## x ラベルの非表示
13  )
```

の右に示したとおり，メインタイトル，サブラベル，x ラベルそれぞれの情報を非表示にしている．「= ""」で何も入力していないことになるため，仮にタイトルに何らかの情報を入力したい場合は，入力したい情報を入れれば良い．結果を示したのが以下である．

2　plot 関数は，散布図などその他のグラフ作成時にも共通して使用する．そのため，例えば散布図の x 軸・y 軸のラベルを表示したい場合は，「xlab = ""」や「ylab = ""」に表示したい変数名を入力すれば良い．

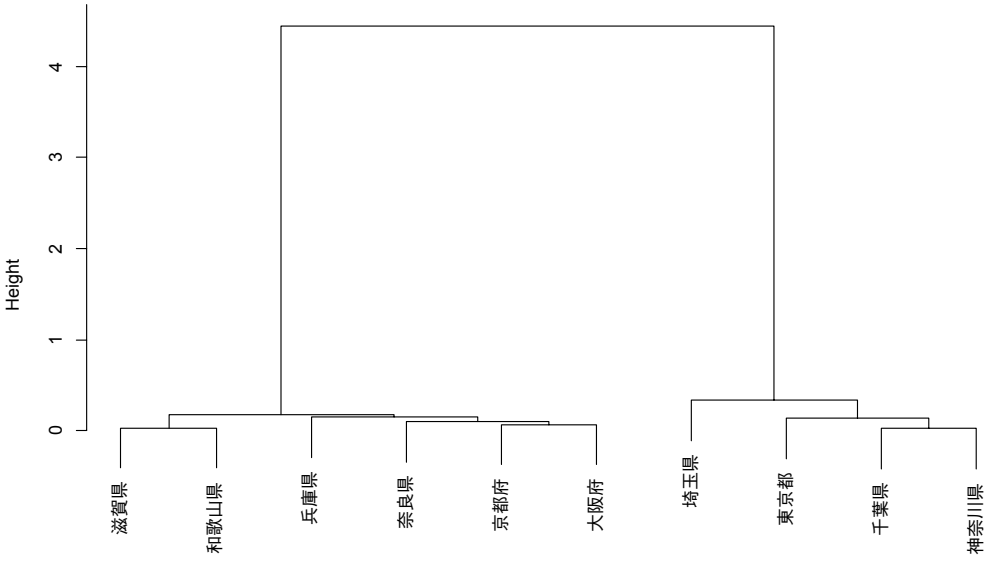


図 4：各情報を修正した樹形図

これで，分析に必要な情報のみが表示されるようになった．

5.3　文字の大きさや余白の調整

図 4 は必要情報のみが表示されるようになったが，文字が小さく，余計な空白が入っている．このような作図にかんする変更を加えるには，par 関数[3]を用いる．例えば，文字の大きさ・空白・線の太さなどを変更するには，以下のコマンドを入力する．

```
14  par(
15   ps=20,                          ## 文字の大きさの調整
16   mar=c(2, 6, 2, 0.5),            ## 空白の調整（下, 左, 上, 右）
17   par(lwd = 2),                   ## 線の太さの調整
18   par(lty = 1)                    ## 線の種類の調整
19  )
```

線の種類では実線を指定しているため，実際の出力結果には変動はない．このパラメータは，「2」でダッシュ，「3」でドットといったように，数字によって線分の種類を指定する．「lty = "dashed"」などとしても変更することもできる．

3　par 関数も，plot 関数と同様その他のグラフを作成する際に使用することができる．ただし，グラフによって調整するべき空白の大きさも変わるため，ここで挙げた設定がそのまま当てはまるわけではない．

結果を示したのが以下である．

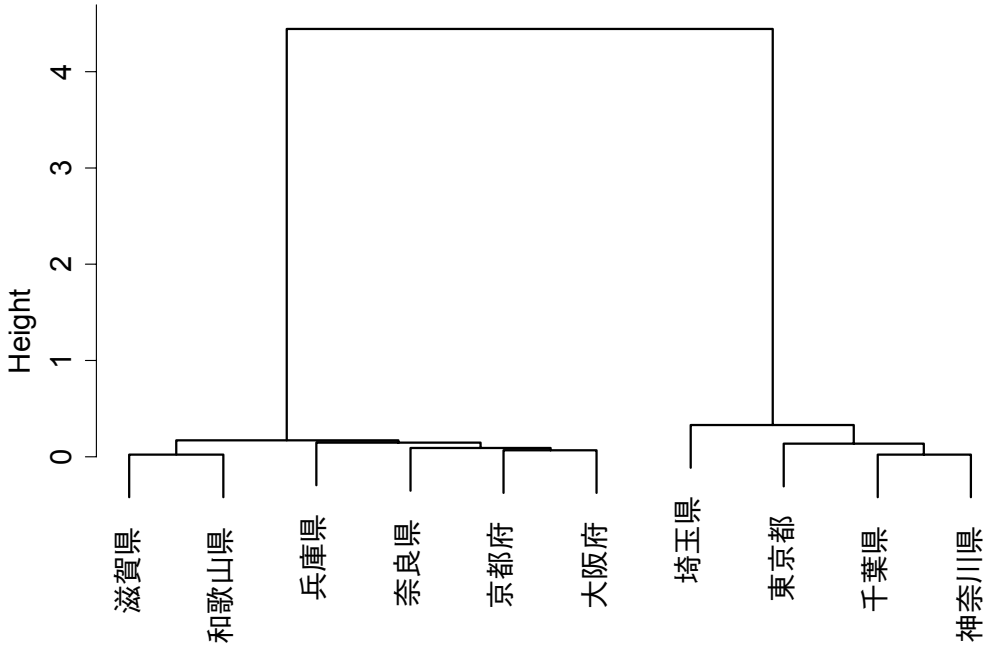


図 5：出力情報を調整した樹形図

これで，文字のポイントが上がり，かなりみやすくなったように思われる．par 関数はその他にも多数のパラメータがあるため，必要に応じてその他の変更を加えても良い．散布図によってプロットするマーカーの種類を変えるといった操作もできる．

さらに，これまで作成してきた樹形図では大きく首都圏・関西圏という 2 種類のクラスターに分かれることが直感的に判断できるため，次に，その情報を樹形図に反映する．ここで使用するのは，rect.hclust 関数である．入力する場合，以下のようにする．

```
20  rect.hclust(test_clst, k = 2, border = "black")
```

「k = 」でクラスター数を指定し，「border = 」で分割線の色を指定している．ここで上記 par 関数の「lty」などを用いれば，線の種類を変えることもできる．出力例は以下である．

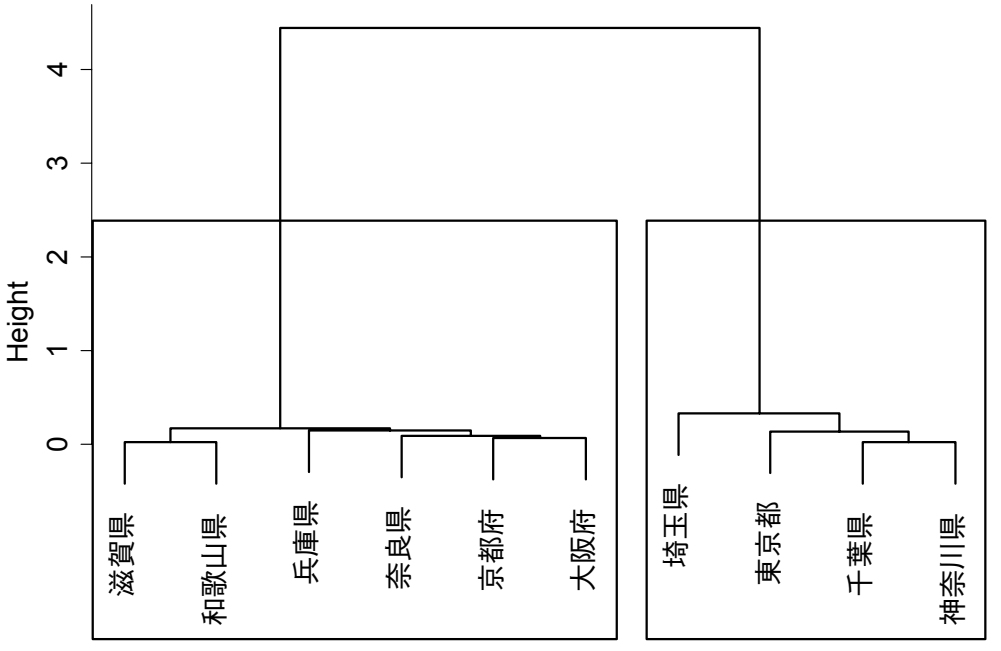


図 6：クラスターを 2 分割した樹形図

5.4　図の出力

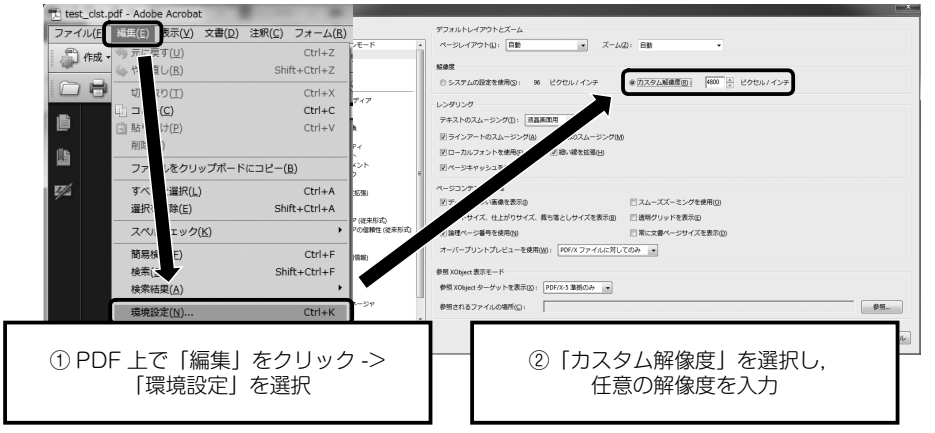
ここまで作成してきた図では，R Graphics 上でメタファイルとしてコピーしたものを貼り付けてきた．これらの図をファイルとして出力したい場合は，別途コマンドを入力する必要がある．PDF で出力する場合は，以下のように入力する．

```
21  dev.copy2pdf(file = "test_clst.pdf", family = "Japan1GothicBBB")
```

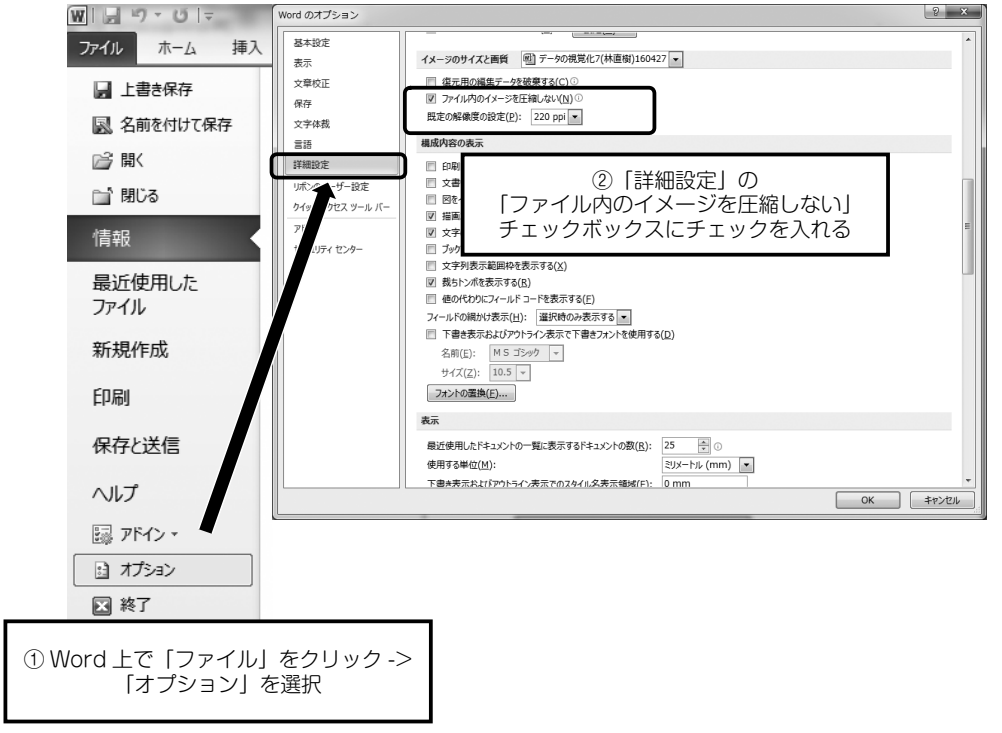
「file = 」がファイル名を指定するパラメータ，「family = 」がフォントを指定するパラメータである．ここではゴシックを指定した．これにより，作業ディレクトリに「test_clst.pdf」という PDF ファイルが出力される．

5.5　PDF の貼り付け

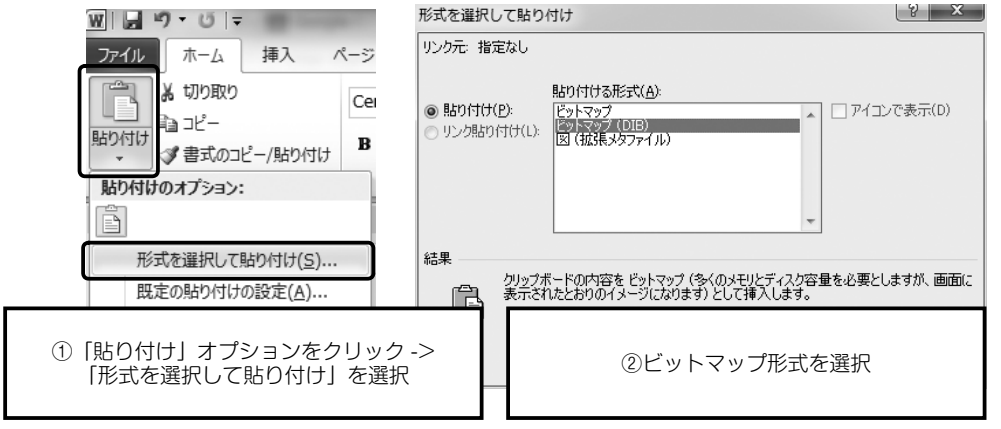
次に，PDF を Word に貼り付ける際のオプションを説明する．作成された PDF ファイルはスナップショットツールにより単純に貼り付けることもできるが，大抵は解像度が低く，ぼやけたような状態になってしまう．そこで，PDF の設定を以下のように変更する．



さらに，Word の設定も以下のように変更する．



この上で，クリップボード上に保存されている画像を以下のように貼り付けると，より鮮明になる．



これらの一連の操作を行い，PDF ファイルを貼り付けたのが以下である．

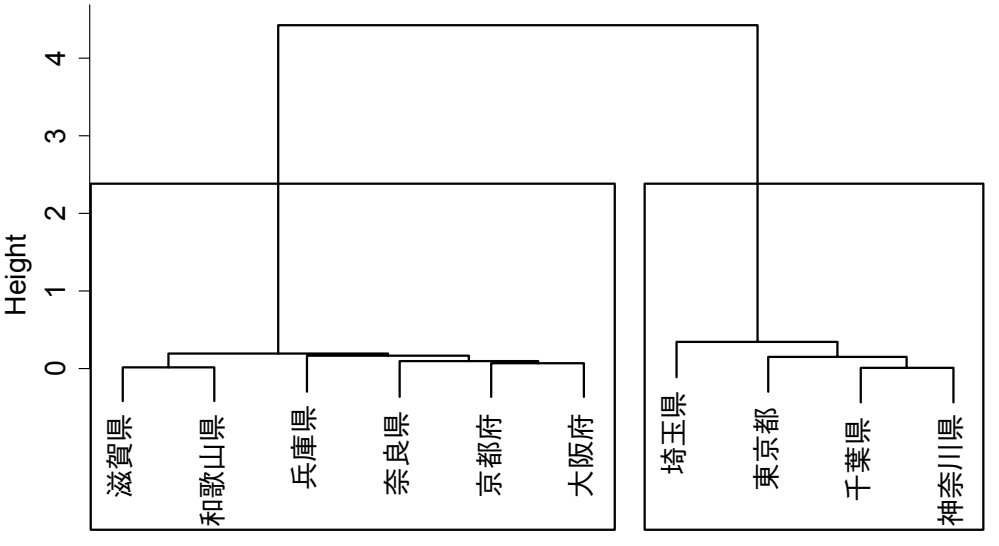


図 7：PDF にしたファイルを貼り付けた樹形図

図 6・7 ともに，初期的な状態よりも大分鮮明になったように思われるため，一応の完成とする．レイアウトを気にする場合は，樹形図全体を右に 90° 回転させるなどしても良い．

6．R でグラフを作成するにあたって

以上，R によるグラフ作成の出力から，修正方法まで記してきた．ここで述べたことは R のグラフ作成機能のごく一部分である．熟達すれば，さらに自在にグラフを描けるようになる．筆者は R と Excel とを使い分けているが，使い方によっては Excel よりも完成度

の高いグラフを作成することができるようにもなるだろう．ただし，本稿で述べてきたように，何も工夫をしないと書き手の意図が十分に伝わらないグラフができてしまう危険性もある．

R を使用する上で重要なのは，とにかく作成と修正を繰り返し，自身が作図を行う際に基盤となるような設定を見つけ出すことである．最初は苦労するかもしれないが，コマンドを保存しておけば別のデータを使用する際にもそのまま適応できるため，結果的に手間と時間が節約されることにもなる．

以上のように，グラフを自分好みにカスタマイズしていけることが，R による作図の利点といえるだろう．

謝辞

データの使用をご承諾くださった田中ゆかり先生，ならびに明治書院に御礼申し上げます．

付記

この研究は科学研究費若手(B)課題番号 16K16846 の一環である．

参考文献

金明哲（2007)『R によるデータサイエンス―データ解析の基礎から最新手法まで―』森北出版．

船尾暢男（2008)『The R Tips 第 2 版』オーム社．

佐藤義治（2009)『シリーズ〈多変量データの統計科学〉2　多変量データの分析―判別分析・クラスター分析―』朝倉書店．

田中ゆかり（2012)「統計ソフトウェア SPSS の利用法―データの読み込みと基礎統計―」荻野綱男・田野村忠温編『講座 IT と日本語研究 8　質問調査法と統計処理』157-203, 明治書院．

山田剛史・杉澤武俊・村井潤一郎（2008)『R によるやさしい統計学』オーム社．

鑓水兼貴（2012)「多変量解析の利用」荻野綱男・田野村忠温編『講座 IT と日本語研究 8 質問調査法と統計処理』205-254, 明治書院．

H.Charles Romesburg 著・西田英郎・佐藤嗣二訳（1992)『実例クラスター分析』内田老鶴圃．

サイト

明治書院「講座 IT と日本語研究」
http://www.meijishoin.co.jp/news/n4066.html（2016 年 5 月 3 日最終確認）

（2016 年 5 月 31 日受付）

Mathematical Linguistics, Vol.30 No.6 (September 2016) pp.378-390.

Tutorial

Data Visualization (6): Making Dendrogram in R statics

HAYASHI Naoki (Nihon University College of Humanities and Sciences)

Abstract:
In this paper, I describe how to create a graph using the statistical software R. When creating a graph, rather than simply pasting the initial output, we alternately list commands to add various changes so that we are able to plot the final output or results. To explain this process, I use a dendrogram. The dendrogram describes how various kinds of information from the initial output can be changed from the command line. It also explains the method of plotting the output or resultant graph onto a Word document. Finally, I also provide a list of points to note when creating a chart in R.

Keywords: R statics, cluster analysis, dendrogram