

catdap2ext (Ver. 0.2.1) マニュアル

1. はじめに

R パッケージ catdap2ext (An extended version of catdap2)は, 「CATDAP 機能強化プラン」(石黒 2021)のうち

- 角度データ (circular data) の扱い
- 条件付き確率のグラフィカル出力

を試みた試作版パッケージである.

R パッケージ catdap の関数 catdap2() は連続型説明変数も扱えるが, 本パッケージはこの関数の機能を拡張した catdap2c() から成る. 時間データや方向データなど角度データと言われる循環性のあるデータを説明変数とする場合に, `表記上の最大値と最小値をまたぐ不等間隔プーリング`も考慮されるようにした.

また, 最適モデルに対するグラフィカル出力は今までの条件付き帯グラフに加え, 目的変数のユーザーが指定する「関心事象 (Events Of Interest, EOI)」に属する確率の説明変数への依存を可視化する「確率地図」および 確率分布棒グラフも出力できるようした. さらに, 目的変数のカテゴリ毎に度数分布棒グラフと連続変数については散布図も表示できるようにした. これにより, ユーザーが視覚的に分かりやすい出力方法を選択可能となった. ただし, 最適なモデルで説明変数が 3 つ以上の場合には 3 番目以降の変数の分布は反映されない.

2. パッケージ catdap2ext のインストールと読み込み

(注記) R はすでにインストールされているとし, あらかじめ RColorBrewer パッケージをインストールしておく必要がある.

2.1 Windows の場合

バイナリファイル catdap2ext_0.2.1.zip を適当なフォルダにダウンロードする.

R (RGui) を起動し, メニュー [Packages] から

-> Install package(s) from local zip files...

-> Select files で ダウンロードした catdap2ext_0.2.1.zip を選択してインストール.

メニュー [Packages] から

-> Load Package..

-> Select one で catdap2ext を選択して読み込む.

2.2 Linux の場合

ソースファイル catdap2ext_0.2.1.tar.gz を適当なディレクトリにダウンロードする。
 ターミナルで R を起動し、

```
> install.packages("ダウンロード先のパス/catdap2ext_0.2.1.tar.gz", repos=NULL)
```

 を実行してインストール。
 インストール先を指定したい場合は、lib="インストール先ディレクトリ" で指定する。
 デフォルトでは、>.libPaths()で最初に表示されるディレクトリにインストールされる。

```
> library(catdap2ext)
```

 で catdap2ext を読み込む。

3. 例題の実行例

3.1 気象データ (windrain)

このデータは以下の 3 項目についての 3 年間の観測データである。

- DAYoftheYEAR: 年間通算日 (1-366)
- rain: 降水量の有無 (0: 0.5 mm 以下 / 1: 1.0 mm 以上)
- winddirection: 風向十六方位 (0-15), 欠測値あり (無風の場合の風向は欠測となる)

このデータについて rain を目的変数とし DAYoftheYEAR と winddirection を角度データとして解析した結果は以下のようになる。最適なモデルでは、DAYoftheYEAR は 12 月下旬から翌年の 2 月下旬までが表記上の最大値 12 月 31 日と最小値 1 月 1 日をまたぐ一つのカテゴリの範囲となっている。

```
data(windrain)
x1 <- catdap2c(windrain, pool = c(3, 2, 3), response.name = "rain",
               accuracy = c(1, 0, 1), missingmark = 10000, print.level = 1)
```

<< List of single explanatory variables (arranged in ascending order of AIC) >>

Response variable : rain (base AIC = 1728.76)

	Explanatory	Number of	Difference		
	variables	categories	A I C	of AIC	Weight
		of exp. var.			

1	winddirection	6	-90.32	0.00	1.00

2	DAYoftheYEAR	9	-86.21	4.11	0.13
---	--------------	---	--------	------	------

<< Two-way tables arranged in ascending order of AIC >>

(rain)			
	0	1	Total
(winddirection)			
1	390 (67.9)	184 (32.1)	574 (100.0)
2	442 (79.5)	114 (20.5)	556 (100.0)
3	134 (90.5)	14 (9.5)	148 (100.0)
4	47 (43.5)	61 (56.5)	108 (100.0)
5	40 (54.8)	33 (45.2)	73 (100.0)
6	2 (100.0)	0 (0.0)	2 (100.0)
Total	1055 (72.2)	406 (27.8)	1461 (100.0)

<Note>

winddirection

category	value range
1	0.00000e+00 - 3.50000e+00
2	3.50000e+00 - 1.05000e+01
3	1.05000e+01 - 1.35000e+01
4	1.35000e+01 - 1.45000e+01
5	1.45000e+01 - 1.50000e+01
6	missing of type 1

	0	1	Total
(DAYoftheYEAR)			
1	80 (94.1)	5 (5.9)	85 (100.0)
2	140 (87.5)	20 (12.5)	160 (100.0)
3	312 (75.0)	104 (25.0)	416 (100.0)
4	34 (53.1)	30 (46.9)	64 (100.0)
5	130 (67.7)	62 (32.3)	192 (100.0)
6	103 (53.6)	89 (46.4)	192 (100.0)
7	60 (62.5)	36 (37.5)	96 (100.0)

8	117 (73.1)	43 (26.9)	160 (100.0)		
9	79 (82.3)	17 (17.7)	96 (100.0)		

Total	1055 (72.2)	406 (27.8)	1461 (100.0)		
<Note>					
DAYoftheYEAR					
category	value range				
1	3.60500e+02 - 1.65000e+01				
2	1.65000e+01 - 5.65000e+01				
3	5.65000e+01 - 1.60500e+02				
4	1.60500e+02 - 1.76500e+02				
5	1.76500e+02 - 2.24500e+02				
6	2.24500e+02 - 2.72500e+02				
7	2.72500e+02 - 2.96500e+02				
8	2.96500e+02 - 3.36500e+02				
9	3.36500e+02 - 3.60500e+02				
<< Summary of subsets of explanatory variables >>					
Response variable : rain					

Explanatory variables		Number of categories of exp. var.	A I C	Difference of AIC	Weight

1	winddirection	20	-208.30	0.00	1.00
	DAYoftheYEAR				
2	winddirection	6	-90.32	117.98	0.00
3	DAYoftheYEAR	9	-86.21	122.09	0.00
4	- - -	0	0.00	208.30	0.00
<< Contingency table constructed by the best subset of explanatory variables >>					
X(1) : rain					
X(2) : winddirection					
X(3) : DAYoftheYEAR					

X X		response variable X(1)		
(2)	(3)	1	2	Total

1	1	106 (90.6)	11 (9.4)	117 (100.0)
1	2	95 (70.4)	40 (29.6)	135 (100.0)
1	3	24 (32.9)	49 (67.1)	73 (100.0)
1	4	32 (45.1)	39 (54.9)	71 (100.0)
1	5	133 (74.7)	45 (25.3)	178 (100.0)
2	1	78 (94.0)	5 (6.0)	83 (100.0)
2	2	204 (82.9)	42 (17.1)	246 (100.0)
2	3	137 (78.3)	38 (21.7)	175 (100.0)
2	4	61 (64.9)	33 (35.1)	94 (100.0)
2	5	96 (90.6)	10 (9.4)	106 (100.0)
3	1	36 (80.0)	9 (20.0)	45 (100.0)
3	2	13 (37.1)	22 (62.9)	35 (100.0)
3	3	3 (37.5)	5 (62.5)	8 (100.0)
3	4	10 (37.0)	17 (63.0)	27 (100.0)
3	5	25 (37.9)	41 (62.1)	66 (100.0)
4	1	0 (0.0)	0 (0.0)	0 (0.0)
4	2	0 (0.0)	0 (0.0)	0 (0.0)
4	3	0 (0.0)	0 (0.0)	0 (0.0)
4	4	0 (0.0)	0 (0.0)	0 (0.0)
4	5	2 (100.0)	0 (0.0)	2 (100.0)

Total		1055 (72.2)	406 (27.8)	1461 (100.0)
<Note>				
X(1) : rain				
	category	variable value		
	1	0		
	2	1		
X(2) : winddirection				
	category	value range		
	1	0.00000e+00 - 3.50000e+00		
	2	3.50000e+00 - 1.35000e+01		
	3	1.35000e+01 - 1.50000e+01		

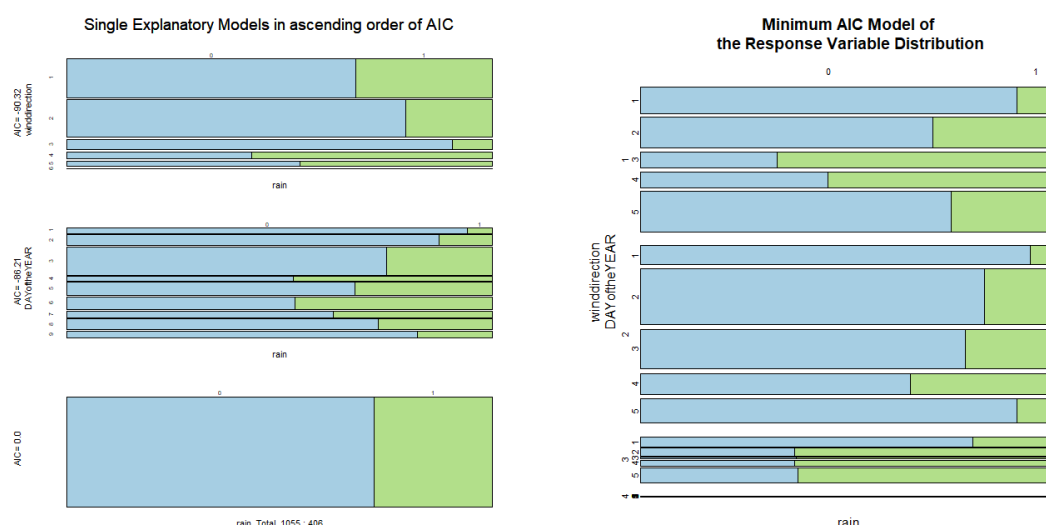
```

4      missing of type 1
X(3) : DAYoftheYEAR
category    value range
1      3.60500e+02 - 5.65000e+01
2      5.65000e+01 - 1.60500e+02
3      1.60500e+02 - 2.24500e+02
4      2.24500e+02 - 2.72500e+02
5      2.72500e+02 - 3.60500e+02

AIC = -208.30
base AIC = 1728.76

```

以下の図は、左側が各説明変数に対する二次元クロス表に対する、右側は AIC 最小モデルに対するプロット出力 (デフォルト: 帯グラフ) であり, catdap パッケージでも同様に出力される.

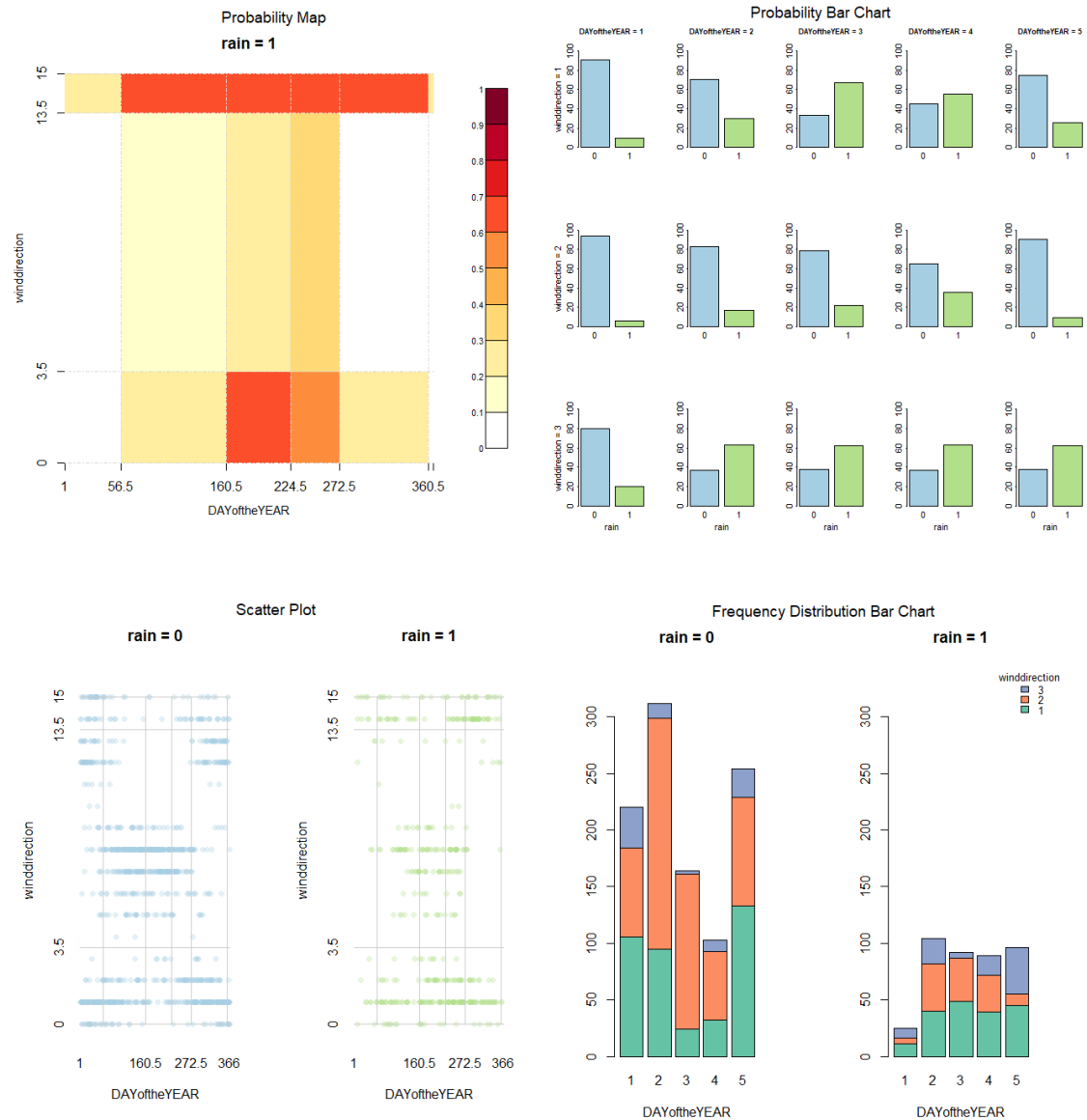


次に本パッケージで追加された 4 種のプロット出力を示す.
 確率分布図は, 最適なモデルのクロス表の比率(上記出力における ()内の値)を濃淡で表現したものである.
 上段左は rain = 1 という事象を関心事象とする「確率地図」である. 梅雨から台風シーズンにかけて北よりの風のときに降雨確率が高いことが視覚的に分かりやすい.

```

plot(x1, type = c("prob", "bar1", "scatter", "bar2"), col = "RD", eoi = 2)

```

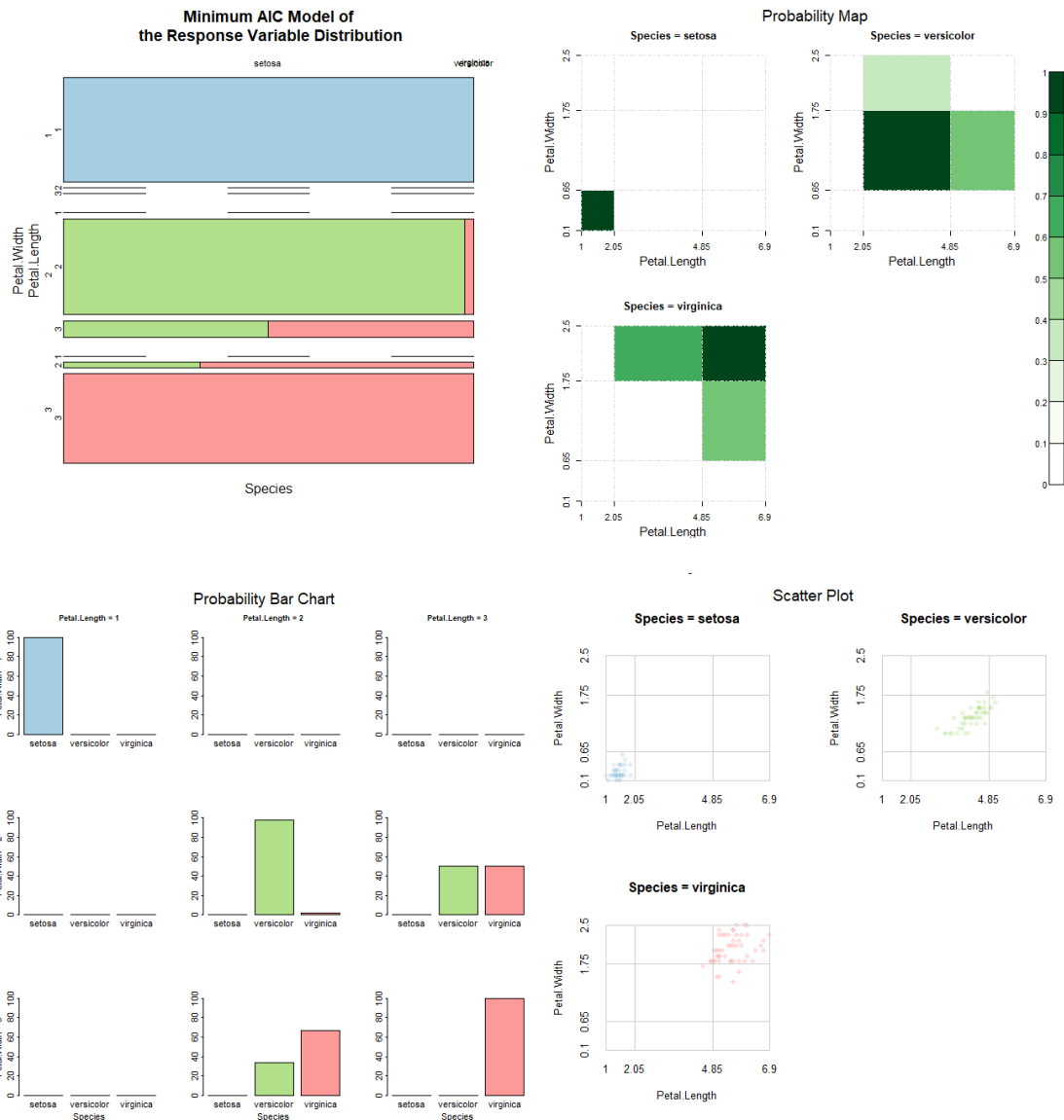


3.2 iris データ

このデータは角度データではないが、AIC 最小モデルに対するプロットと追加されたタイプのプロット出力のみ示す。

以下は、Species を目的変数とした場合の AIC 最小モデルの説明変数が Petal.Width と Petal.Length となった結果の出力である。

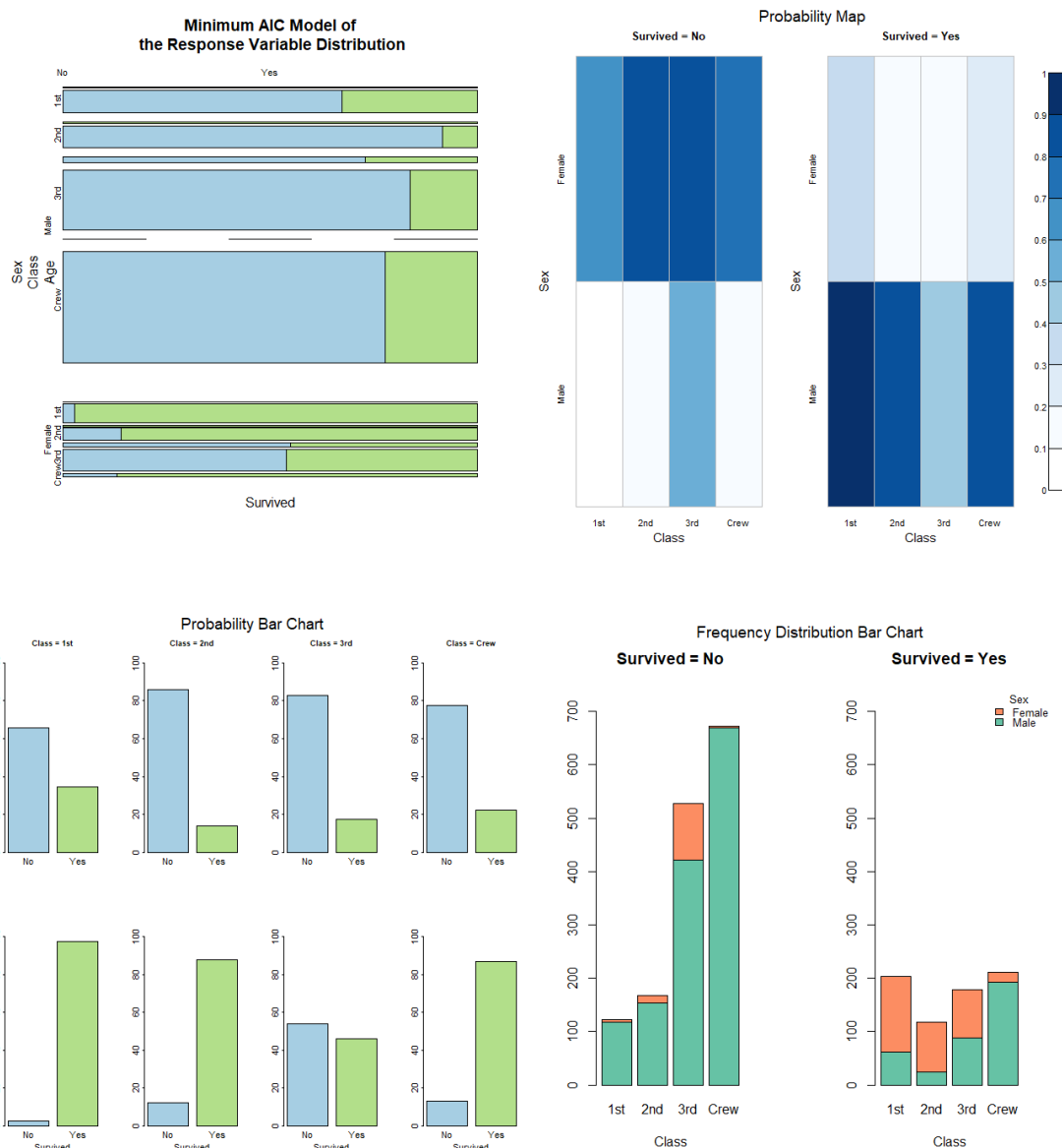
```
x2 <- catdap2c(iris, pool = c(0,0,0,0,2), response.name = "Species",
               accuracy = c(0.1, 0.1, 0.1, 0.1, 0))
plot(x2, type = c("prob", "bar1", "scatter"), col = "GN")
```



3.3 Titanic データ

目的変数も説明変数もカテゴリカルデータの場合の例である。散布図は連続型変数の分布を視覚的に表現するのに適しているが、カテゴリカルデータのみの場合には分かりにくい。下段右は散布図の代わりに目的変数別の棒グラフを表示した例である。

```
d1 <- as.data.frame(Titanic)
TitaData <- d1[rep(seq_len(nrow(d1)), d1$Freq),1:4]
row.names(TitaData) <- NULL
x3 <- catdap2c(TitaData, c(2,2,2,2), "Survived", c(0,0,0,0))
plot(x3, type = c("prob", "bar1", "bar2"), col = "BU")
```



4. 参考文献

- [1] Y.Sakamoto and H.Akaike (1978). Analysis of Cross-Classified Data by AIC. Ann. Inst. Statist. Math., 30, pp.185-197.
- [2] K.Katsura and Y.Sakamoto (1980). A Categorical Data Analysis Program Package, Computer Science Monographs, No.14. The Institute of Statistical Mathematics, Tokyo.
- [3] Y.Sakamoto (1985). Categorical Data Analysis by AIC, Kluwer Academic publishers.
- [4] (株) NTT データ数理システム (2015). 情報量統計学的データ可視化ツール.
<http://hdl.handle.net/10787/3614>
- [5] 石黒 真木夫 (2016) CATDAP マニュアル. <http://hdl.handle.net/10787/3821>
- [6] 石黒 真木夫 (2016) 統計モデル可視化. <http://hdl.handle.net/10787/3823>
- [7] 石黒 真木夫 (2021) CATDAP 機能強化プラン. <http://hdl.handle.net/10787/00034178>