データの価値を究めるTRAD

データサイエンスコンソーシアム，慶應義塾大学

柴田　里程

データの価値評価

- 手元に蓄積したデータ
  - どれだけの利用価値があるのか
  - どう活用できるか
  - だれが活用できるのか
- データの流通
  - 適切な価格
  - 需要先
  - マーケット

データ解析の一つの目的はこのような価値を見極めること

TRAD (TextilePlot, R and DandD)

- データサイエンスコンソーシアムの長年の蓄積と大勢の努力の賜物
- データサイエンスの基礎理論と実践の反映
- 業務使用に耐えるだけの完成度

- データサイエンスの健全な発展を願って無償で公開 (Windows, Mac)
  - http://datascience.jp
  - 日々アップデート, なにか気づいたことがあれば
    query@datascience.jp までご一報のほど.

Visual Excel

- 表計算ソフト
  - 簡単につかえ便利
  - 数字や文字の羅列を眺めていても，どう解析したらよいのか何のアイディアも浮かばない
  - とりあえずは，いろいろな図表を作ってみる(図表の海に溺れる）
  - 適当な方法を適用して結果らしきものを出す（自信，確信はない）
- R
  - 駆使できるようになるにはかなりの修練が必要
  - 修練を積んでも手間がかかることには変わりない
    - データの全体像をつかむには手間と時間がかかる
    - データを眺めるというより，データを扱うための高度な機能を備えたソフトウエア

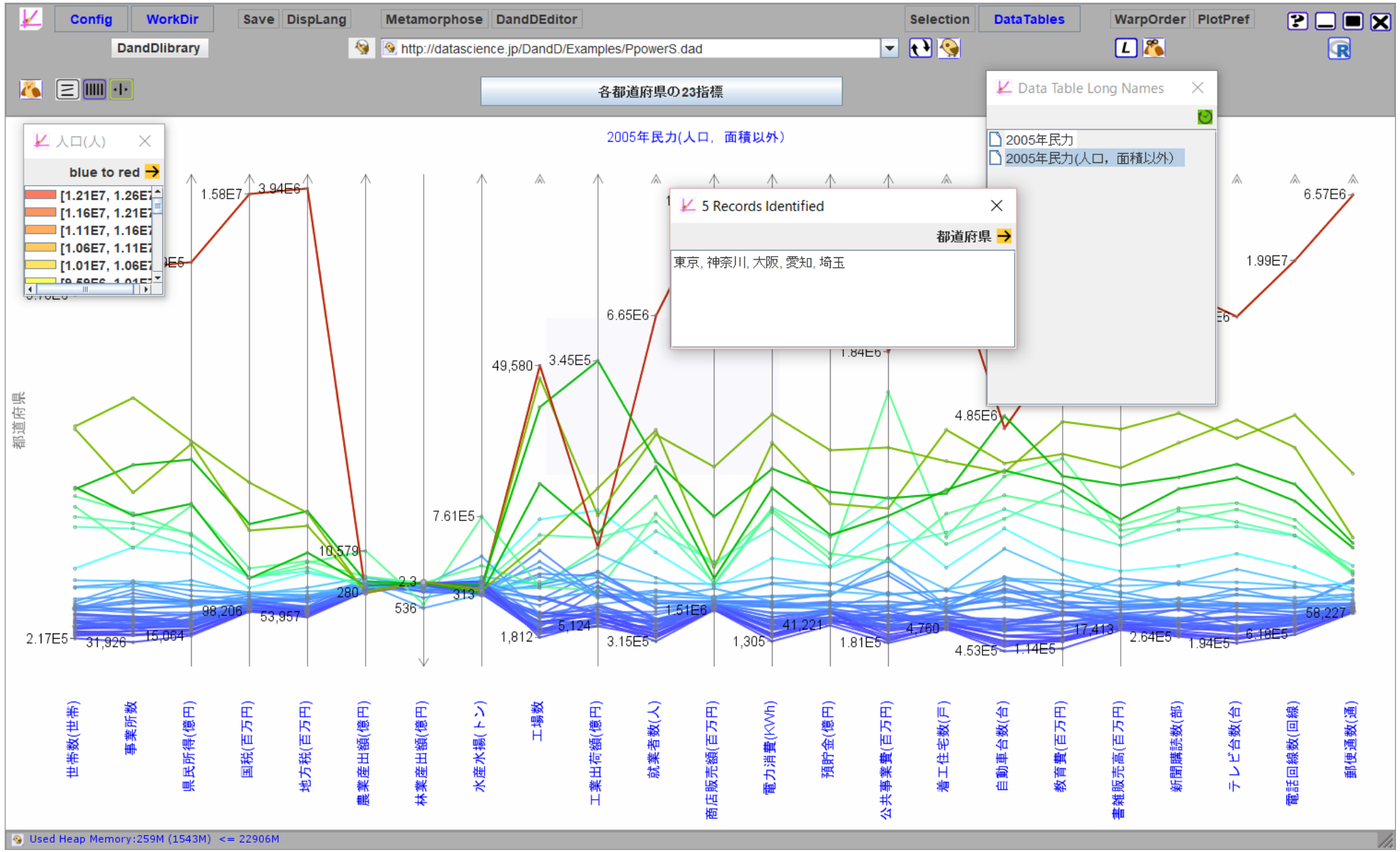
一年間の TRAD 研究開発の軌跡

- インストールの簡便化
- カラーリング
- Logicals to Mark
- Projections
- 高速化とメモリー使用量の削減 (Excel と同等かそれ以上の性能）
- Parallel Coordinate Plot の導入
- 線形モデル当てはめ結果の Parallel Coordinate Plot による表示

インストール

- あらかじめ Java Runtime Environment をインストールしておく（無償）
- TRADのインストールはワンクリック (Windows, MacOS)
  - 40MB 程度
  - 本体は約1MB, ソースコードで約 5 万行
  - 立ち上げ時メモリー消費量は約 100MB
- すでに R がインストールされていれば，自動的に連携
- アイコン  をクリックするだけで起動

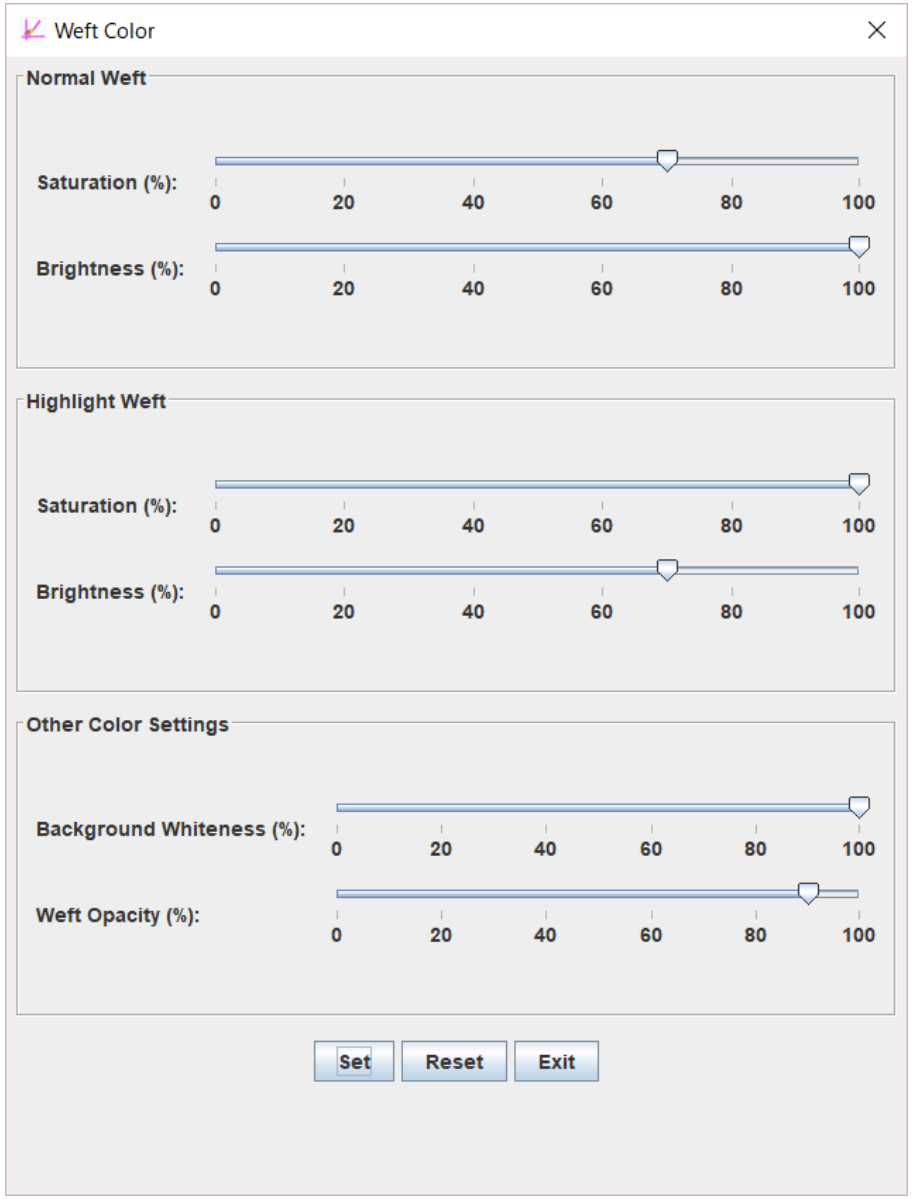
- オフラインマニュアルも電子ブック(ePub) 形式で提供


Config
WorkDir
Save
DispLang
Metamorphose
DandDEditor
Selection
DataTables
WarpOrder
PlotPref
DandDlibrary
http://datascience.jp/DandD/Examples/PpowerS.dad
各都道府県の23指標
2005年民力(人口，面積以外)
人口(人)
blue to red
[1.21E7, 1.26E7
[1.16E7, 1.21E7
[1.11E7, 1.16E7
[1.06E7, 1.11E7
[1.01E7, 1.06E7
Data Table Long Names
2005年民力
2005年民力(人口，面積以外)
5 Records Identified
都道府県
東京, 神奈川, 大阪, 愛知, 埼玉
都道府県
1.58E7
3.94E6
6.65E6
1.84E6
6.57E6
1.99E7
49,580
3.45E5
4.85E6
7.61E5
10,579
2.3
280
313
536
2.17E5
31,926
15,064
98,206
53,957
1,812
5,124
3.15E5
1.51E6
1,305
41,221
1.81E5
4,760
4.53E5
1.14E5
17,413
2.64E5
1.94E5
6.18E5
58,227
世帯数(世帯)
事業所数
県民所得(億円)
国税(百万円)
地方税(百万円)
農業産出額(億円)
林業産出額(億円)
水産水揚(トン)
工場数
工業出荷額(億円)
就業者数(人)
商店販売額(百万円)
電力消費(KWh)
預貯金(億円)
公共事業費(百万円)
着工住宅数(戸)
自動車台数(台)
教育費(百万円)
書雑販売高(百万円)
新聞購読数(部)
テレビ台数(台)
電話回線数(回線)
郵便通数(通)
Used Heap Memory:259M (1543M) <= 22906M

カラーリング (HSBモデル）

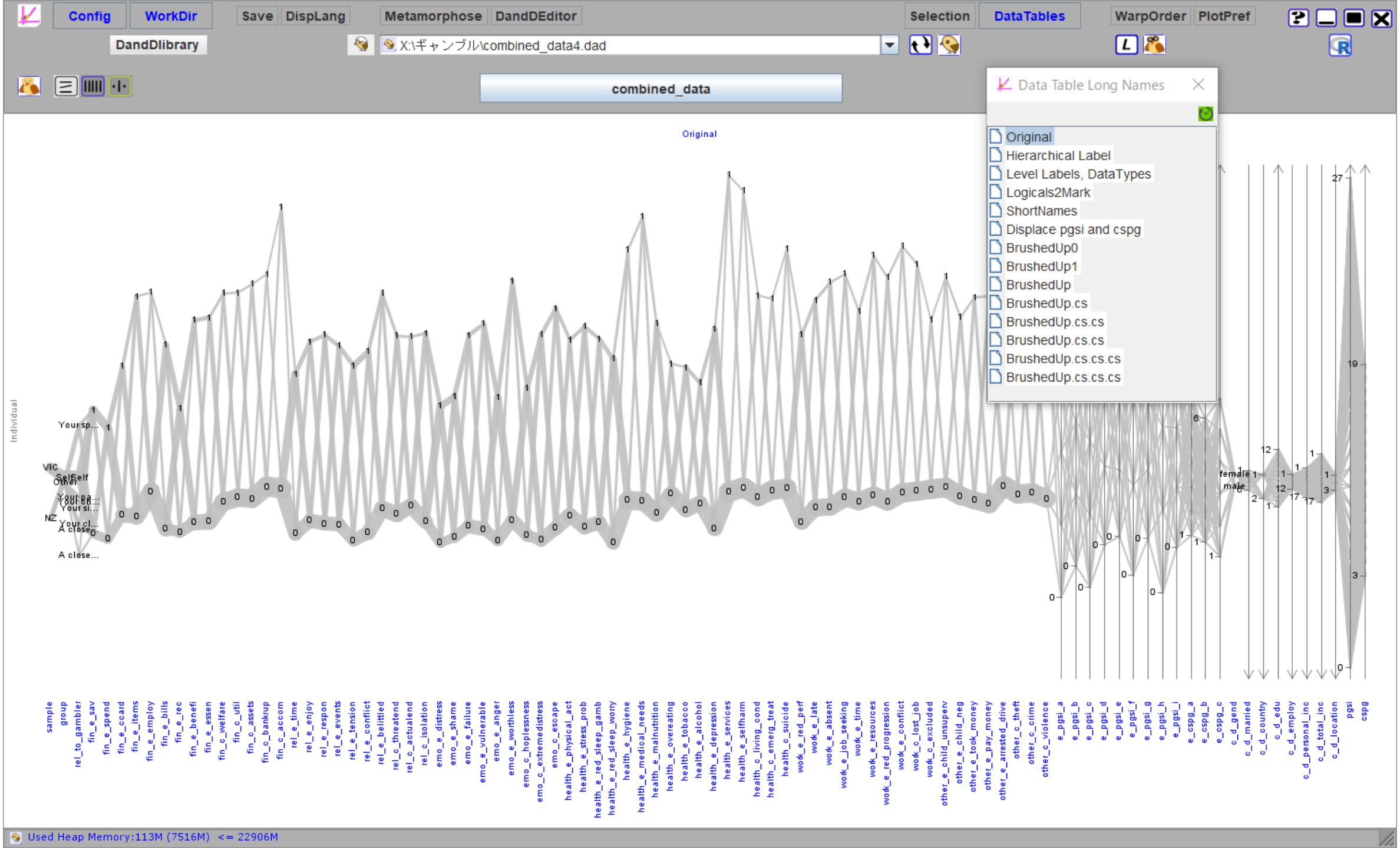
- 各記録あるいは水準に色相(Hue) を割り当てる
  - ハイライトと同じようにマウス操作で割り当てる
  - 割り当ての基準となるデータベクトルを選ぶ
- Lint (2つの軸に挟まれた部分での Weft の断片）
  - 画面の解像度に合わせて Thinning を行う段階で色相を合成する
  - ハイライトは彩度(Saturation)の違いで表現する
  - 基本の彩度や明度（Brightness) はユーザが自由に変更できる
- 描画は画面の下側から行う. ディフォルトの不透明度は90 %

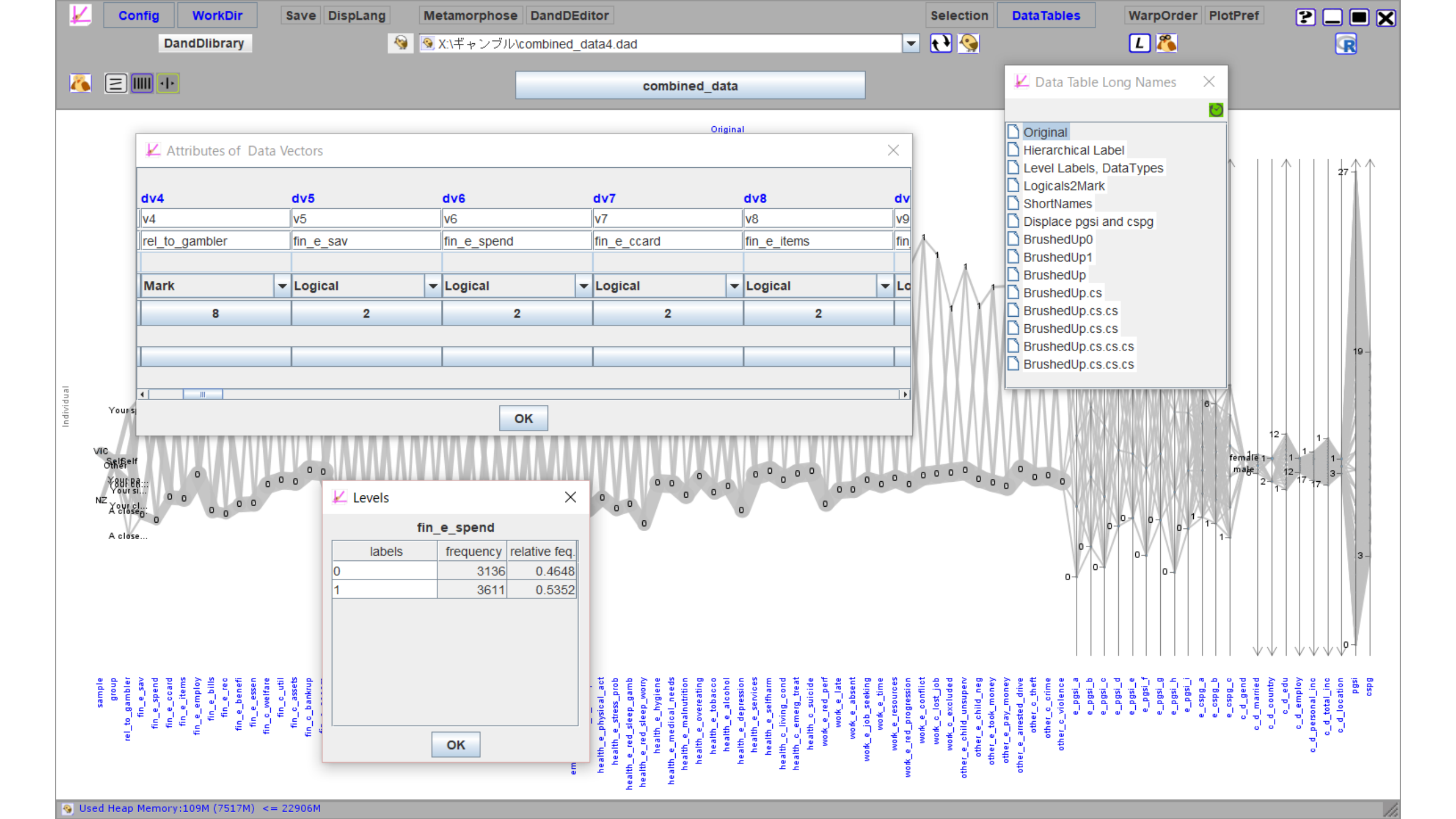
- 軸上のノードの色も同様な手順で行う. 必ずしも Lint 色と同等になるとは限らない

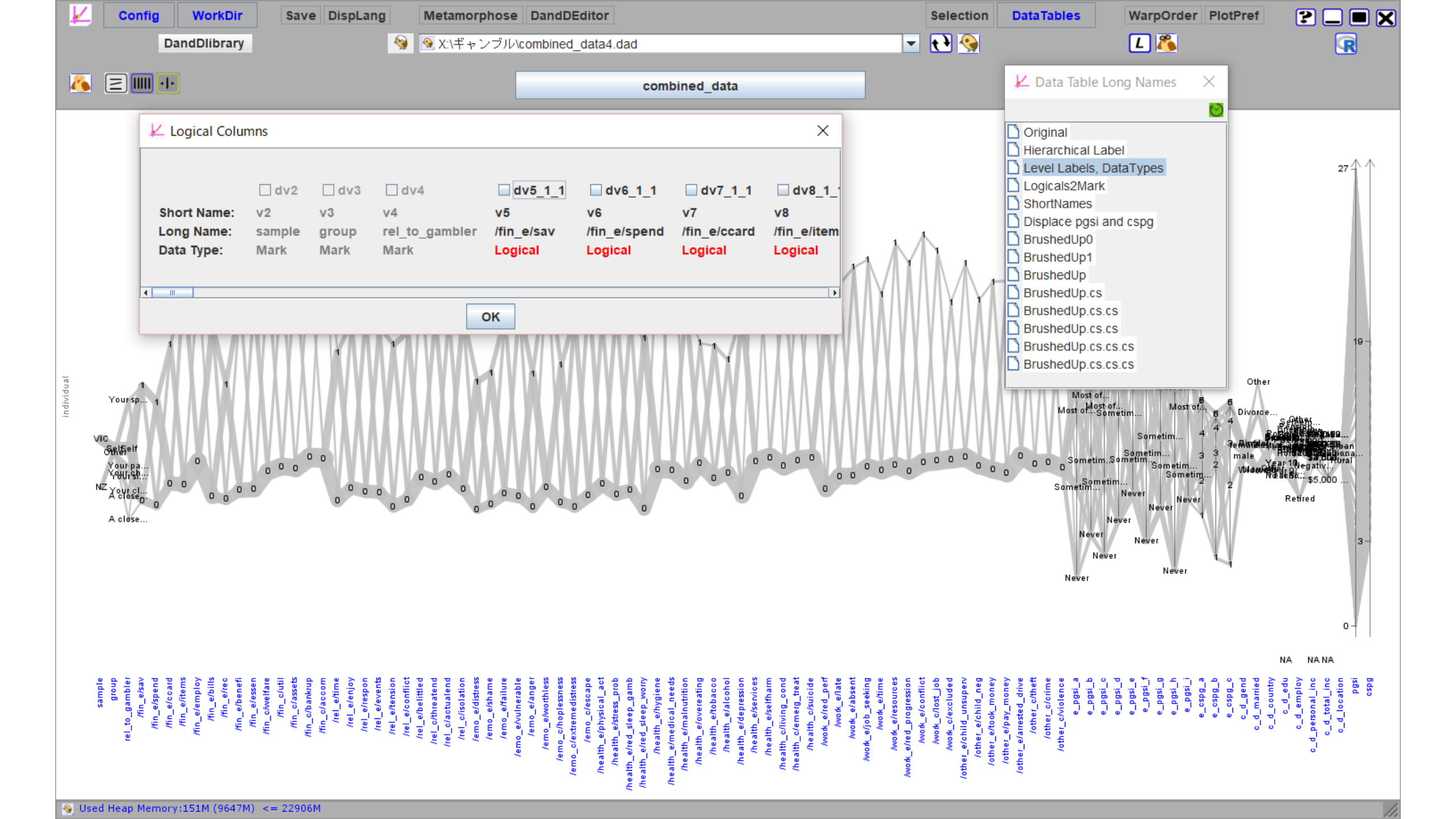

Weft Color
Normal Weft
Saturation (%):
0 20 40 60 80 100
Brightness (%):
0 20 40 60 80 100
Highlight Weft
Saturation (%):
0 20 40 60 80 100
Brightness (%):
0 20 40 60 80 100
Other Color Settings
Background Whiteness (%):
0 20 40 60 80 100
Weft Opacity (%):
0 20 40 60 80 100
Set
Reset
Exit

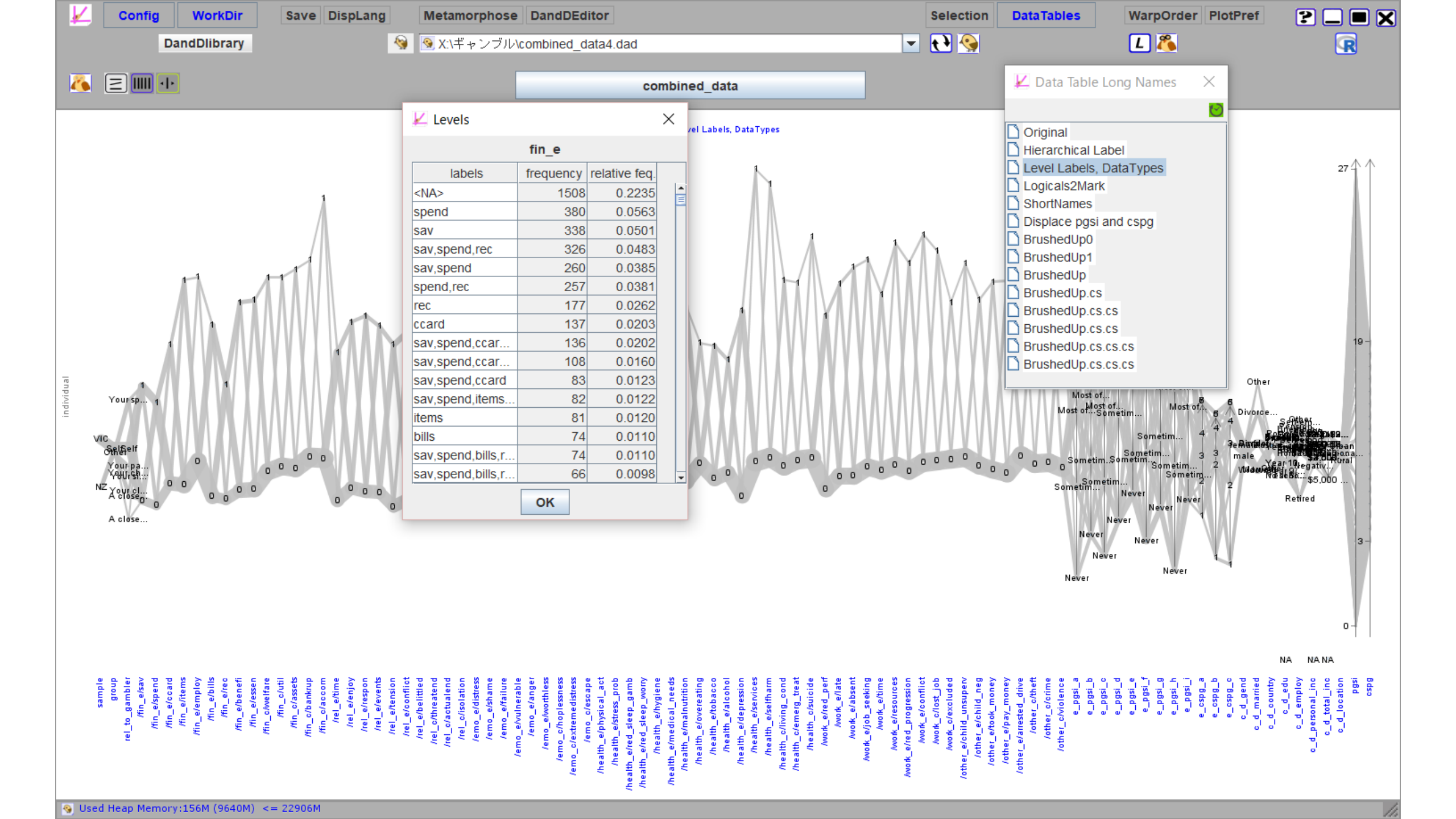
Logicals to Mark

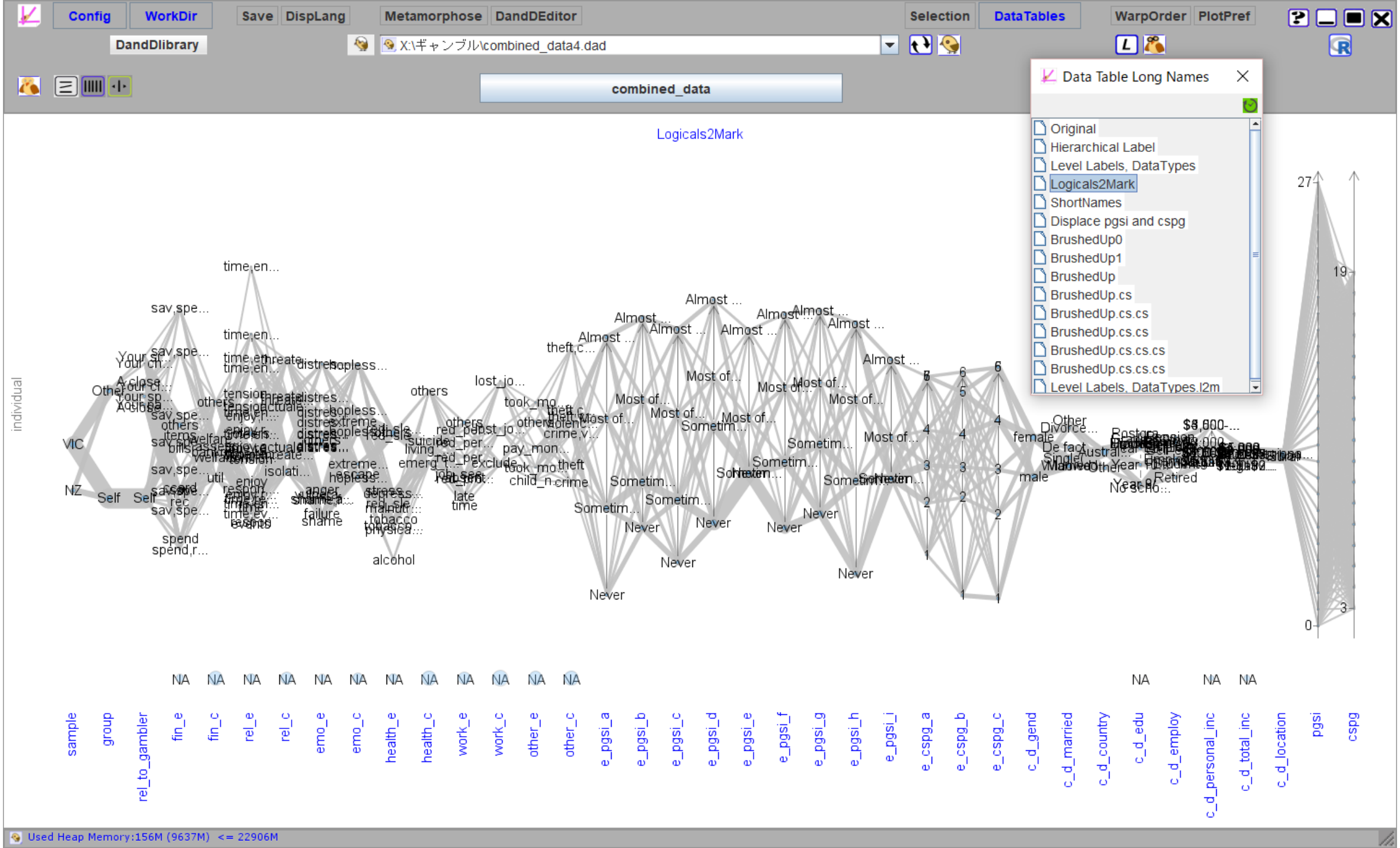
- 複数選択可のアンケート回答
  - 選択したかどうかをあらわす TRUE (1) あるいは FALSE(0) の列
  - このままでは疎でありすぎるので，Mark 型に直したい
  - しかも回答項目が多いのでシステマティックに直したい

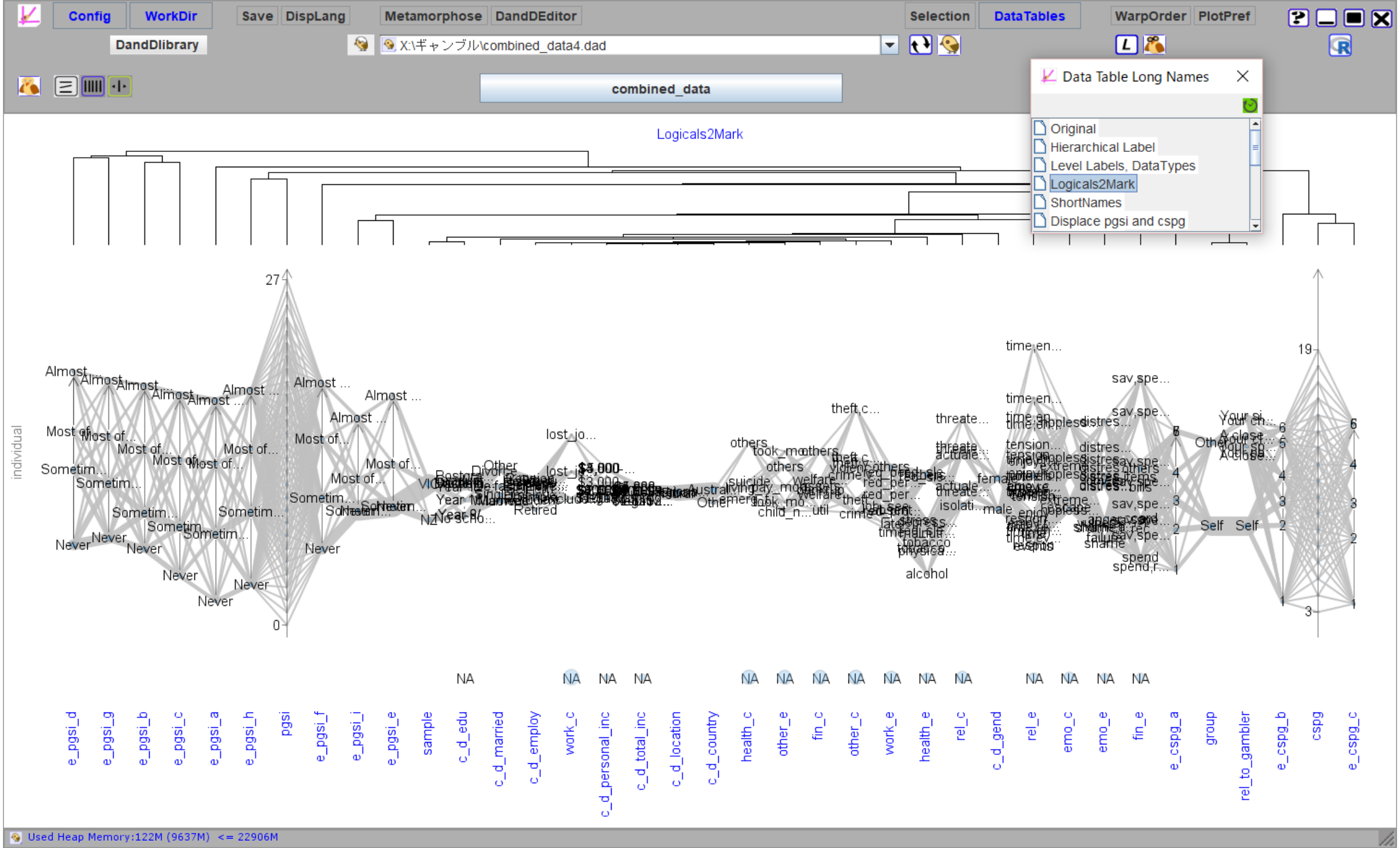

Config
WorkDir
Save
DispLang
Metamorphose
DandDEditor
Selection
DataTables
WarpOrder
PlotPref
DandDlibrary
X:\ギャンブル\combined_data4.dad
combined_data
Data Table Long Names
Original
Hierarchical Label
Level Labels, DataTypes
Logicals2Mark
ShortNames
Displace pgsi and cspg
BrushedUp0
BrushedUp1
BrushedUp
BrushedUp.cs
BrushedUp.cs.cs
BrushedUp.cs.cs
BrushedUp.cs.cs.cs
BrushedUp.cs.cs.cs
Original
Individual
Used Heap Memory:113M (7516M) <= 22906M


Config
WorkDir
Save
DispLang
Metamorphose
DandDEditor
Selection
DataTables
WarpOrder
PlotPref
DandDlibrary
X:\ギャンブル\combined_data4.dad
combined_data
Data Table Long Names
Original
Hierarchical Label
Level Labels, DataTypes
Logicals2Mark
ShortNames
Displace pgsi and cspg
BrushedUp0
BrushedUp1
BrushedUp
BrushedUp.cs
BrushedUp.cs.cs
BrushedUp.cs.cs
BrushedUp.cs.cs.cs
BrushedUp.cs.cs.cs
Attributes of Data Vectors
dv4
dv5
dv6
dv7
dv8
v4
v5
v6
v7
v8
rel_to_gambler
fin_e_sav
fin_e_spend
fin_e_ccard
fin_e_items
Mark
Logical
8
2
OK
Levels
fin_e_spend
labels
frequency
relative feq.
0
3136
0.4648
1
3611
0.5352
OK
Individual
female
male
Used Heap Memory:109M (7517M) <= 22906M


Config
WorkDir
Save
DispLang
Metamorphose
DandDEditor
Selection
DataTables
WarpOrder
PlotPref
DandDlibrary
X:\ギャンブル\combined_data4.dad
combined_data
Logical Columns
dv2
dv3
dv4
dv5_1_1
dv6_1_1
dv7_1_1
dv8_1_1
Short Name:
v2
v3
v4
v5
v6
v7
v8
Long Name:
sample
group
rel_to_gambler
/fin_e/sav
/fin_e/spend
/fin_e/ccard
/fin_e/item
Data Type:
Mark
Mark
Mark
Logical
Logical
Logical
Logical
OK
Data Table Long Names
Original
Hierarchical Label
Level Labels, DataTypes
Logicals2Mark
ShortNames
Displace pgsi and cspg
BrushedUp0
BrushedUp1
BrushedUp
BrushedUp.cs
BrushedUp.cs.cs
BrushedUp.cs.cs
BrushedUp.cs.cs.cs
BrushedUp.cs.cs.cs
individual
Used Heap Memory:151M (9647M) <= 22906M

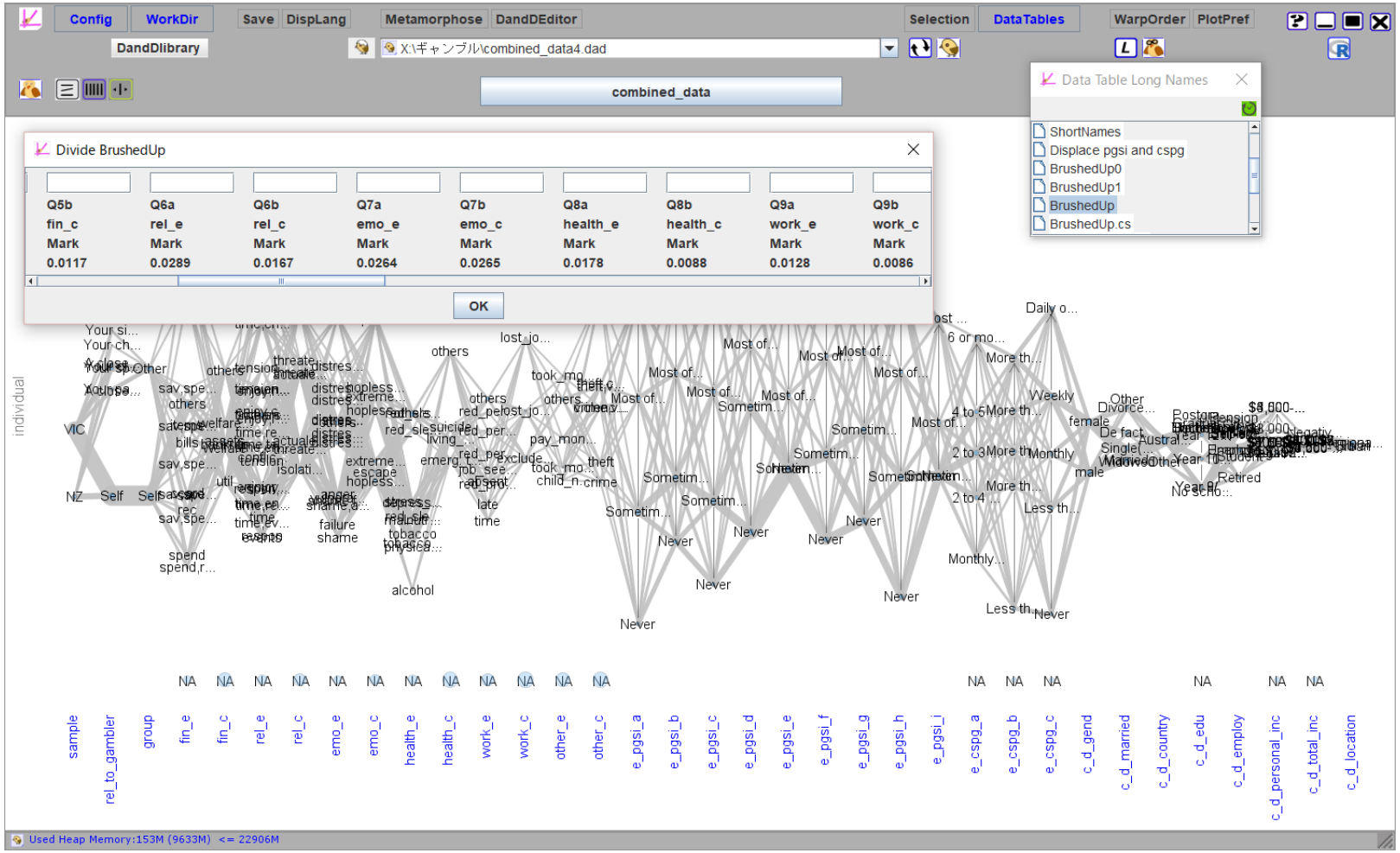

Config
WorkDir
Save
DispLang
Metamorphose
DandDEditor
Selection
DataTables
WarpOrder
PlotPref
DandDlibrary
X:\ギャンブル\combined_data4.dad
combined_data
Levels
fin_e
labels | frequency | relative feq.
<NA> | 1508 | 0.2235
spend | 380 | 0.0563
sav | 338 | 0.0501
sav,spend,rec | 326 | 0.0483
sav,spend | 260 | 0.0385
spend,rec | 257 | 0.0381
rec | 177 | 0.0262
ccard | 137 | 0.0203
sav,spend,ccar... | 136 | 0.0202
sav,spend,ccar... | 108 | 0.0160
sav,spend,ccard | 83 | 0.0123
sav,spend,items... | 82 | 0.0122
items | 81 | 0.0120
bills | 74 | 0.0110
sav,spend,bills,r... | 74 | 0.0110
sav,spend,bills,r... | 66 | 0.0098
OK
Data Table Long Names
Original
Hierarchical Label
Level Labels, DataTypes
Logicals2Mark
ShortNames
Displace pgsi and cspg
BrushedUp0
BrushedUp1
BrushedUp
BrushedUp.cs
BrushedUp.cs.cs
BrushedUp.cs.cs
BrushedUp.cs.cs.cs
BrushedUp.cs.cs.cs
Used Heap Memory:156M (9640M) <= 22906M

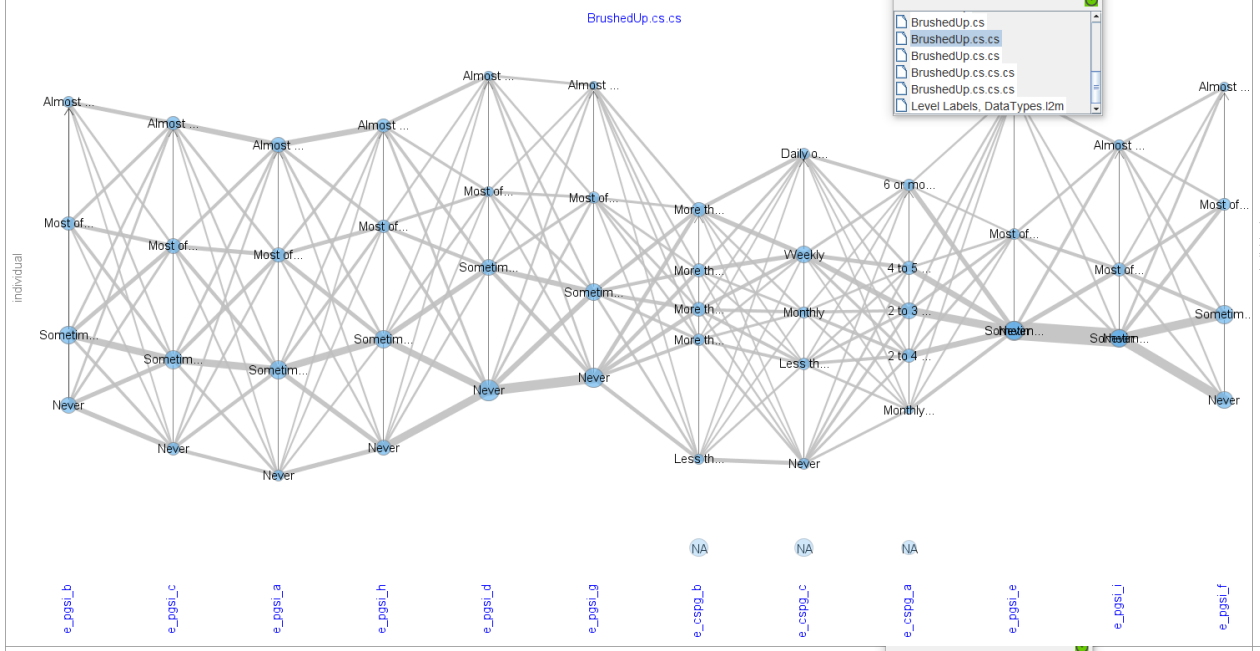

Config
WorkDir
Save
DispLang
Metamorphose
DandDEditor
Selection
DataTables
WarpOrder
PlotPref
DandDlibrary
X:\ギャンブル\combined_data4.dad
combined_data
Data Table Long Names
Original
Hierarchical Label
Level Labels, DataTypes
Logicals2Mark
ShortNames
Displace pgsi and cspg
BrushedUp0
BrushedUp1
BrushedUp
BrushedUp.cs
BrushedUp.cs.cs
BrushedUp.cs.cs
BrushedUp.cs.cs.cs
BrushedUp.cs.cs.cs
Level Labels, DataTypes.l2m
Logicals2Mark
individual
sample
group
rel_to_gambler
fin_e
fin_c
rel_e
rel_c
emo_e
emo_c
health_e
health_c
work_e
work_c
other_e
other_c
e_pgsi_a
e_pgsi_b
e_pgsi_c
e_pgsi_d
e_pgsi_e
e_pgsi_f
e_pgsi_g
e_pgsi_h
e_pgsi_i
e_cspg_a
e_cspg_b
e_cspg_c
c_d_gend
c_d_married
c_d_country
c_d_edu
c_d_employ
c_d_personal_inc
c_d_total_inc
c_d_location
pgsi
cspg
Used Heap Memory:156M (9637M) <= 22906M

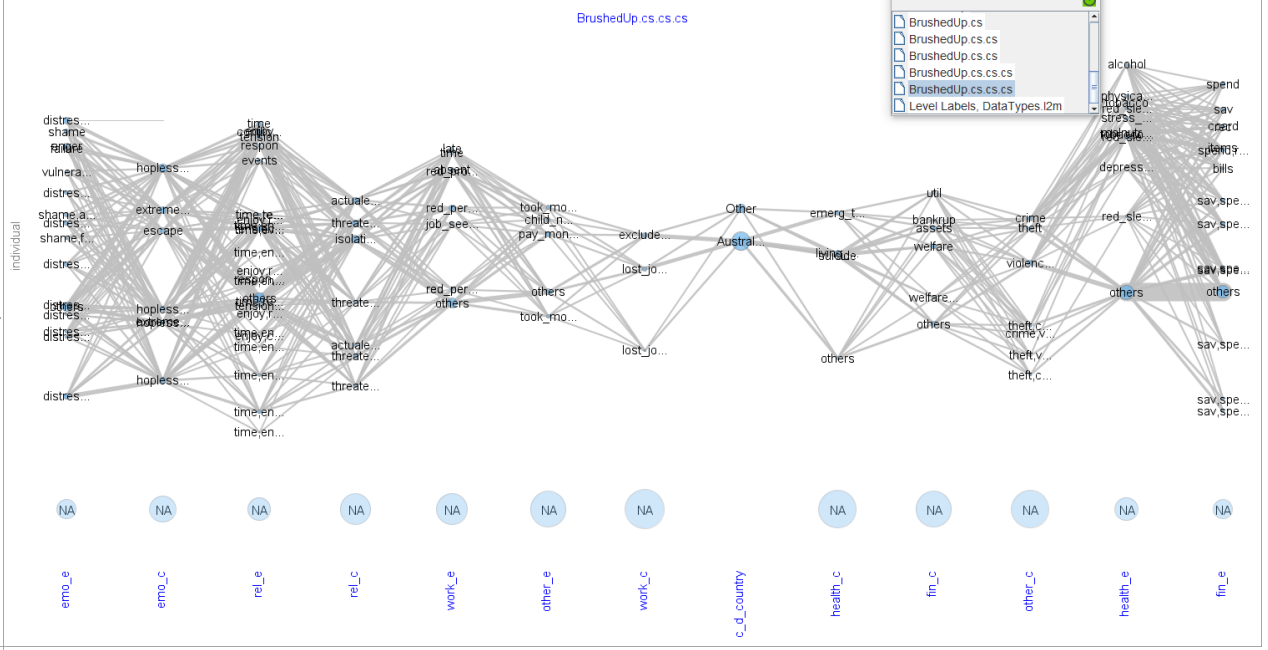

Config
WorkDir
Save
DispLang
Metamorphose
DandDEditor
Selection
DataTables
WarpOrder
PlotPref
DandDlibrary
X:\ギャンブル\combined_data4.dad
combined_data
Data Table Long Names
Original
Hierarchical Label
Level Labels, DataTypes
Logicals2Mark
ShortNames
Displace pgsi and cspg
Logicals2Mark
individual
27
0
19
1
3
e_pgsi_d
e_pgsi_g
e_pgsi_b
e_pgsi_c
e_pgsi_a
e_pgsi_h
pgsi
e_pgsi_f
e_pgsi_i
e_pgsi_e
sample
c_d_edu
c_d_married
c_d_employ
work_c
c_d_personal_inc
c_d_total_inc
c_d_location
c_d_country
health_c
other_e
fin_c
other_c
work_e
health_e
rel_c
c_d_gend
rel_e
emo_c
emo_e
fin_e
e_cspg_a
group
rel_to_gambler
e_cspg_b
cspg
e_cspg_c
Used Heap Memory:122M (9637M) <= 22906M

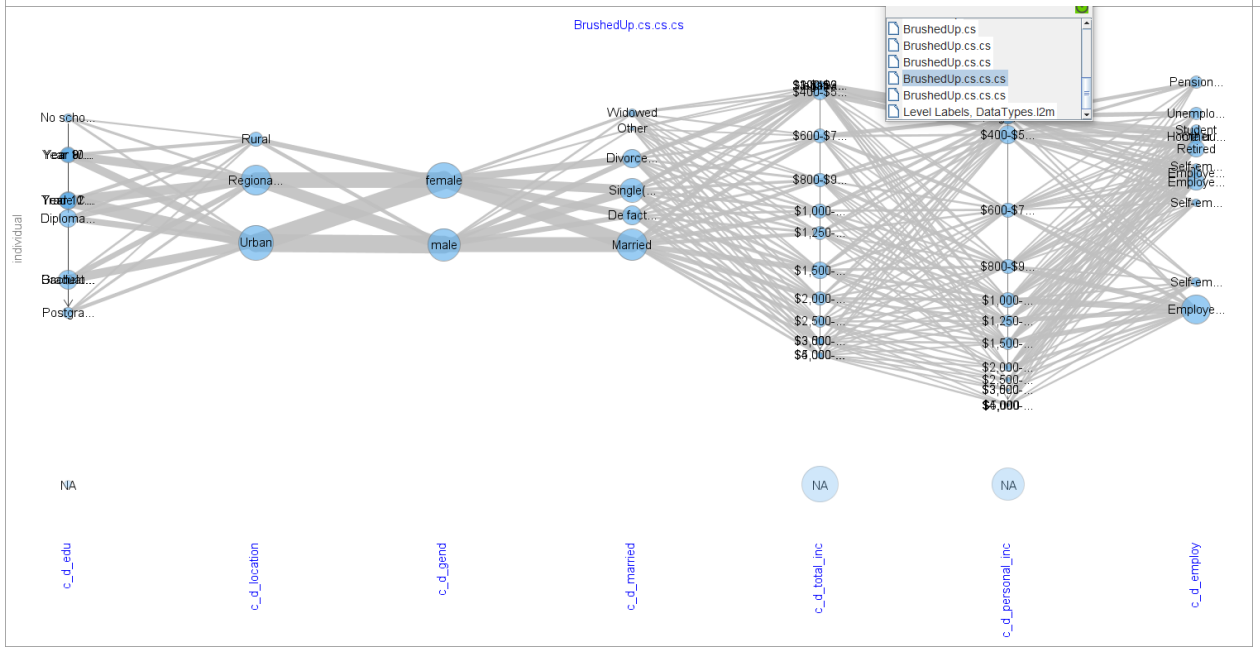
Projections

- 変量（軸）をいくつか選んで，一つのデータテーブルからいくつかのサブデータテーブルを作り出す




BrushedUp.cs.cs
BrushedUp.cs
BrushedUp.cs.cs
BrushedUp.cs.cs
BrushedUp.cs.cs.cs
BrushedUp.cs.cs.cs
Level Labels, DataTypes.l2m
individual
Almost ...
Most of...
Sometim...
Never
Daily o...
6 or mo...
Weekly
4 to 5 ...
Monthly
2 to 3 ...
Less th...
2 to 4 ...
Monthly...
More th...
NA
e_pgsi_b
e_pgsi_c
e_pgsi_a
e_pgsi_h
e_pgsi_d
e_pgsi_g
e_cspg_b
e_cspg_c
e_cspg_a
e_pgsi_e
e_pgsi_i
e_pgsi_f


BrushedUp.cs.cs.cs
BrushedUp.cs
BrushedUp.cs.cs
BrushedUp.cs.cs
BrushedUp.cs.cs.cs
BrushedUp.cs.cs.cs
Level Labels, DataTypes.l2m
individual
distres...
hopless...
extreme...
escape
time
events
actuale...
threate...
isolati...
red_per...
job_see...
others
took_mo...
child_n...
pay_mon...
exclude...
lost_jo...
Other
Austral...
emerg_t...
bankrup
assets
welfare
util
crime
theft
violenc...
alcohol
physica...
depress...
spend
sav
card
bills
sav,spe...
others
NA
emo_e
emo_c
rel_e
rel_c
work_e
other_e
work_c
c_d_country
health_c
fin_c
other_c
health_e
fin_e


BrushedUp.cs.cs.cs
BrushedUp.cs
BrushedUp.cs.cs
BrushedUp.cs.cs
BrushedUp.cs.cs.cs
BrushedUp.cs.cs.cs
Level Labels, DataTypes.l2m
individual
No scho...
Diploma...
Postgra...
Rural
Regiona...
Urban
female
male
Widowed
Other
Divorce...
Single(...
De fact...
Married
$400-$5...
$600-$7...
$800-$9...
$1,000-...
$1,250-...
$1,500-...
$2,000-...
$2,500-...
$3,000-...
$3,500-...
$4,000-...
$4,500-...
$5,000-...
Pension...
Unemplo...
Student
Retired
Self-em...
Employe...
NA
c_d_edu
c_d_location
c_d_gend
c_d_married
c_d_total_inc
c_d_personal_inc
c_d_employ

高速化とメモリー使用量削減

- 高速化
  - 隠しプログラミングテクニック(Java)
  - 最終的にもっとも時間を費やすのは最後の描画
    - イメージとして保存して再利用
- メモリー使用量の削減
  - ガベージコレクションの効果的な適用
  - 負荷の少ない計算なら結果を保存せず必要になったときに求める(トレードオフ)
- 目安
  - 30列30万行 CSV ファイル(200M) : 描画まで28秒, メモリー 1GB

PlotType の導入

- TextilePlot 
  - 各軸の位置尺度を水平性規準で定める
- Parallel Coordinate Plot
  - 各軸の位置尺度は任意
    - 画面の高さいっぱいまで伸ばす 
    - ママ（位置＝0, 尺度=1) 
    - 位置･尺度を与える (例 : 線形モデル当てはめ結果の表示） 

まとめ

- どのような機能が必要で，どんな形で実装すればよいのかわかっているわけではない
- 実験を重ね，改良を重ねるしかない
- 芸術の側面，美しく楽しい環境
- 様々なデータにする知見の蓄積
- 利用者からのフィードバック
- データの価値評価のための基本ソフトウエアになることを期待