ビッグデータの要約としてのMDS

DATAEXPLORING　出口慎二

ビッグデータ

- ここでは「東日本大震災ビッグデータワークショップ - Project 311 -」で使用したデータを使用.
- おおよそ, 2,000万ツイート×4日のデータを使用.
  - サイズにして, テキストで16GBほど？
  - ロウデータは規約上すでに無い. ここでは当時作った集計データを再利用.
- MDSで使うのは集計データ.
  - 10語の距離行列×4日. → SMACOF
  - 10語の距離行列×4日. → ALSCAL
  - 10語の出現頻度×4日. → 対応分析(ポアソン分布)
  - 10語の出現頻度の時間帯別平均→ 平均値表の対応分析(正規分布)
- 集計データであれば, 4日間8,000万ツイートのデータ処理でも特に問題は無い.
  - 読み込みデータが大きすぎる場合は(桁落ちする場合は)指数化すれば操作できるだろう(この程度であれば, 特に問題は無い. たとえば, Excel vba で長整数型(long)は, -2,147,483,648 ～ 2,147,483,647).

データの要約

- 2段階の処理
  - 1段階目：集計
  - 2段階目：MDS
    - ALSCAL：SPSS
    - SMACOF：R
    - 対応分析：ポアソン分布／正規分布とも手計算(ExcelVBA)，特異値分解は青木(*)を利用．
      注）青木繁伸，http://aoki2.si.gunma-u.ac.jp/，最終アクセス日，2003/10/31．（現在は無い）
- 平均値表の対応分析
  - ポアソン分布（近似）する2相2元表の分析（普通の対応分析）に対して，
    正規分布する2相2元表の分析．
    - ポアソン分布する場合　・・・平均／標準偏差に周辺度数を利用（ポアソン分布だから）．
    - 正規分布する場合　・・・平均／標準偏差は算術平均／標準偏差を使用
      （データから計算しないで，平均値表から重みづけ計算）．
  - ビッグデータの処理に適している．
  - 相関係数行列を得て，その主成分分析をすることも可能かもしれない？相関係数を計算できるなら，重回帰も．

INDSCALモデル（stress）

■ ストレス

◆ ALSCALプログラム

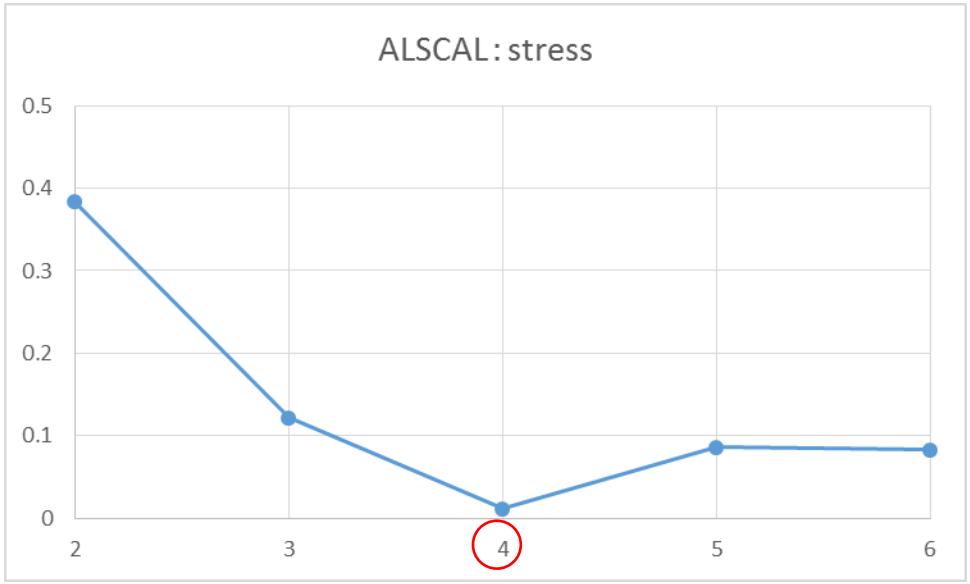

図1. stress（ALSCAL）

◆ SMACOFプログラム

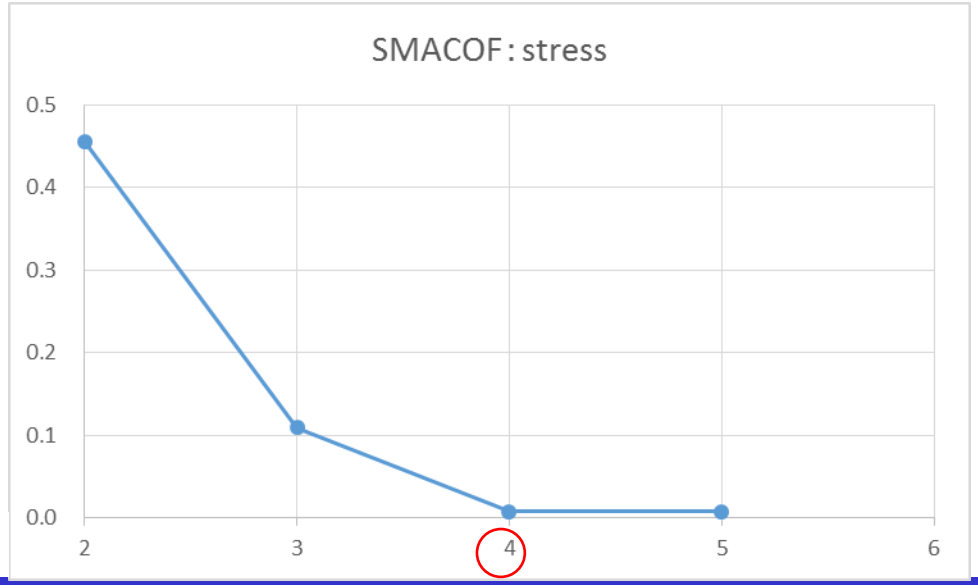

図2. stress（SMACOF）

INDSCALモデル（共通対象布置）

■ 共通対象布置

◆ ALSCALプログラム

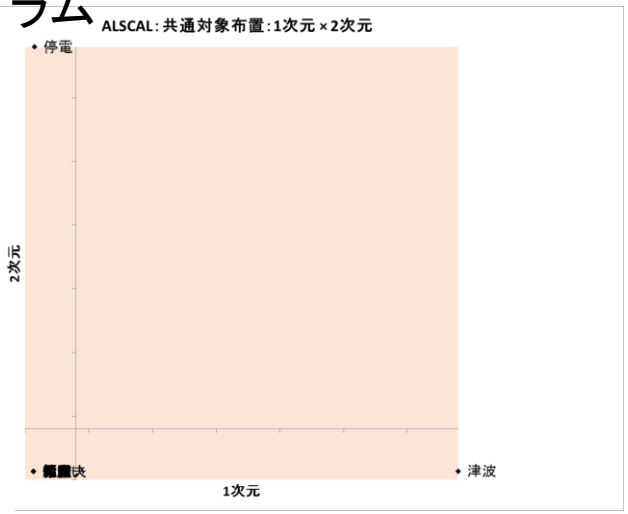

図3. 共通対象布置
1×2次元（ALSCAL）

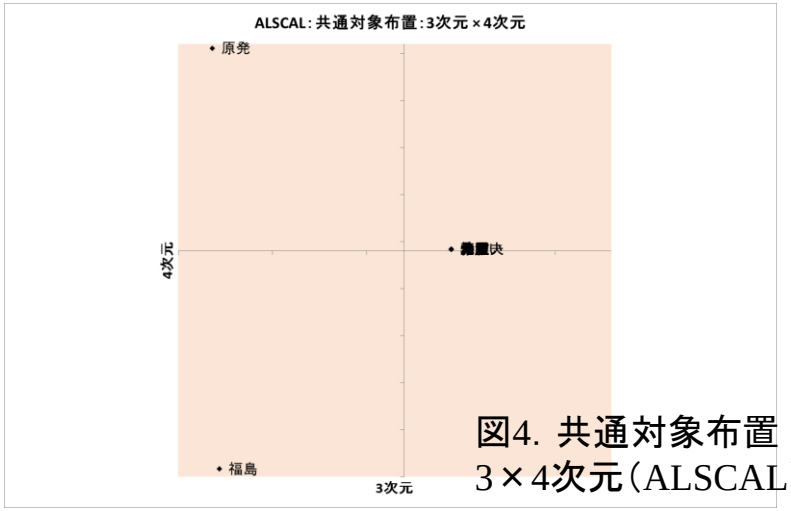

図4. 共通対象布置
3×4次元（ALSCAL）

◆ SMACOFプログラム

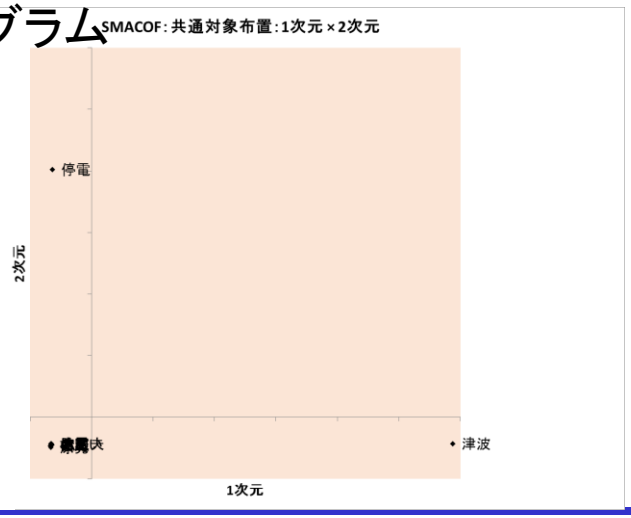

図5. 共通対象布置
1×2次元（SMACOF）

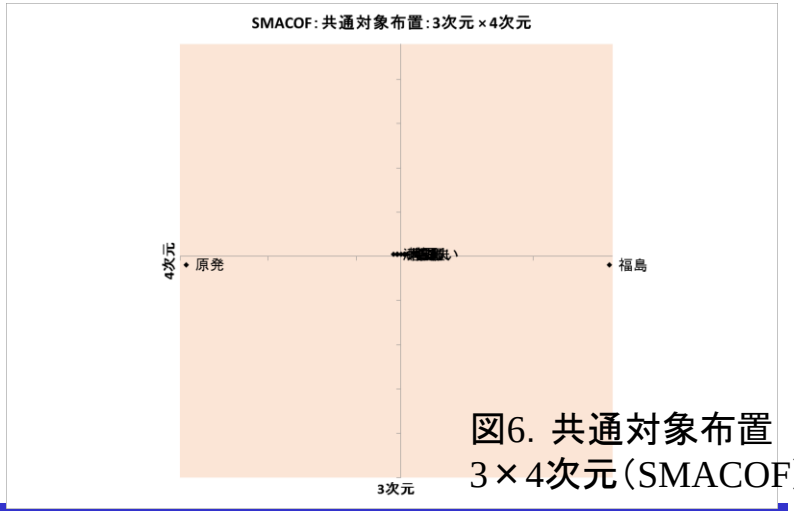

図6. 共通対象布置
3×4次元（SMACOF）

INDSCALモデル（重み布置）

■ 重み布置

◆ ALSCALプログラム

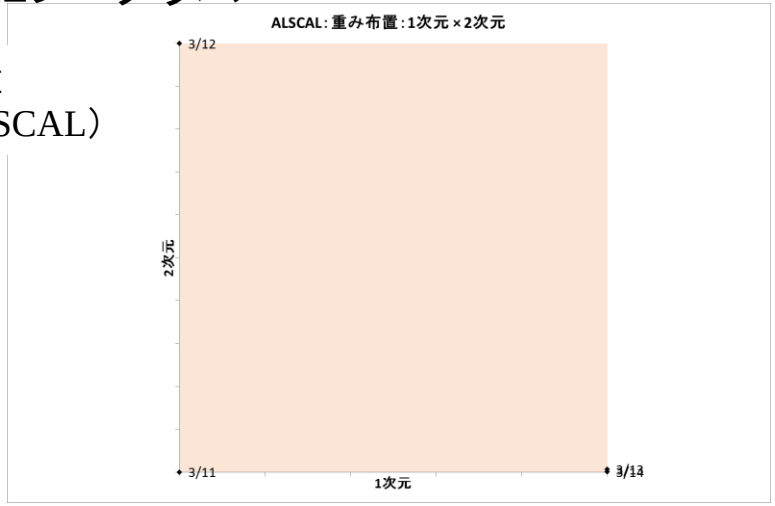

図7. 重み布置
1×2次元（ALSCAL）

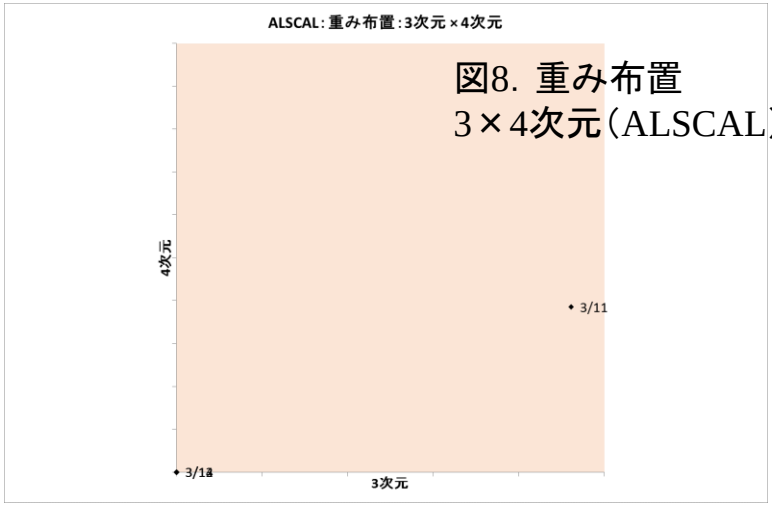

図8. 重み布置
3×4次元（ALSCAL）

◆ SMACOFプログラム

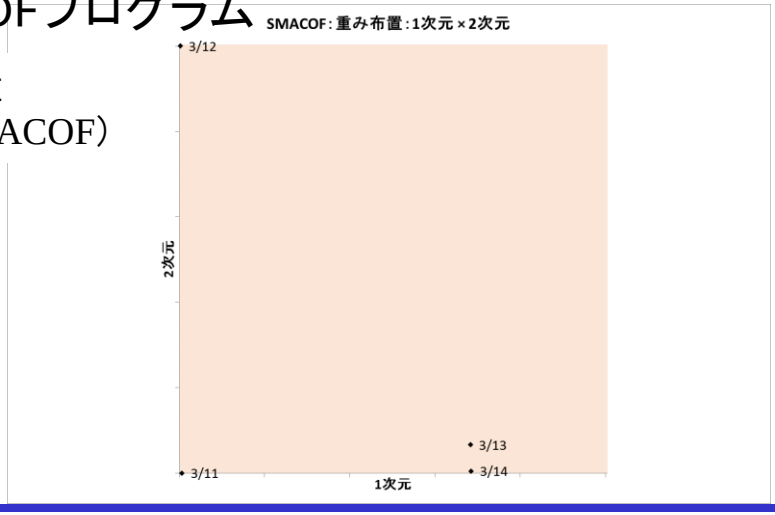

図9. 重み布置
1×2次元（SMACOF）

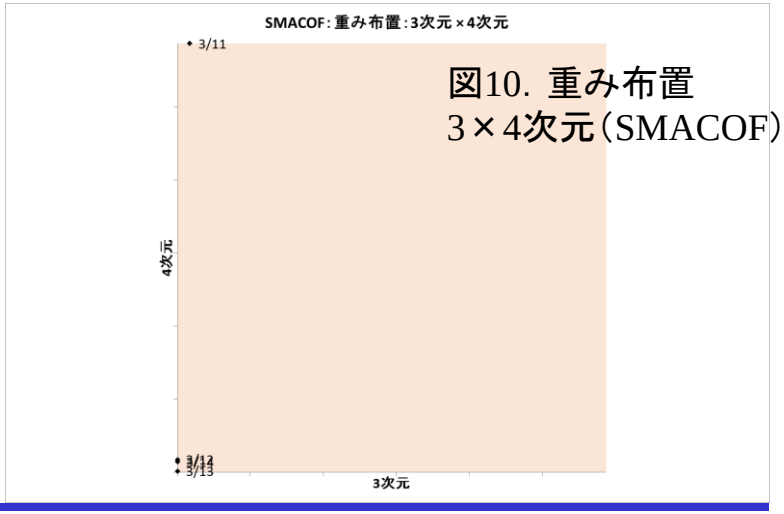

図10. 重み布置
3×4次元（SMACOF）

対応分析

■ ポアソン分布

◆ 日付×語

※行と列の同時付置なので，位置ではなく角度を見る．

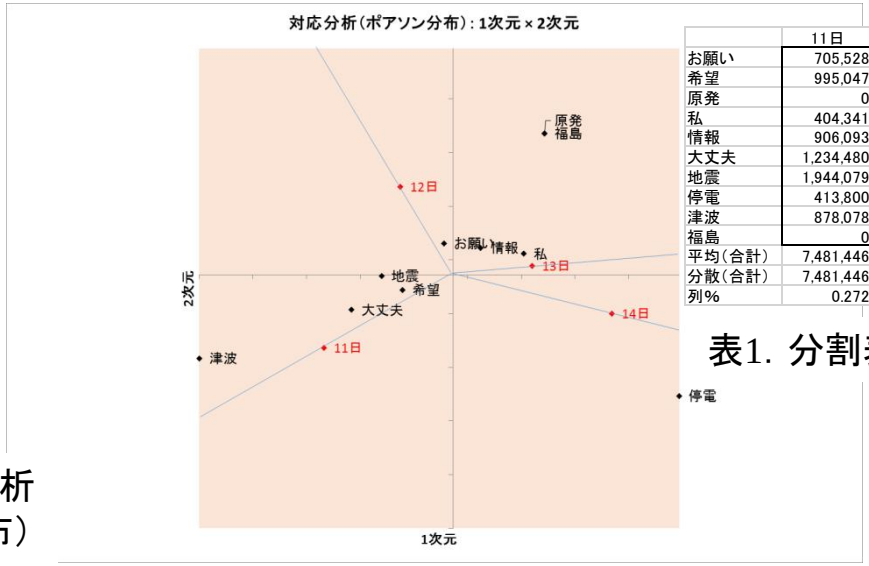

図11. 対応分析（ポアソン分布）

	11日	12日	13日	14日	平均(合計)	分散(合計)	行%
お願い	705,528	972,918	696,708	481,112	2,856,266	2,856,266	0.104
希望	995,047	832,246	605,534	412,002	2,844,829	2,844,829	0.104
原発	0	669,276	281,615	399,926	1,350,817	1,350,817	0.049
私	404,341	718,852	652,397	692,478	2,468,068	2,468,068	0.090
情報	906,093	1,380,632	1,116,394	946,273	4,349,392	4,349,392	0.158
大丈夫	1,234,480	870,197	391,298	349,829	2,845,804	2,845,804	0.104
地震	1,944,079	1,825,111	929,797	721,983	5,420,970	5,420,970	0.197
停電	413,800	0	902,605	1,569,237	2,885,642	2,885,642	0.105
津波	878,078	431,471	0	0	1,309,549	1,309,549	0.048
福島	0	551,723	290,374	309,689	1,151,786	1,151,786	0.042
平均(合計)	7,481,446	8,252,426	5,866,722	5,882,529	27,483,123		
分散(合計)	7,481,446	8,252,426	5,866,722	5,882,529		27,483,123	
列%	0.272	0.300	0.213	0.214			1.000

表1. 分割表（ポアソン分布）

■ 正規分布

◆ 時間×語

※時間を平均して日を計算．

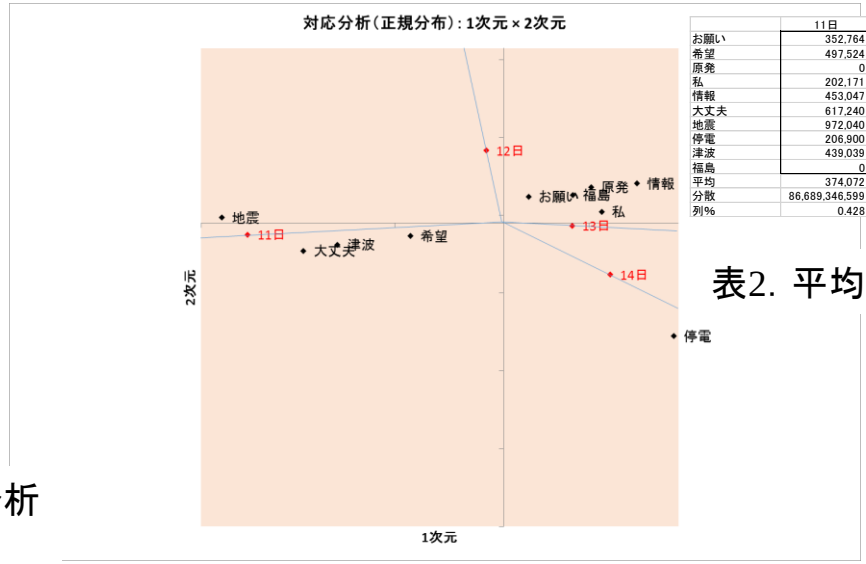

図12. 対応分析（正規分布）

	11日	12日	13日	14日	平均	分散	行%
お願い	352,764	243,230	174,177	120,278	222,612	10,060,940,397	0.102
希望	497,524	208,062	151,384	103,001	239,992	31,320,055,889	0.110
原発	0	167,319	70,404	99,982	84,426	4,812,532,000	0.039
私	202,171	179,713	163,099	173,120	179,526	274,563,235	0.082
情報	453,047	345,158	279,099	236,568	328,468	8,893,756,781	0.150
大丈夫	617,240	217,549	97,825	87,457	255,018	61,798,361,964	0.117
地震	972,040	456,278	232,449	180,496	460,316	130,699,958,660	0.211
停電	206,900	0	225,651	392,309	206,215	25,844,644,382	0.094
津波	439,039	107,868	0	0	136,727	43,204,648,904	0.063
福島	0	137,931	72,594	77,422	71,987	3,186,872,039	0.033
平均	374,072	206,311	146,668	147,063	218,529		
分散	86,689,346,599	15,823,688,235	7,487,115,339	11,651,046,235		36,948,829,782	
列%	0.428	0.236	0.168	0.168			1.000

表2. 平均値表（正規分布）

結果の解釈

- ALSCALとSMACOF
  - 互いにStressで比較したが，同程度.
  - 布置も似たり寄ったり.
  - 3/11（原発・福島）が3次元に出るか，4次元に出るか，の差.

- ポアソン分布の対応分析と正規分布の対応分析
  - 基本的には同様.
    - ポアソン分布について:「津波」は3/11.「地震」は3/11～12. 度数が多いのか原点より.「原発」は3/12以降.「停電」は3/14.
    - 正規分布について:「津波」や「地震」は3/11.「原発」「福島」は3/12～13.「停電」は3/14.

まとめ

- 大きいデータとMDS
  - いったん元の大きなデータを「集計」し，これに対して処理をするMDSは，大きなデータの処理に向いているのではないか．
  - ただし，基本的にはMDSは「記述的」多変量解析．
- 対応分析の別の考え方
  - 対応分析で特異値分解する行列は，ポアソン分布する値の表の標準化行列であると考えると，正規分布する値の表であれば，それを特異値分解する，として，たとえば平均値表の対応分析も考えることが可能ではないか．

	11日	12日	13日	14日	平均(合計)	分散(合計)	行%
お願い	705,528	972,918	696,708	481,112	2,856,266	2,856,266	0.104
希望	995,047	832,246	605,534	412,002	2,844,829	2,844,829	0.104
原発	0	669,276	281,615	399,926	1,350,817	1,350,817	0.049
私	404,341	718,852	652,397	692,478	2,468,068	2,468,068	0.090
情報	906,093	1,380,632	1,116,394	946,273	4,349,392	4,349,392	0.158
大丈夫	1,234,480	870,197	391,298	349,829	2,845,804	2,845,804	0.104
地震	1,944,079	1,825,111	929,797	721,983	5,420,970	5,420,970	0.197
停電	413,800	0	902,605	1,569,237	2,885,642	2,885,642	0.105
津波	878,078	431,471	0	0	1,309,549	1,309,549	0.048
福島	0	551,723	290,374	309,689	1,151,786	1,151,786	0.042
平均(合計)	7,481,446	8,252,426	5,866,722	5,882,529	27,483,123		
分散(合計)	7,481,446	8,252,426	5,866,722	5,882,529		27,483,123	
列%	0.272	0.300	0.213	0.214			1.000

表3. 分割表(ポアソン分布)(表1の再掲)

	11日	12日	13日	14日	平均	分散	行%
お願い	352,764	243,230	174,177	120,278	222,612	10,060,940,397	0.102
希望	497,524	208,062	151,384	103,001	239,992	31,320,055,889	0.110
原発	0	167,319	70,404	99,982	84,426	4,812,532,000	0.039
私	202,171	179,713	163,099	173,120	179,526	274,563,235	0.082
情報	453,047	345,158	279,099	236,568	328,468	8,893,756,781	0.150
大丈夫	617,240	217,549	97,825	87,457	255,018	61,798,361,964	0.117
地震	972,040	456,278	232,449	180,496	460,316	130,699,958,660	0.211
停電	206,900	0	225,651	392,309	206,215	25,844,644,382	0.094
津波	439,039	107,868	0	0	136,727	43,204,648,904	0.063
福島	0	137,931	72,594	77,422	71,987	3,186,872,039	0.033
平均	374,072	206,311	146,668	147,063	218,529		
分散	86,689,346,599	15,823,688,235	7,487,115,339	11,651,046,235		36,948,829,782	
列%	0.428	0.236	0.168	0.168			1.000

表4. 平均値表(正規分布)(表2の再掲)

観測値(O)と期待値(E)の残差(O−E)を行／列の標準偏差で割って標準化する．(※ポアソン分布の場合，行／列計が平均でもあり標準偏差でもある)．

極座標→xy座標系

■ 対応分析の1×2次元図は，極座標表現になっている？

◆ 行と列の布置をするとき，度数が小さいセルは図の外側に付置される．

- 行計／列計が小さく，度数が小さいセル→標準化後の値は大きい→外側に付置される

（図では目立つが，観測度数が小さかったという意味．絶対寄与／相対寄与の計算には大きな影響を持つが，基本的にはそれ以上の以上の意味は？？）
・・・原点からの遠さ

- 原点からの距離が度数（稀さ）を現すだけだとすると，各点の持つ情報は，残りは角度

・・・角度

（図では目立つが，観測度数が小さかったという意味．絶対寄与／相対寄与の計算には大きな影響を持つが，基本的にはそれ以上の以上の意味は？？）

→これは極座標ではないか？

極座標表現した散布図

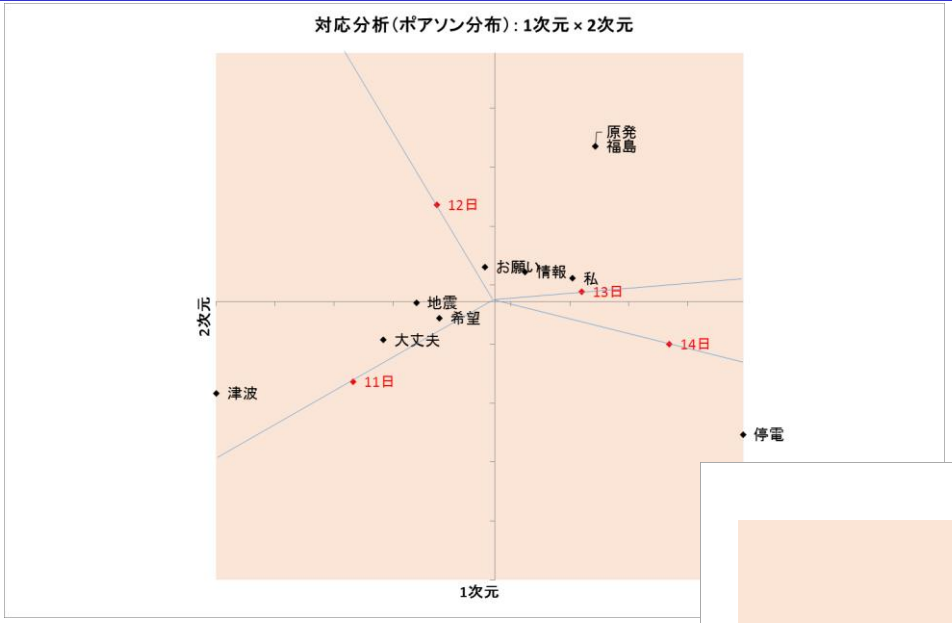

図13. 第1次元×第2次元の散布図

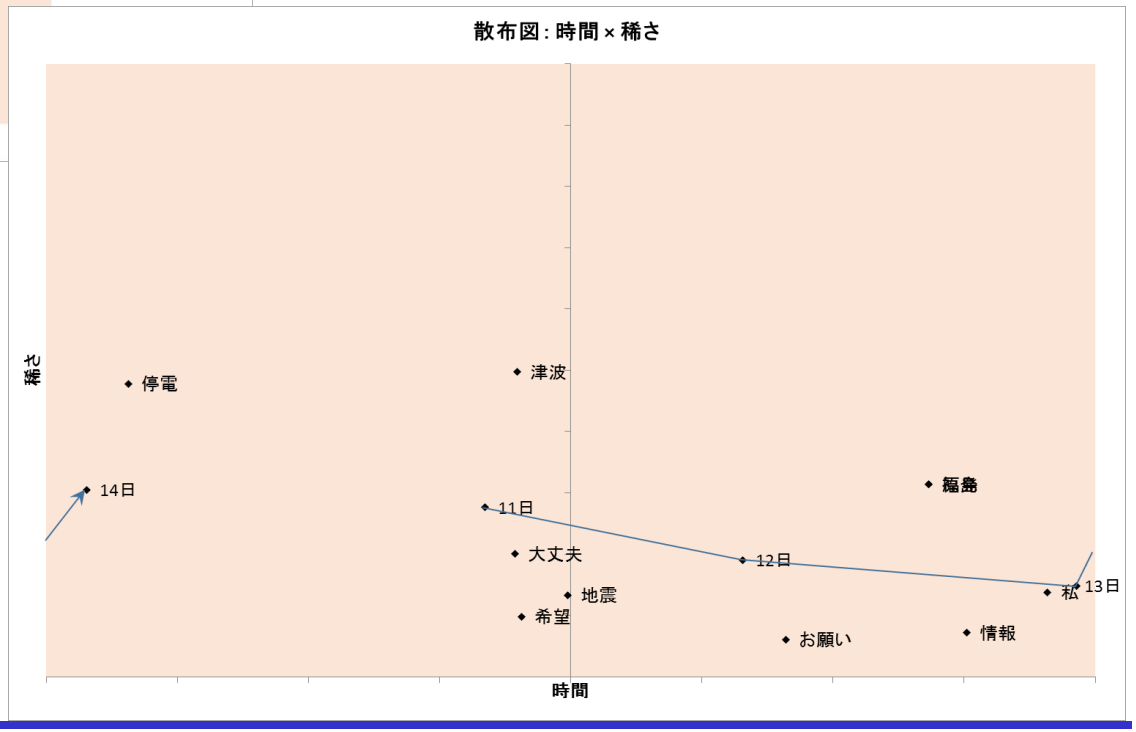

図14. 角度と距離の散布図

座標値

■ 極座標

- もとの x(1次元)，y(2次元)の座標値を，極座標(r,θ)に変換.
- 変換した数値を，普通の(直交座標系の)散布図の x(1次元)，y(2次元)として，散布図を描画.
- 何と呼ぶのか？(とりあえず「角度と距離の散布図」)
- この場合，馬蹄形問題は「問題」ではない.
- この程度の点の数だとピンとこないが，もう少し点が多いと・・・(次ページ)

	1次元	2次元
お願い	−0.034	0.116
希望	−0.188	−0.057
原発	0.341	0.527
私	0.264	0.077
情報	0.103	0.100
大丈夫	−0.378	−0.131
地震	−0.266	−0.005
停電	0.842	−0.452
津波	−0.944	−0.312
福島	0.341	0.527
11日	−0.481	−0.272
12日	−0.196	0.327
13日	0.295	0.033
14日	0.592	−0.146

表5. 2次元の座標値

	1次元	2次元		時間	稀さ
福島	0.341	0.527	福島	122.925	0.627
原発	0.341	0.527	原発	122.921	0.627
13日	0.295	0.033	13日	173.680	0.297
私	0.264	0.077	私	163.654	0.275
情報	0.103	0.100	情報	135.838	0.144
お願い	−0.034	0.116	お願い	73.863	0.121
12日	−0.196	0.327	12日	59.103	0.381
停電	0.842	−0.452	停電	−151.769	0.956
14日	0.592	−0.146	14日	−166.170	0.609
希望	−0.188	−0.057	希望	−16.832	0.196
地震	−0.266	−0.005	地震	−1.070	0.266
大丈夫	−0.378	−0.131	大丈夫	−19.131	0.401
11日	−0.481	−0.272	11日	−29.485	0.552
津波	−0.944	−0.312	津波	−18.303	0.994

表6. 座標値を極座標に変換

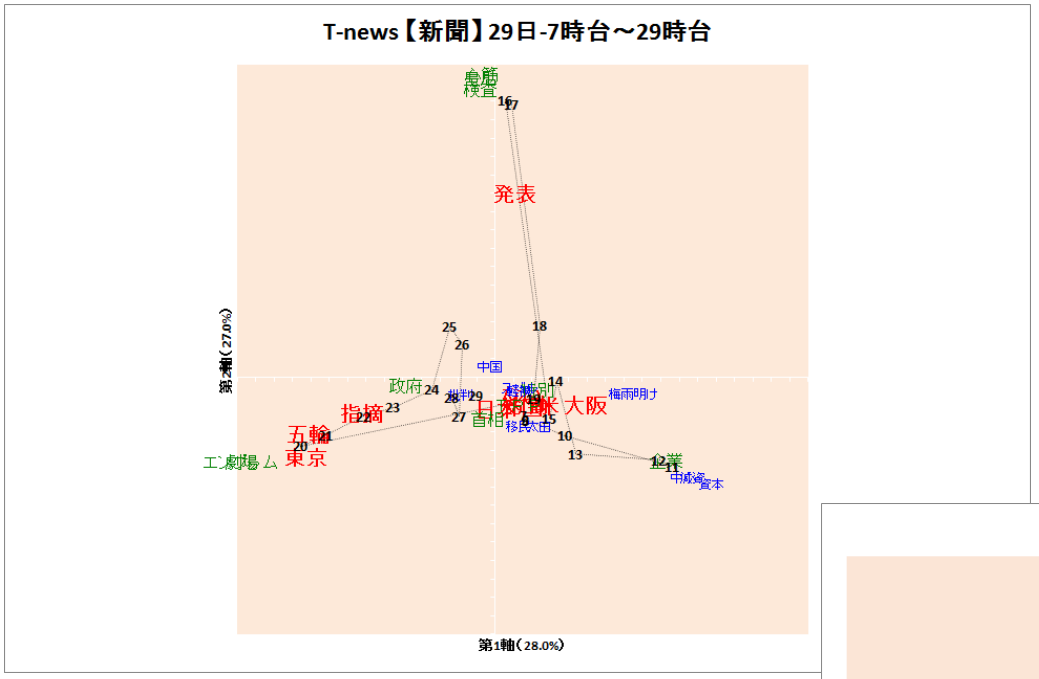

図15. 第1次元×第2次元の散布図

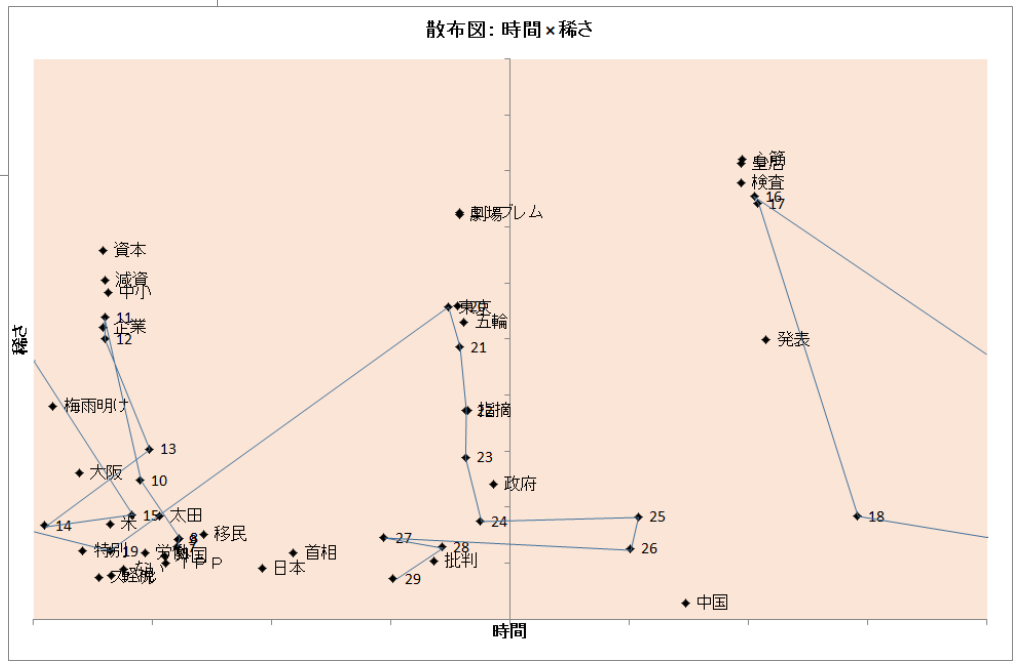

図16. 角度と距離の散布図

参考

- standard coordinate ／ principal coordinate
  - 特異ベクトルに固有値(特異値の平方根，固有値)を掛けてスコアを求めるのが一般的
    → principal coordinate
    (Rのca(グリネーカー)は standard coordinate を使っているので，固有値は掛かっていない)
- Project 311
  - 私の報告は下記参照.
    http://t-news.jp/311/index.html