時系列モデルとして式 (1) と式 (2) で表される ARMA-GARCH（Autoregressive Moving Average - Generalized Autoregressive Conditional Heteroscedasticity）モデルを用いる．

$$y_t = \sum_{i=1}^{p} a_i y_{t-i} + \sum_{i=1}^{q} b_i \varepsilon_{t-i} + c + \varepsilon_t \qquad \varepsilon_t \sim N(0, h_t) \quad i.i.d \tag{1}$$

$$h_t = \omega + \sum_{i=1}^{r} \alpha_i \varepsilon_{t-i}^2 + \sum_{i=1}^{s} \beta_i h_{t-i} \tag{2}$$

1　クラスタ数の決定

クラスタリングを行うにあたり，あらかじめクラスタ数を設定しておく必要がある．ここでは，様々あるクラスタリング指標のうち，PseudoF と金尻先輩がこれに改良を加えた二つの合わせて三つのうちどの指標を用いるのか，またその指標のもと定まる最適クラスタ数を求める．

クラスタリング手法は階層的クラスタリングを用い，距離関数をユークリッド距離，クラスタ融合手法をウォード法とした．

クラスタリングにおける要素集合 C の代表点には式 (3) で表されるメドイドを用いる．dist($\boldsymbol{x}$,$\boldsymbol{y}$) は要素 $\boldsymbol{x}$ と $\boldsymbol{y}$ との間のユークリッド距離であり，要素 $\boldsymbol{x} = [w_{x1}, w_{x2}, ..., w_{xn}]$ として，式 (4) で表される．

$$（メドイド）= \arg\min_{\boldsymbol{x} \in C} \sum_{\boldsymbol{y} \in C - \{\boldsymbol{x}\}} dist(\boldsymbol{x}, \boldsymbol{y}) \tag{3}$$

$$dist(\boldsymbol{x}, \boldsymbol{y}) = \sqrt{\sum_{i=1}^{n} (w_{xi} - w_{yi})^2} \tag{4}$$

PseudoF は式 (5) で表されるように重み付きクラスタ間普遍分散をクラスタ内分散で除したものとなっており，この値が大きいほどクラスタ間は疎でクラスタ内は蜜であることを意味する．一般的にクラスタ間は疎でクラスタ内は蜜となったクラスタリング結果は良いものとされるので，PseudoF 値を大きくするようなクラスタ数を設定する．

$$\text{PseudoF} = \frac{\frac{1}{k-1} \sum_{i=1}^{k} n_i \, dist(\boldsymbol{m_i}, \boldsymbol{m})^2}{\frac{1}{N-k} \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in C_i - \{\boldsymbol{m_i}\}} dist(\boldsymbol{x}, \boldsymbol{m_i})^2} \tag{5}$$

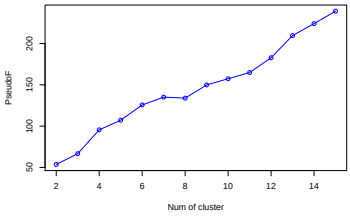
k はクラスタ数とし，各クラスタ集合を $C_1, C_2, ..., C_k$ と表す．n_i はクラスタ C_i に属する要素数であり，N は全要素数である．また，$\boldsymbol{m_i}$ はクラスタ C_i のメドイドであり，$\boldsymbol{m}$ は全要素に対するメドイドである．

しかし，式 (5) における分子部分のクラスタ間分散は各クラスタの代表点と全要素の代表点との間で定義されており，クラスタ間の距離は反映されていない．そこで，式 (5) におけるクラスタ間分散をクラスタ間の距離を用いて定めるように改良した式 (6) と式 (7) も用いる．また，式 (6) と式 (7) では，分母の小ささが全体の値の大きさに決定的な影響を与えないように分母部分に 1 を

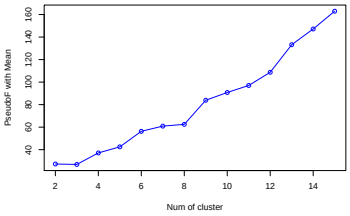
加算している．

$$\text{PseudoF with Mean} = \frac{\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{k} n_i \; dist(\boldsymbol{m_i}, \boldsymbol{m_j})^2}{1 + \sum_{i=1}^{k}\sum_{\boldsymbol{x}\in C_i - \{\boldsymbol{m_i}\}} dist(\boldsymbol{x}, \boldsymbol{m_i})^2} \tag{6}$$

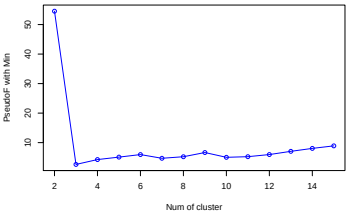
$$\text{PseudoF with Min} = \frac{\sum_{i=1}^{k} n_i \; \min\{dist(\boldsymbol{m_i}, \boldsymbol{m_j}), j \neq i\}^2}{1 + \sum_{i=1}^{k}\sum_{\boldsymbol{x}\in C_i - \{\boldsymbol{m_i}\}} dist(\boldsymbol{x}, \boldsymbol{m_i})^2} \tag{7}$$



(a) PseudoF



(b) PseudoF with Mean



(c) PseudoF with Min

図 1: 実測値における最適クラスタ数指標

表 1: 実測値における最適クラスタ数

指標	最適クラスタ数
PseudoF	15
PseudoF with Mean	15
PseudoF with Min	2

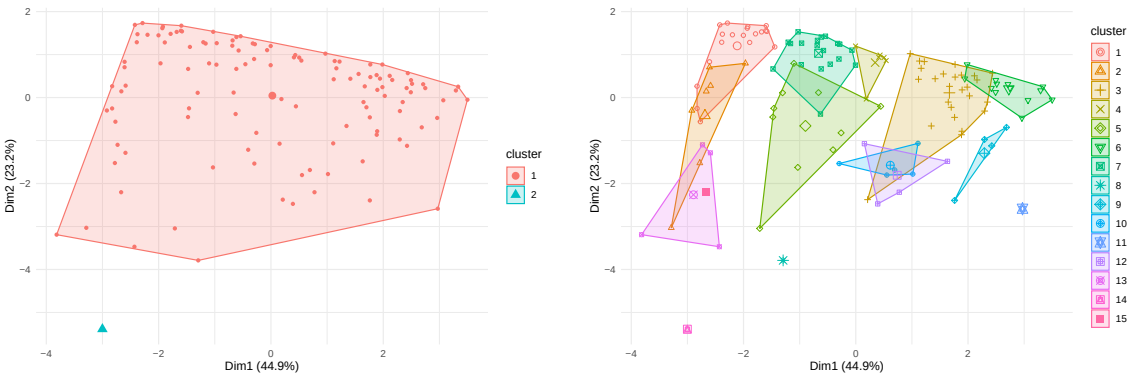
1.1 実測値に対して

実測値に対してこれら三つの指標を用いた結果は図 1 となった．縦軸に指標の値をとり，横軸をクラスタ数とした．表 1 は各指標から定まる最適クラスタ数をまとめたものであり，図 2 はそれらの最適クラスタ数でクラスタリングを行ったときの主成分散布図である．

図 1(a)，図 1(b) の指標値は単調に増加しており，クラスタ数が 2 から 15 までの範囲では最適クラスタ数は 15 となった．図 1(c) の指標値はクラスタ数 2 の場合に突出して大きくなっており，最適クラスタ数は 2 となった．

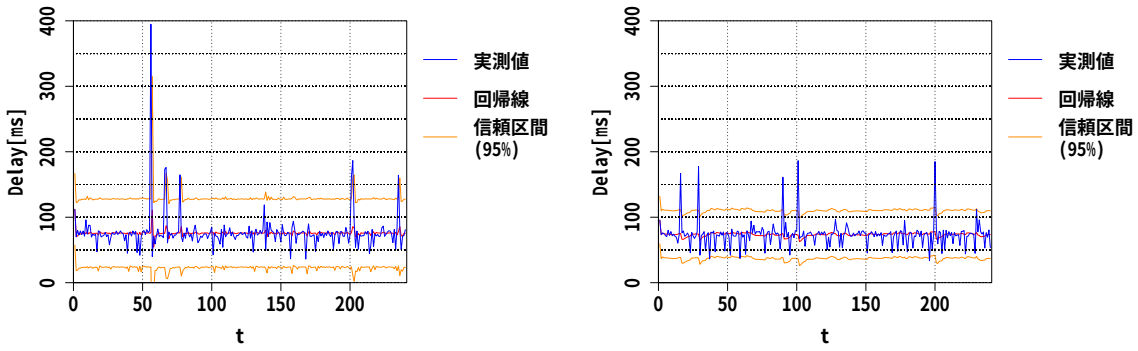
このような結果となるのは，クラスタ数 2 の時に形成される二つのクラスタ間は大きく離れているのだが，クラスタ数を増やすにつれてこれらのクラスタから分離して形成される新たなクラスタは，全要素に対するメドイドや分離前に属していなかったクラスタからは離れているものの，分離前に属していたクラスタからは近いためだと思われる．したがって，クラスタ数 3 以上においてはクラスタの細分化が起こっていると考えられ，必要以上の細分化は一つのクラスタ内で完結した共通の特徴を損なわせるため生じないほうが良いという観点では，指標 PseudoF with Min が良いように思われる．

図 2(a) より，クラスタ数 2 において形成される二つのクラスタのうち一方は要素数が 1 でもう一方のクラスタからは大きく離れた分布に位置していた．この要素は 3/15（日）7:00-8:00 における計測データであった．この計測データに対する ARMA-GARCH(2,2,1,1) による回帰結果は図 3 となっており，その回帰におけるパラメータは表 2 となっていた．比較のため 3/22（日）7:00-8:00 の計測データに対するものを載せている．図 3 と表 2 から，3/15（日）7:00-8:00 における回帰では，約 400ms のスパイク的な応答遅延が発生しておりこれは他のスパイク的な応答遅延が約 200ms であることと比べても大きな値である．これにより，ノイズ項が従う正規分布の分散の回帰における定数項 ω の値が異常に大きくなりすぎていることが読み取れる．この大きな 400ms 程度の遅延はほとんど起こらないものの異常ではないと考えられるため，回帰を行うにあたり，スパイク的な応答遅延を除くもしくは何らかの上限値を設けて置換するなどが必要かもしれない．も



(a) クラスタ数 2　　(b) クラスタ数 15

図 2: 実測値におけるクラスタリング結果の主成分散布図



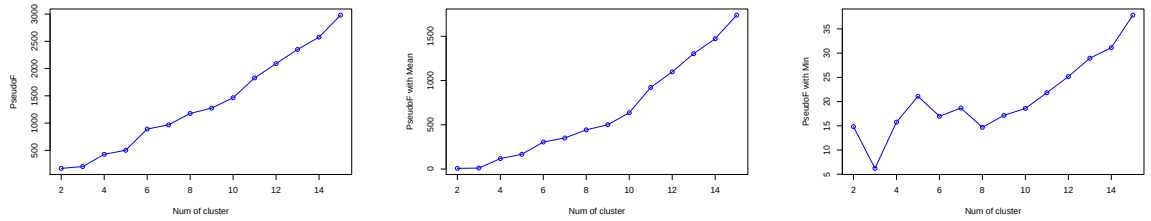
(a) 3/15（日）7:00-8:00　　(b) 3/22（日）7:00-8:00

図 3: 実測値に対する ARMA-GARCH(2,2,1,1) での回帰結果

しくは，標準化を行うなどして，ある一つのパラメータの大きさがクラスタリングにおける距離関数にて決定的な影響を与えないようにする必要があるだろう．

表 2: 回帰結果におけるパラメータ

	3/15（日）7:00-8:00	3/22（日）7:00-8:00
c	37.37	26.86
a_1	0.5457	0.5265
a_2	-0.0392	0.1082
b_1	-0.4351	-0.6159
b_2	-0.0635	-0.1387
ω	703.18	55.78
α_1	0.0995	$\ll 0.0001$
β_1	$\ll 0.0001$	0.8390



(a) PseudoF (b) PseudoF with Mean (c) PseudoF with Min

図 4: 変動値における最適クラスタ数指標

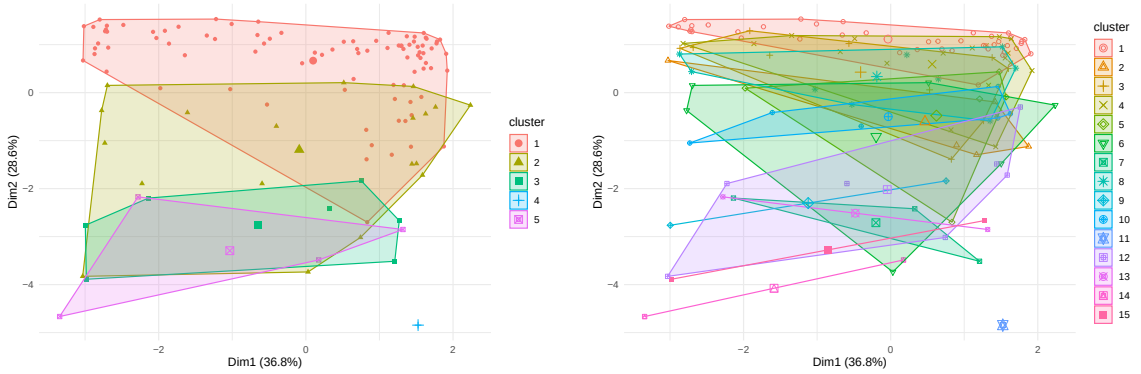
1.2 変動値に対して

変動値に対してこれら三つの指標を用いた結果は図 4 となった．表 3 は各指標から定まる最適クラスタ数をまとめたものであり，図 5 はそれらの最適クラスタ数でクラスタリングを行ったときの主成分散布図である．

変動値におけるクラスタリングにおいて，図 4(a) と図 4(b) は実測値の場合と同様に単調増加していたが， 4(c) はクラスタ数 5 にて極大値を取っていた．主成分分析はより少ない次元数でデータを特徴づけるものであるため，適切なクラスタ数においては主成分散布図上においてもクラスタ間に重なりがなく，クラスタ内の要素は固まっていることが望ましい．図 5(a) からその傾向を強く見受けることができないが，赤色で表されるクラスタ 1 に属する要素のうち第二主成分が大きなものなどでは他のクラスタと重なりがないため，ある程度は適切なクラスタ数であると考えられる．

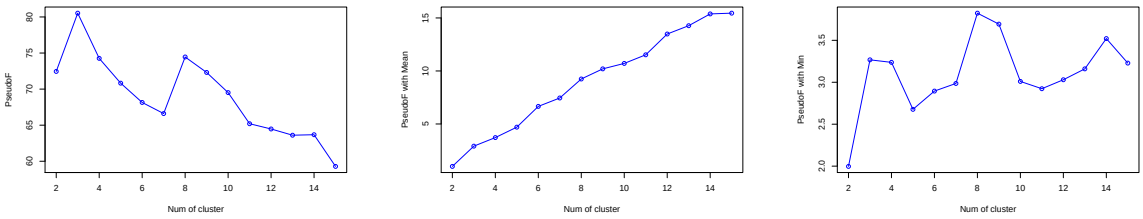
表 3: 変動値における最適クラスタ数

指標	最適クラスタ数
PseudoF	15
PseudoF with Mean	15
PseudoF with Min	5,15



(a) クラスタ数 5　　(b) クラスタ数 15

図 5: 変動値におけるクラスタリング結果の主成分散布図



(a) PseudoF　　(b) PseudoF with Mean　　(c) PseudoF with Min

図 6: 実測値の主成分における最適クラスタ数指標

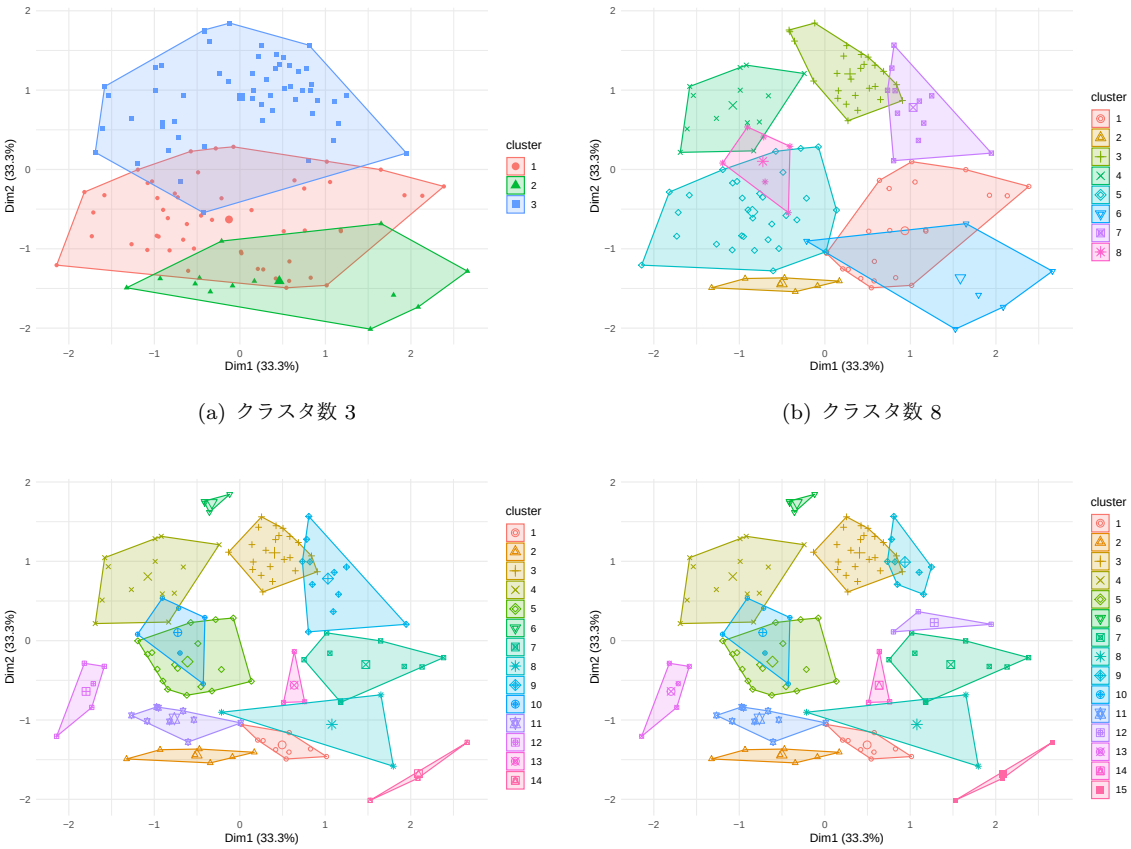
1.3　実測値の主成分に対して

実測値の主成分に対してこれら三つの指標を用いた結果は図 6 となった．表 4 は各指標から定まる最適クラスタ数をまとめたものであり，図 7 はそれらの最適クラスタ数でクラスタリングを行ったときの主成分散布図である．主成分を用いて行ったクラスタリングにおける主成分散布図の第一主成分・第二主成分は，クラスタリングに用いた第一主成分・第二主成分と一致する．

図 6(b) は変わらず単調増加しているものの，図 6(a) においては，クラスタ数 3 と 8 において極大値をとっていた．したがって，ARMA-GARCH モデルの回帰結果から得られるパラメータを標準化しさらに主成分分析を行った結果得られる主成分を用いることで，クラスタ数が多すぎない範囲で，全データ集合の代表点から各クラスタが散らばった良いクラスタリング結果が得られるよ

表 4: 実測値の主成分における最適クラスタ数

指標	最適クラスタ数
PseudoF	3,8
PseudoF with Mean	15
PseudoF with Min	3,8,14



(a) クラスタ数 3

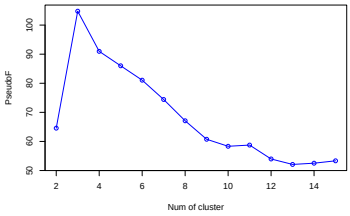
(b) クラスタ数 8

(c) クラスタ数 14

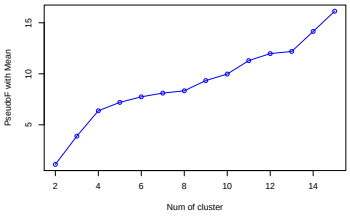
(d) クラスタ数 15

図 7: 実測値の主成分におけるクラスタリング結果の主成分散布図

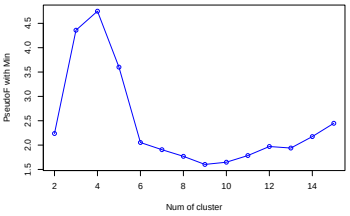
うだ．さらに，図 6(c) が極大値をとるクラスタ数は図 6(a) が極大値をとるクラスタ数と一致している．このことから，主成分を用いたクラスタリングにおいては，指標 PseudoF with Min は，指標 PseudoF を兼ねているようだ．



(a) PseudoF



(b) PseudoF with Mean



(c) PseudoF with Min

図 8: 変動値の主成分における最適クラスタ数指標

表 5: 変動値の主成分における最適クラスタ数

指標	最適クラスタ数
PseudoF	3
PseudoF with Mean	15
PseudoF with Min	4

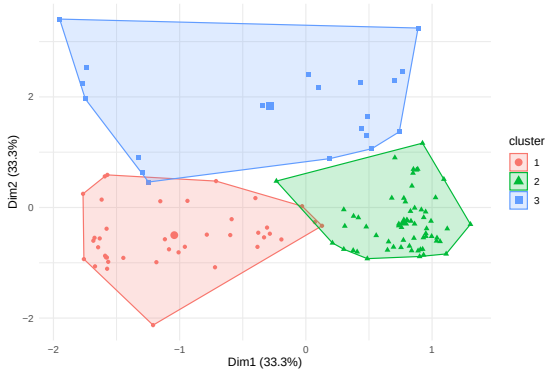
1.4 変動値の主成分に対して

変動値の主成分に対してこれら三つの指標を用いた結果は図 8 となった．表 5 は各指標から定まる最適クラスタ数をまとめたものであり，図 9 はそれらの最適クラスタ数でクラスタリングを行ったときの主成分散布図である．

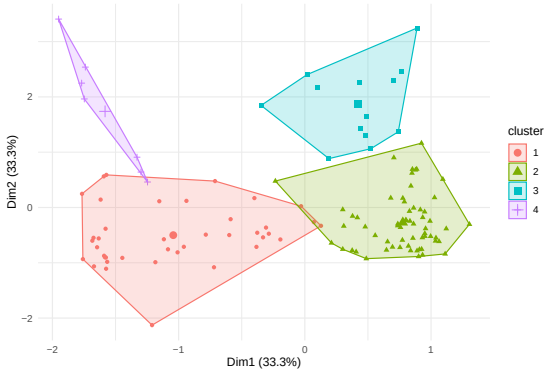
実測値の主成分に対してのものと同様に，図 8(b) は単調増加しており，図 8(a) と図 8(c) はほぼ同じクラスタ数の時に極大値をとっていた．したがって，指標としては PseudoF か PseudoF with Min が適切と考える．

1.5 結論

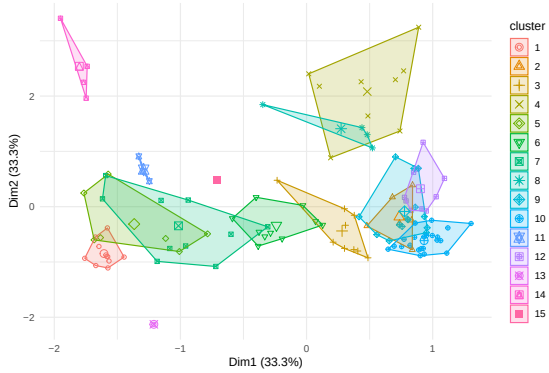
クラスタリングは，標準化後の主成分を用いて行う方が良いだろう．また，クラスタ数を定める指標としては PseudoF か PseudoF with Min が良いだろう．ただ，PseudoF with Min では，変動値においても極大値を見いだせており，PseudoF のように値が単調に増加し最適クラスタ数が見つからないことがなかったため，PseudoF with Min がより適切だと考える．



(a) クラスタ数 3



(b) クラスタ数 4



(c) クラスタ数 15

図 9: 変動値の主成分におけるクラスタリング結果の主成分散布図

2　クラスタリング結果

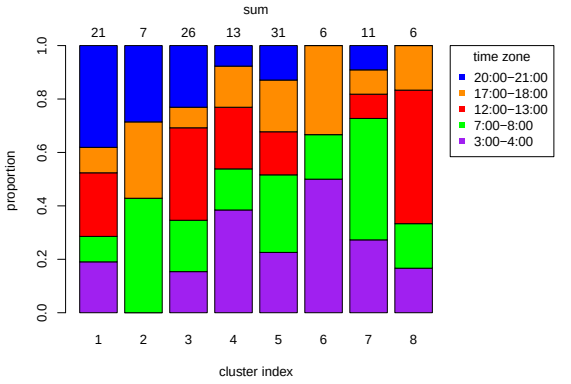
クラスタリング結果を，(a) 横軸をクラスタ番号とし計測した時間帯ごとに色分けした積み上げ棒グラフ，(b) 横軸をクラスタ番号とし計測した曜日ごとに色分けした積み上げ棒グラフ，(c) 横軸を計測した曜日と時間帯とし属するクラスタ番号ごとに色分けした積み上げ棒グラフ，を示す．縦軸は全て割合としており，各棒の上部にその要素数を示している．

2.1　実測値の主成分に対して

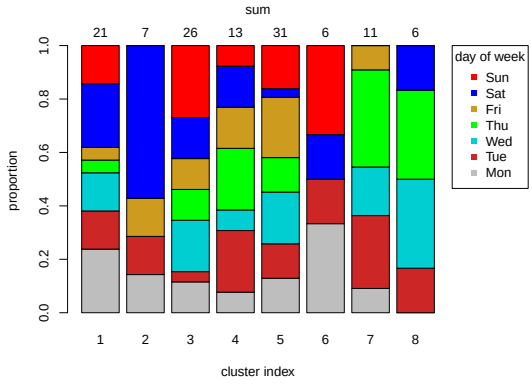
1.3 章よりクラスタ数を 8 とした場合を図 10 に示す．

図 10(a) のクラスタ番号 1 には 20:00 - 21:00 に計測したデータが多く属しており，図 10(c) より全ての曜日の 20:00 - 21:00 の計測データのうち少なくとも一つはクラスタ番号 1 に属していたため，このクラスタには 20:00 - 21:00 の計測データが属しやすいといえそうだ．ただ，他のクラスタ番号において全曜日の同一時間帯に計測されたデータが属するクラスタは見受けられなかった．

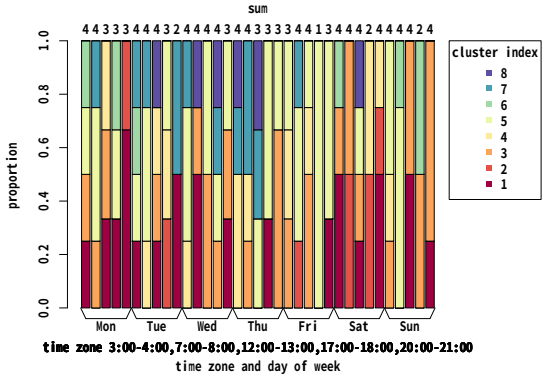
図 10(a) のクラスタ番号 1 と 3 は，利用者が多いと考えられる 12:00 - 13:00，17:00-18:00，20:00-21:00 の計測データの割合が多い．さらに，図 10(b) のクラスタ番号 1 と 3 は，利用者が



(a) 時間帯ごとに属するクラスタ



(b) 曜日ごとに属するクラスタ



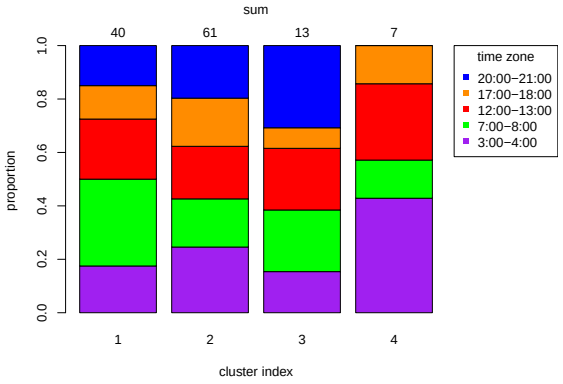
(c) 各時間帯と曜日が属するクラスタ

図 10: 実測値の主成分をもとにクラスタ数 8 で行ったクラスタリング結果

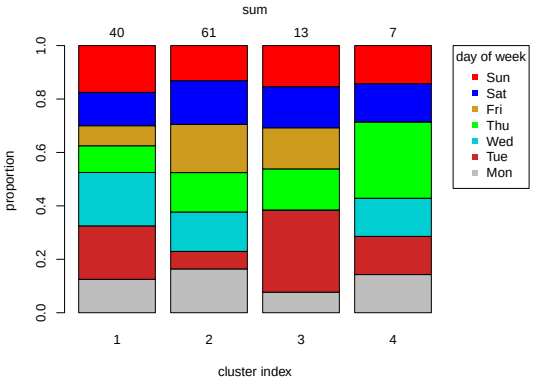
多いと考えられる土曜日と日曜日の割合が多い．これは図 10(c) からも読み取れる．したがって，クラスタ番号 1 と 3 には，利用者が多い状況で計測されたデータが属しやすい傾向があるのではないか．

逆に図 10(a) のクラスタ番号 7 は，利用者が少ないと考えられる 3:00 - 4:00，7:00 - 8:00 の計測データが多く，図 10(b) のクラスタ番号 7 には利用者が多いと考えられる土曜日と日曜日に計測されたデータは属していなかった．したがって，クラスタ番号 7 には，利用者が多くない状況で計測されたデータが属しやすい傾向があるのではないか．

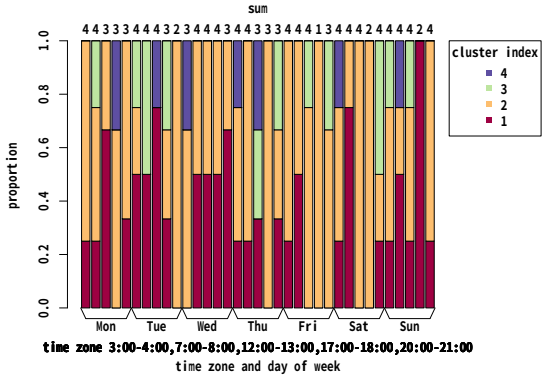
要素数が少ないクラスタであるクラスタ番号 2 と 6 と 8 のそれぞれにおいて，同一時間帯や同一曜日，または同一の時間帯と曜日のデータが，この他のデータ集合から外れたクラスタに属する傾向があるかを調べたが，この結果からはそのような傾向は見受けられなかった．



(a) 時間帯ごとに属するクラスタ



(b) 曜日ごとに属するクラスタ



(c) 各時間帯と曜日が属するクラスタ

図 11: 変動値の主成分をもとにクラスタ数 4 で行ったクラスタリング結果

2.2 変動値の主成分に対して

1.4 章よりクラスタ数を 4 とした場合を図 11 に示す.

変動値の主成分を用いたクラスタリングでは，最適クラスタ数が 4 と少なく，さらに多くのデータが図 11(a) のクラスタ番号 1 か 2 に属していた．図 11(c) からはどの曜日のどの時間帯においてもまんべんなくクラスタ番号 1 か 2 に属していることが見て取れる．またこれらのクラスタから外れた要素数の少ないクラスタのクラスタ番号 3 か 4 に属するデータには，計測時間帯や曜日に共通性は見て取れない．したがって，変動値はおおよそクラスタ番号 1 か 2 で代表され，そこから外れる場合もあるがそこに共通性はなさそうだ．よって，変動値の主成分を用いたクラスタリングを用いた異常検知方法としては，現場の無線機器で一定の時間幅の取得データをもとにクラスタリングを行い，要素数の多いクラスタから外れ続けた場合に異常とみなすことなどが考えられる．

3　まとめと今後の課題

実測値の主成分を用いたクラスタリング結果においては，曜日や時間帯に応じた傾向が存在する可能性が見受けられた．そのため障害検知手法としては，例えば，前もって行ったクラスタリング結果において形成された曜日や時間帯ごとの傾向を持つクラスタの代表点を，その曜日や時間帯における ARMA-GARCH モデルのパラメータのテンプレートとし，それを用いて行った応答遅延の予測値から実測値が大きく外れ続けた場合に異常を検知することが考えられる．または，現場の無線機器で逐次クラスタリングを行った結果が，その計測時における曜日や時間帯の傾向を持つ代表点のクラスタに属さなかった場合に異常を検知することも考えられる．しかし，今回得られたクラスタリング結果においては，曜日や時間帯に応じた傾向を顕著に見て取ることができなかった．これには様々な原因が考えられるがそのうちの一つとして，単発的に発生する応答遅延によりモデルの回帰精度が悪くなっていることが考えられる．また，時系列モデルの回帰において，スパイク的な応答遅延を予測することができないため，異常に大きな遅延が発生したり，頻繁に大きな遅延が発生したりする異常は検知できないと思われる．そのため，実用的な異常検知には，単発的な応答遅延の発生頻度やそのスパイク性を表す別の手法と組み合わせることが必要であろう．