成長パターンのクラスタリングによる林木成長予測

筑波大学　社会工学系

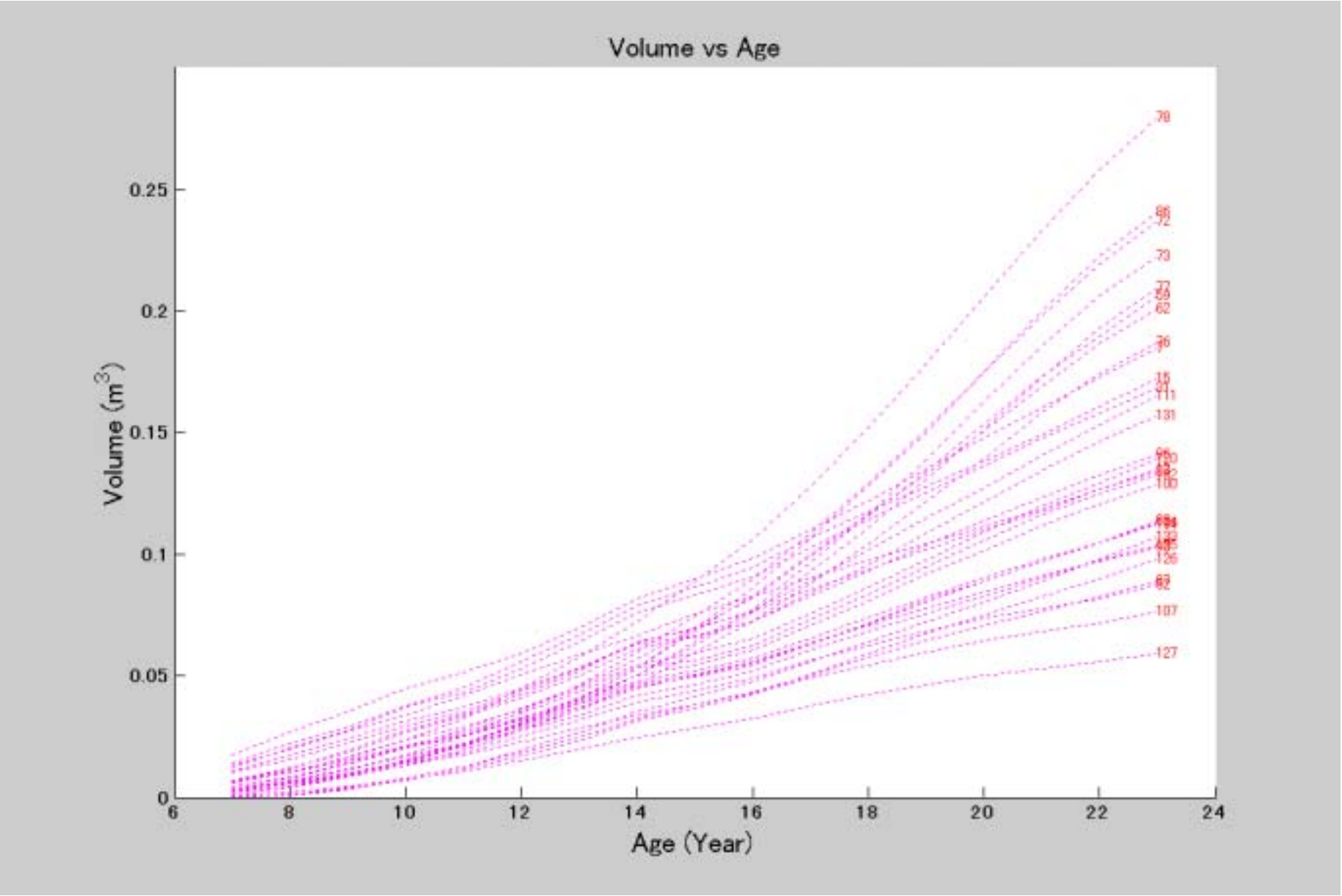
柳原 宏和

東北大学大学院　環境科学研究科

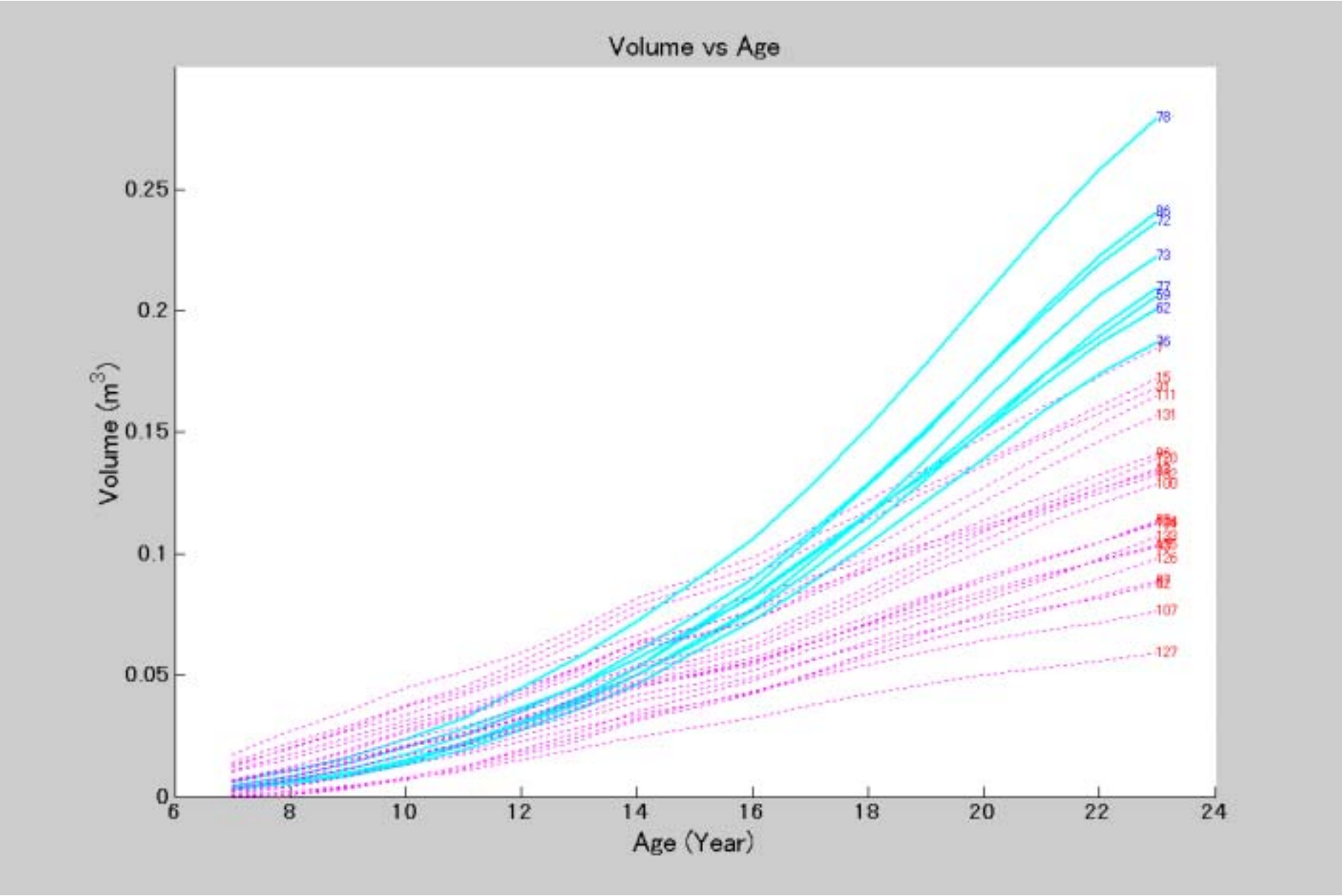
吉本 敦

2004, 3/13 - 14, FORMATH NAGOYA 2004 in 名古屋

成長データ(その1)


Volume vs Age
Volume (m^3)
0
0.05
0.1
0.15
0.2
0.25
6
8
10
12
14
16
18
20
22
24
Age (Year)

成長データ(その2)


Volume vs Age
Volume (m^3)
0
0.05
0.1
0.15
0.2
0.25
6
8
10
12
14
16
18
20
22
24
Age (Year)
78
73
62
131
107
127

問題点

成長パターンが複数ある場合, すべて同じ成長パターンで解析をすると推定結果にバアイスが生じる！

伐採木からのデータから成長パターンを分類.

- ***k*-平均法等で分類できるのは伐採木だけである.**

現在わかる情報から非伐採木の成長パターンも分類し成長予測をできないか？

PAIC (Satoh, 1997) **を用いた外挿でのモデル評価により非伐採木の成長パターンを分類.**

非伐採木の成長予測

伐採木の成長予測 :
個々の林木の成長データに成長曲線をあてはめ予測.

我々が知りたいのは伐採していない林木の成長である!

非伐採木の成長予測 :
林分内の林木の成長の成長曲線のパラメータに DBH 等の現在わかっている情報を共変量に持つ構造を仮定 ($\xi_i = \Theta' \boldsymbol{x}_i$)

- Θ を伐採木で推定, その後非伐採木の共変量で成長曲線のパラメータを推定.
- Yanagihara and Yoshimoto (2004) ではランダム効果を導入したモデルで推定 ($\xi_i = \Theta' \boldsymbol{x}_i + \beta_i$).

目的

- Sample Plot **に複数の成長パターンが存在する場合の成長予測**.
  1. **個々の林木の成長曲線の係数に関する** *k*-**平均法** (MacQueen, 1967) **によるクラスタリング**.
  2. **クラスタリングを考慮に入れた, 個々の成長曲線のパラメータの推定値を応答変数に**, DBH**等を共変量にもつ多変量線形モデルによる係数の推定**.
  3. PAIC (Satoh, 1997) **を用いたモデル評価による非伐採木のクラスタリング**.

  - **成長曲線には** Richards (1958) **の成長関数を用いた**.

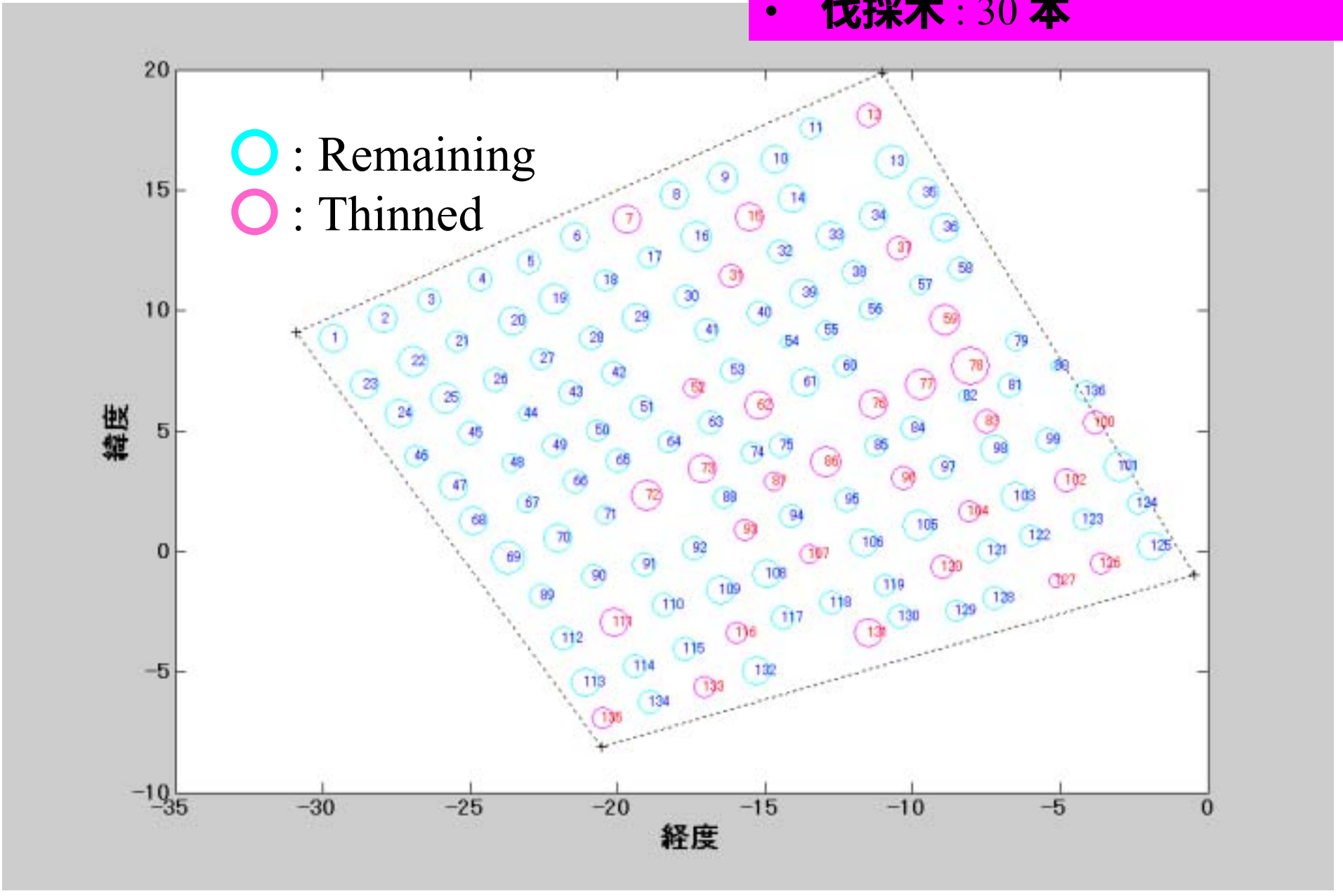
Outline

1. データとSample Plot
2. *k*-平均法を用いた伐採木の成長パターンのクラスタリング.
3. クラスターを考慮に入れた多変量線形モデルとその評価.
4. モデルの外挿とPAIC を用いた非伐採木の成長パターンのクラスタリング.
5. まとめとこれからの課題.

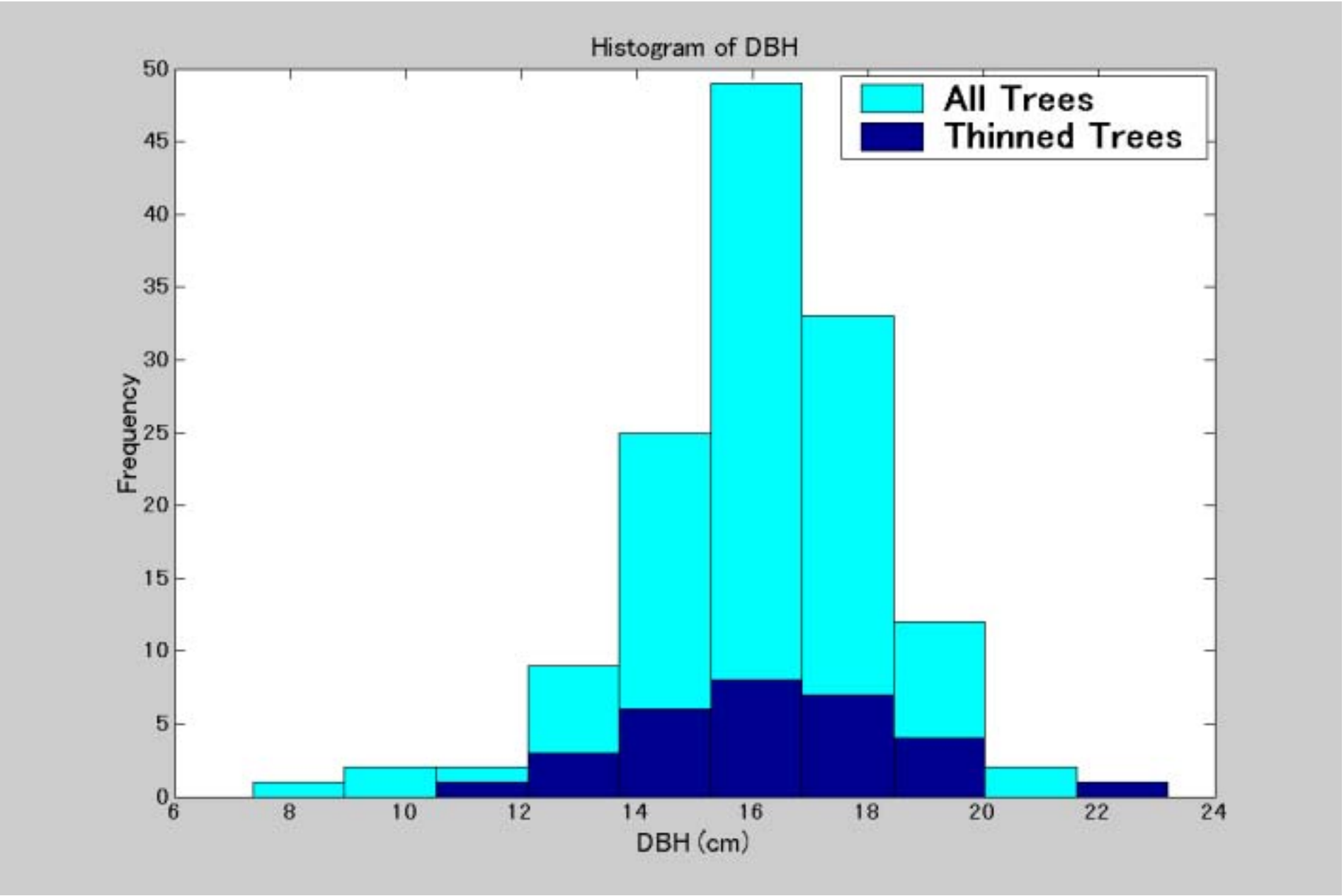
１．データとSample Plot

Sample Plot (その1)

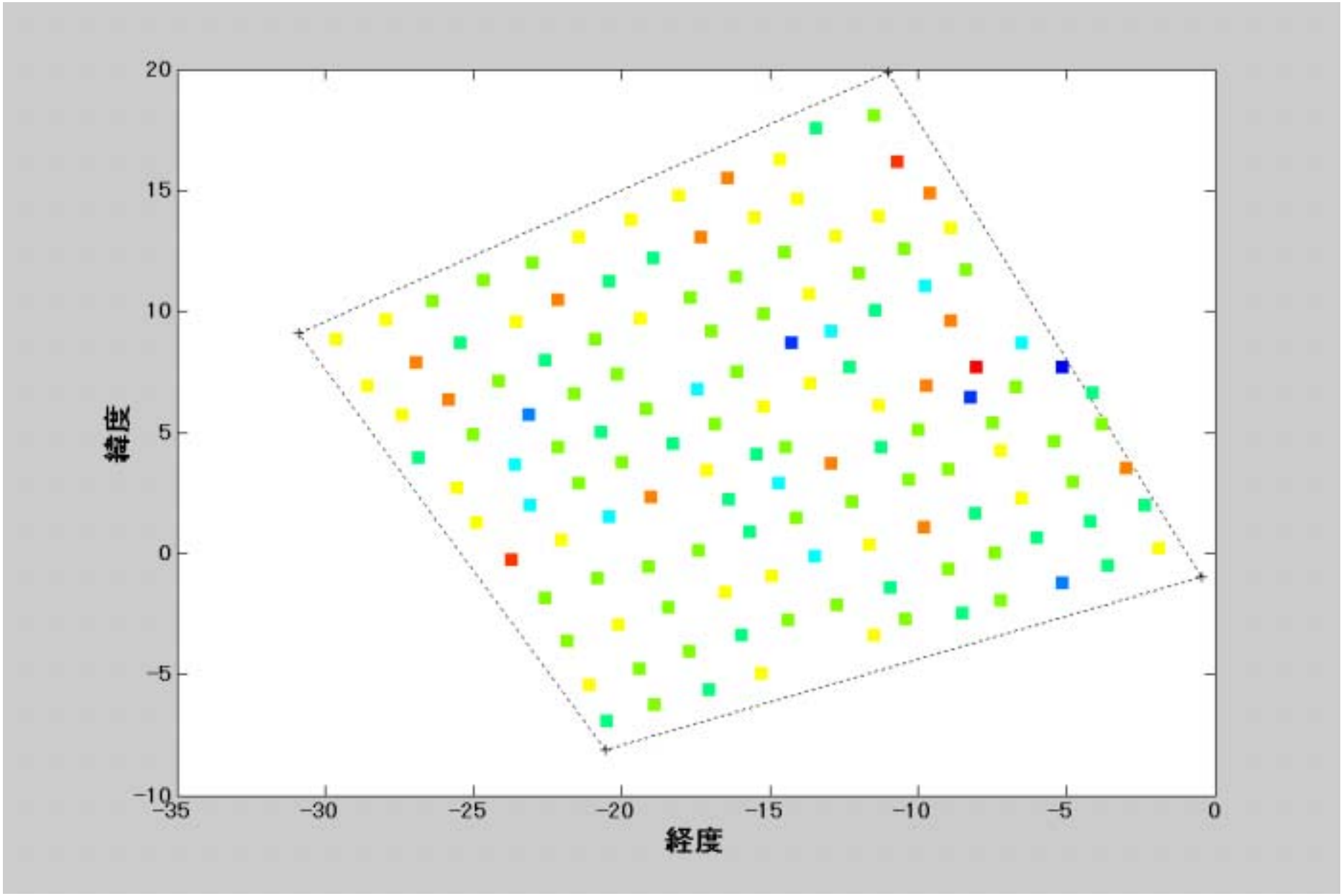
- 試験林 :福岡県八女群星野村.
- 林齢 : 23 年
- 調査区域 : 計 136 本 (0.0466 *ha*)
- 伐採木 : 30 本



直径分布


Histogram of DBH
All Trees
Thinned Trees
Frequency
DBH (cm)
0
5
10
15
20
25
30
35
40
45
50
6
8
10
12
14
16
18
20
22
24

Sample Plot (その2)



応答変数の計算

1. それぞれの材積の成長データに, 個別に成長曲線を当てはめる.

   - あてはめる成長曲線は Richards (1958) の成長曲線. ただしパラメータ空間を $(0, \infty]$ から $[-\infty, \infty]$ に変換する (Yanagihara and Yoshimoto, 2004, 柳原, 吉本, 能本, 2003, 参照).

$$f(t, \boldsymbol{\theta}) = e^{\mu} \left\{1 - \exp\left(-e^{\alpha} t\right)\right\}^{\exp(\beta)}.$$

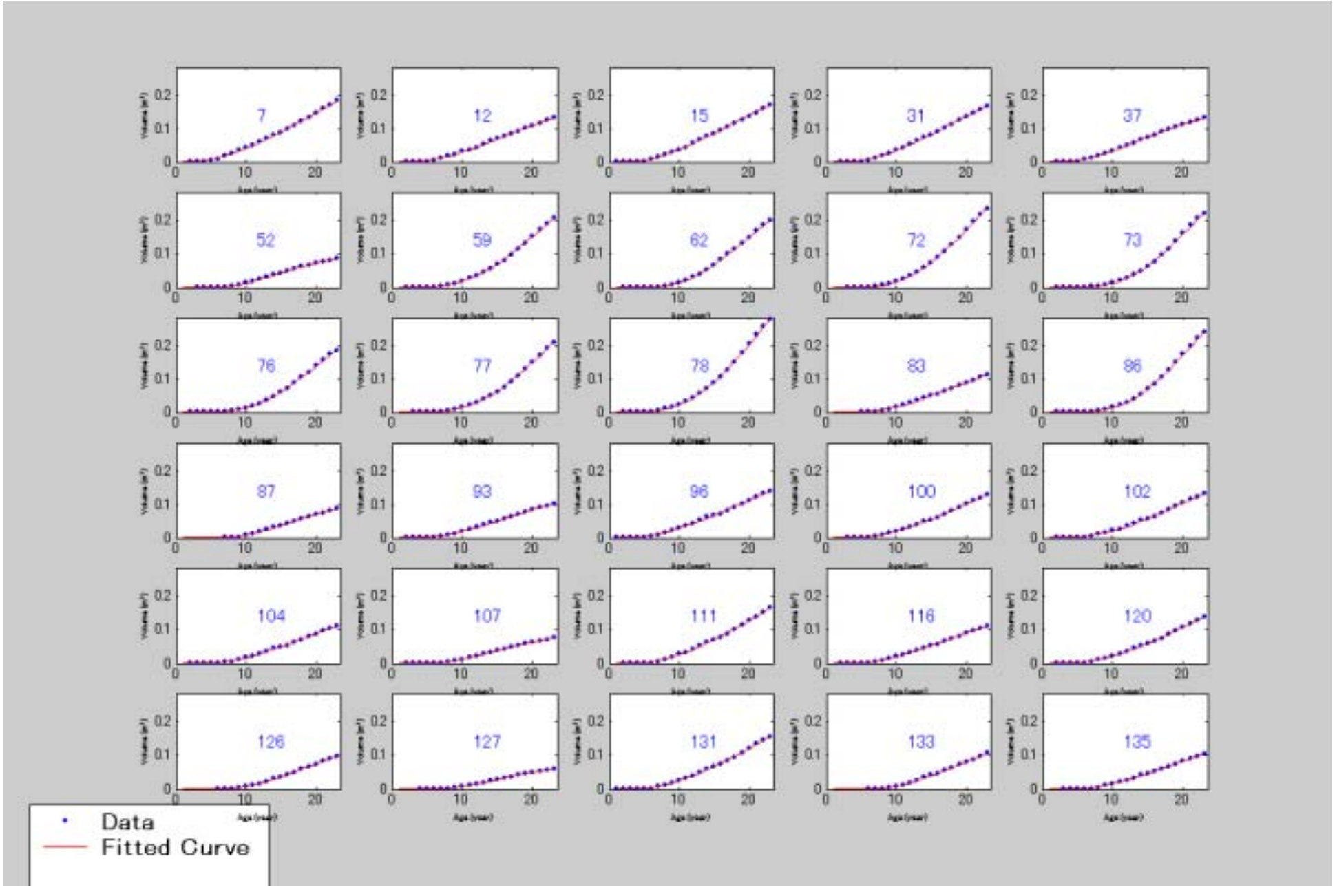
   - パラメータの推定には最小二乗法を使い, 残差平方和を最小にするパラメータ探索には SPIDER アルゴリズム (Ohtaki and Izumi, 1999) を用いた. 多変量 SPIDER アルゴリズムについては Yanagihara and Yoshimoto (2004), 柳原, 吉本, 能本 (2003) 参照.

2. i 番目の林木に関して得られた推定量を $\hat{\mu}_i$, $\hat{\alpha}_i$, $\hat{\beta}_i$ とする. このとき応答変数を

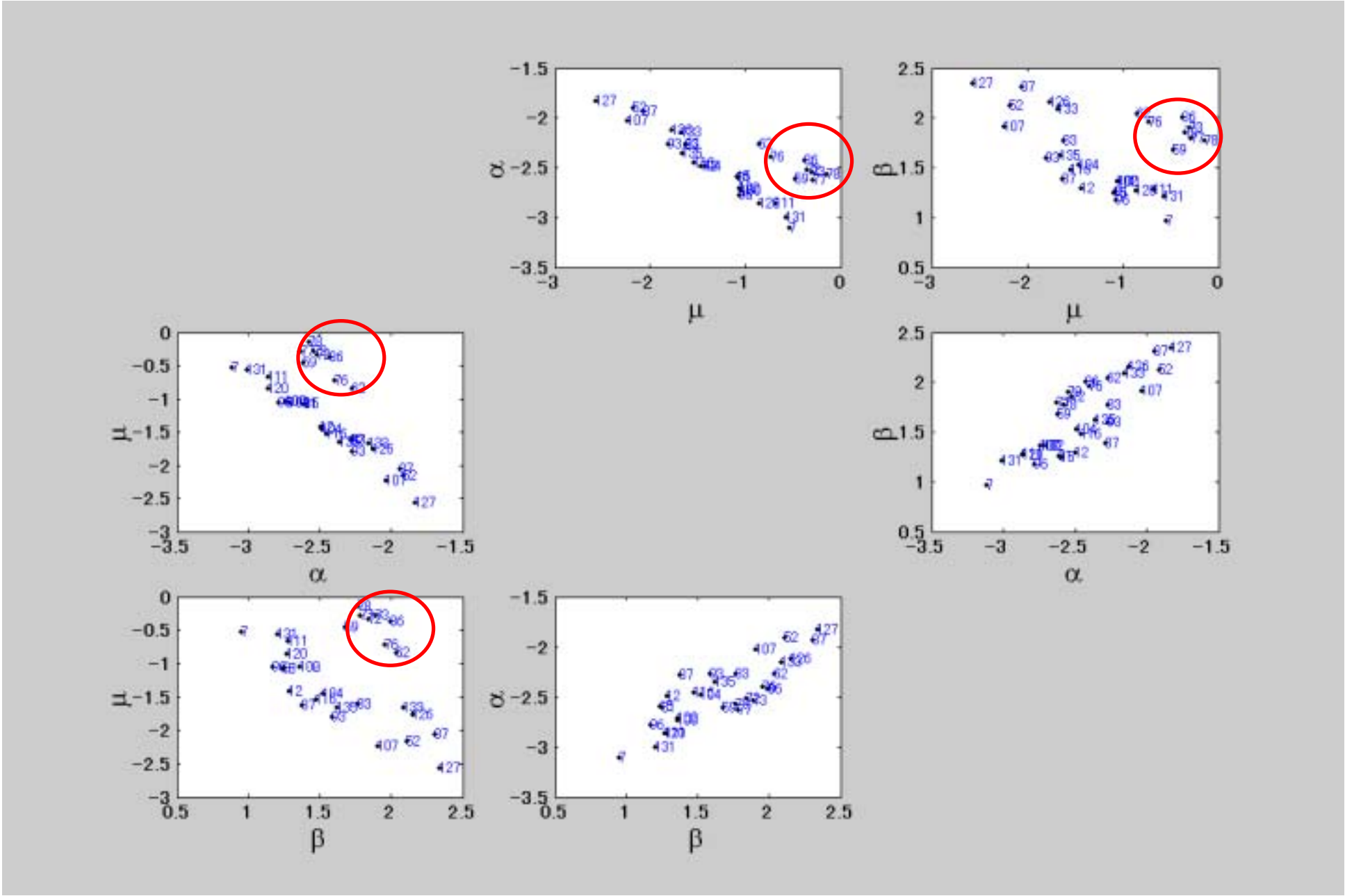
$$\boldsymbol{y}_i = (y_{i1}, y_{i2}, y_{i3})' = (\hat{\mu}_i, \hat{\alpha}_i, \hat{\beta}_i)',$$

   として, それらを並べた行列 $Y = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)'$ を観測値行列とする.

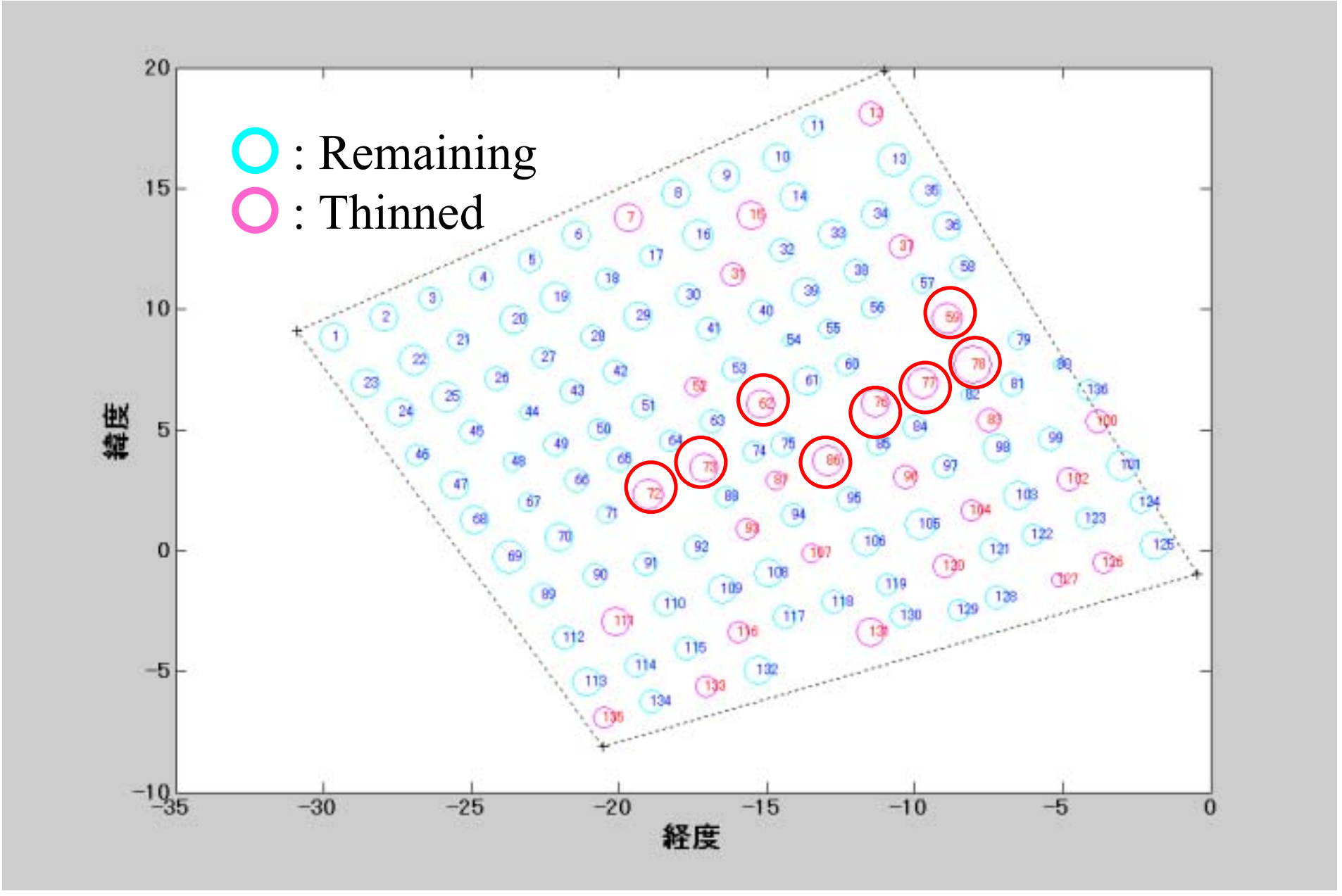
当てはめ結果


7
12
15
31
37
52
59
62
72
73
76
77
78
83
86
87
93
96
100
102
104
107
111
116
120
126
127
131
133
135
Data
Fitted Curve

係数の散布図



立木位置と係数の関係



２．k-平均法を用いた伐採木の成長パターンのクラスタリング

k-平均法

(MacQueen, 1967)

クラスター内の変動を小さくなる分割を探す

分割 G のクラスター	C_1	$\cdots$	C_g
観測値の個数	n_1	$\cdots$	n_g
重心	$\bar{\boldsymbol{y}}_1$	$\cdots$	$\bar{\boldsymbol{y}}_g$
共分散	S_1	$\cdots$	S_g

$$\bar{\boldsymbol{y}}_k = \frac{1}{n_k}\sum_{i\in C_k}\boldsymbol{y}_i,$$

$$S_k = \frac{1}{n_k}\sum_{i\in C_k}(\boldsymbol{y}_i-\bar{\boldsymbol{y}}_k)(\boldsymbol{y}_i-\bar{\boldsymbol{y}}_k)'.$$

クラスター内平方和積和行列：$W(G)=\sum_{r=1}^{g} n_r S_r.$

分割 G 内のクラスター C_k に属する観測値 $\boldsymbol{y}_i$ をクラスター C_t に移動させる.

分割 G^* のクラスター	C_1	$\cdots$	C_k	$\cdots$	C_t	$\cdots$	C_g
観測値の個数	n_1	$\cdots$	n_k-1	$\cdots$	n_t+1	$\cdots$	n_g
重心	$\bar{\boldsymbol{y}}_1$	$\cdots$	$\bar{\boldsymbol{y}}_k^*$	$\cdots$	$\bar{\boldsymbol{y}}_t^*$	$\cdots$	$\bar{\boldsymbol{y}}_g$
共分散	S_1	$\cdots$	S_k^*	$\cdots$	S_t^*	$\cdots$	S_g

クラスター内平方和積和行列：$W(G^*)=\sum_{r=1}^{g} n_r S_r - n_k S_k - n_k S_t + (n_k-1)S_k^* + (n_t+1)S_t^*.$

$$\left(W(G^*)=W(G)-\frac{n_k}{n_k-1}(\boldsymbol{y}_i-\bar{\boldsymbol{y}}_k)(\boldsymbol{y}_i-\bar{\boldsymbol{y}}_k)'+\frac{n_t}{n_t+1}(\boldsymbol{y}_i-\bar{\boldsymbol{y}}_t)(\boldsymbol{y}_i-\bar{\boldsymbol{y}}_t)'\right).$$

tr($W(G^*)$)≦tr($W(G)$) (または $|W(G^*)|\leqq|W(G)|$) であればクラスターを更新する.

アルゴリズム

Step 1. 初期分割 $G, C_1, \ldots, C_g$ を決定する.

Step 2. 分割 G 内のクラスター C_k に属する観測値 $\boldsymbol{y}_i$ をクラスター C_t に移動させ, $\mathrm{tr}(W(G^*)) < \mathrm{tr}(W(G))$, または $|W(G^*)| < W(G)$ であれば分割を更新する. この不等式は

$$\mathrm{tr}(W(G^*)) < \mathrm{tr}(W(G)) \quad \Leftrightarrow \quad \frac{n_t}{n_t+1}||\boldsymbol{y}_i - \bar{\boldsymbol{y}}_t||^2 < \frac{n_k}{n_k-1}||\boldsymbol{y}_i - \bar{\boldsymbol{y}}_k||^2,$$

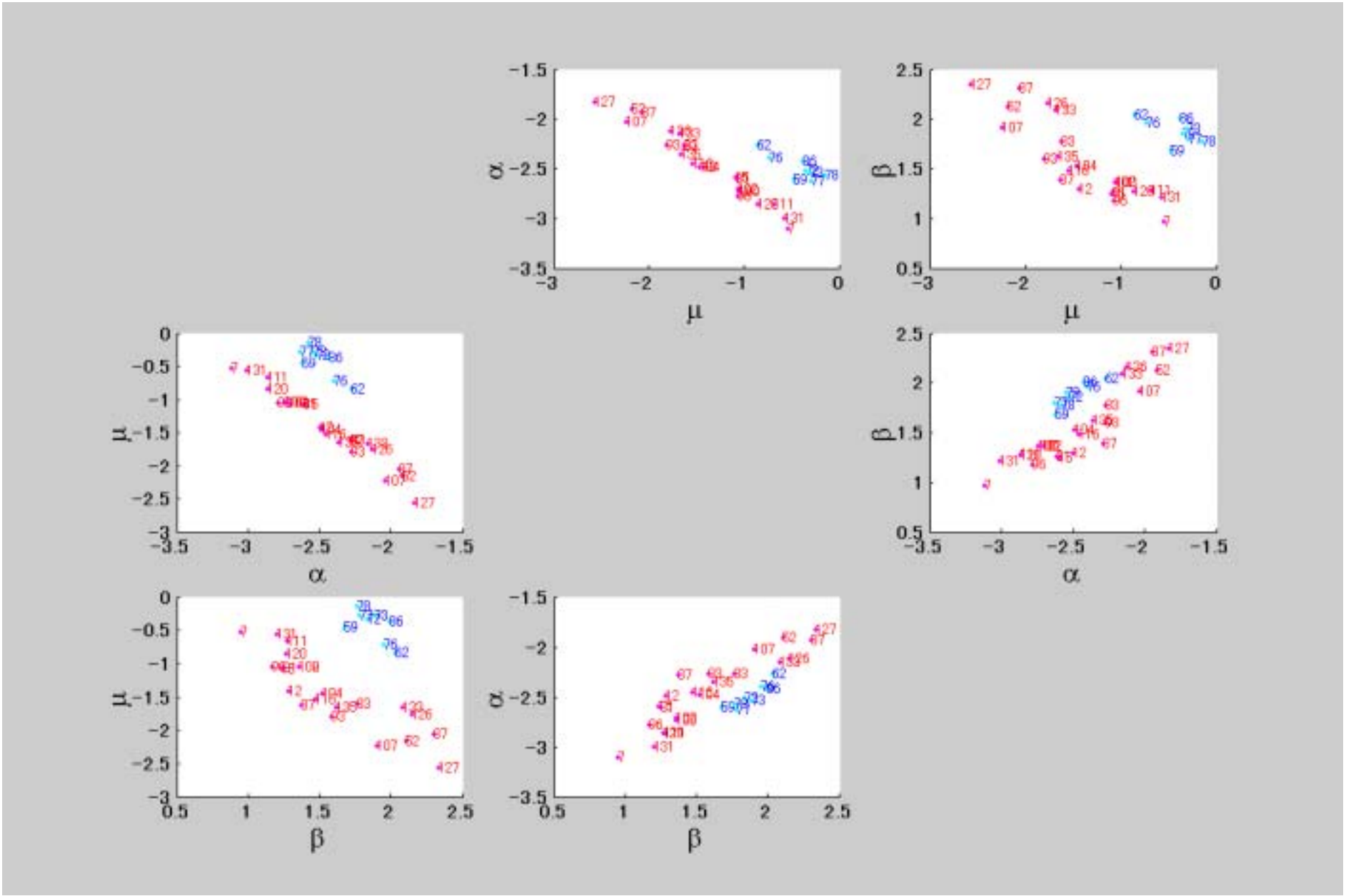
$$|W(G^*)| < |W(G)| \quad \Leftrightarrow \quad \frac{n_t}{n_t+1}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_t)'W(G)^{-1}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_t)$$

$$< \frac{n_k}{n_k-1}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_k)'W(G^*)^{-1}(\boldsymbol{y}_i - \bar{\boldsymbol{y}}_k).$$

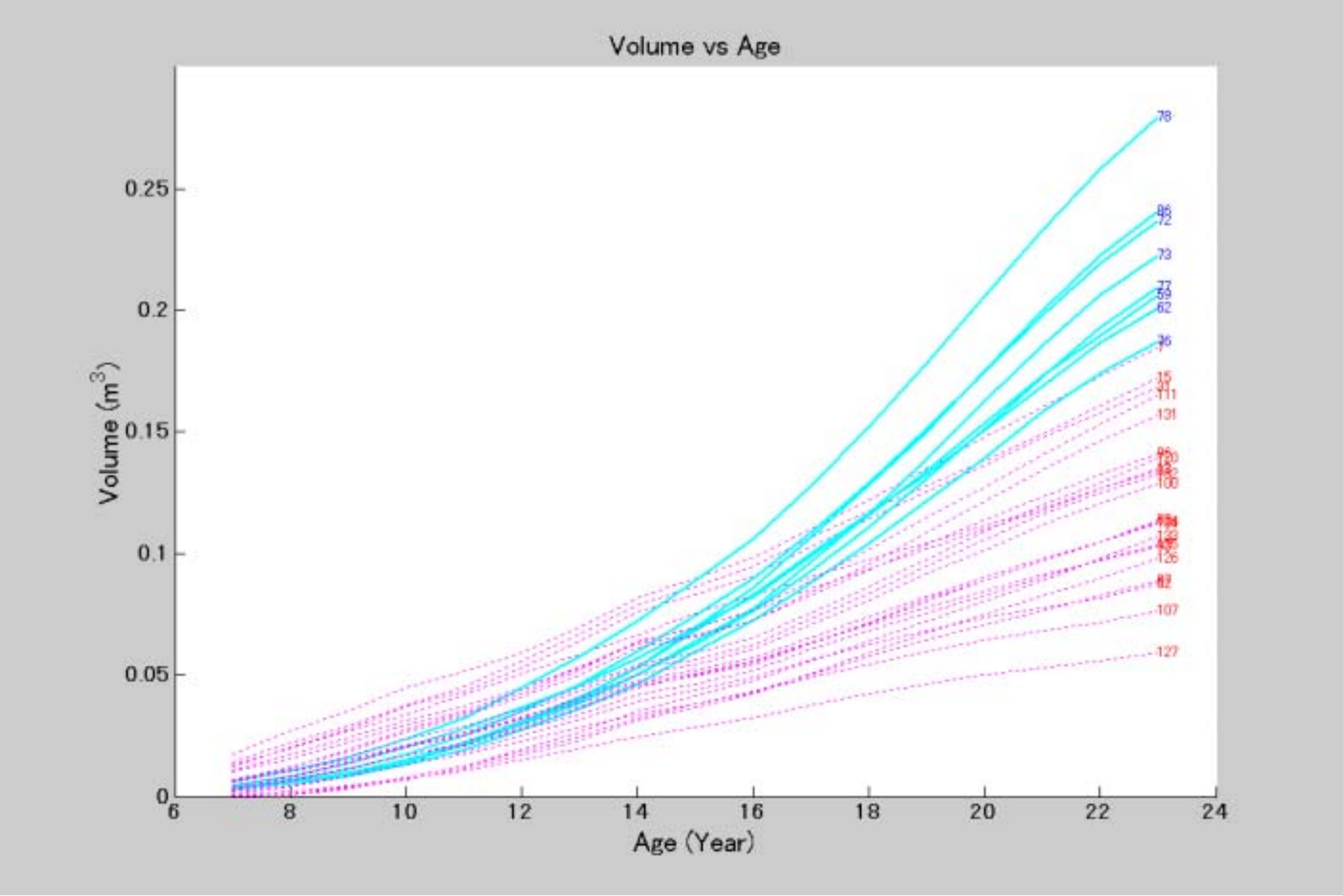
Step 3. Step 2 の操作を分割が収束するまで繰り返す.

注意 初期分割により収束先が変わることもあるので, 初期分割を乱数により割り当て, それを複数回繰り返す. その中で $\mathrm{tr}(W(G))$, または $|W(G)|$ を最小にする分割を最適な分割とする (Gordon, 1981 または Everitt, 1993 等参照).

クラスタリング結果



成長データでのクラスタリング結果



３．クラスターを考慮に入れた多変量線形モデルとその評価

グループに関するモデル

- 1-グループでのモデル：

$$\boldsymbol{y}_i \sim \ i.d.\ \mathrm{N}_p(\Theta'\boldsymbol{x}_i, \Sigma),\ (i = 1, \ldots, n).$$

$X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ とすると, $Y \sim \ \mathrm{N}_{n\times p}(X\Theta, \Sigma \otimes I_n)$.

- g-グループでのモデル：$\boldsymbol{y}_i$ が l 番目のグループに属するとすると,

$$\boldsymbol{y}_i \sim \ i.d.\ \mathrm{N}_p(\Theta_l'\boldsymbol{x}_i, \Sigma),\ (i = 1, \ldots, n).$$

$$\Theta = \begin{pmatrix} \Theta_1 \\ \vdots \\ \Theta_g \end{pmatrix}, \boldsymbol{r}_i = (\boldsymbol{0}', \ldots, \boldsymbol{x}_i', \ldots, \boldsymbol{0}')' = \boldsymbol{\delta}_i \otimes \boldsymbol{x}_i,$$

とおくと, $\Theta'\boldsymbol{r}_i = \Theta_l'\boldsymbol{x}_i$. $X = [\boldsymbol{r}_1, \ldots, \boldsymbol{r}_n]'$ とすると, $Y \sim \ \mathrm{N}_{n\times p}(X\Theta, \Sigma\otimes I_n)$.

同じモデルとして評価可能

正規多変量線形モデル

正規多変量線形モデル：

$$\boldsymbol{y}_i \sim \ i.d.\ \mathrm{N}_p(\Theta'\boldsymbol{x}_i, \Sigma), \quad (i = 1, \ldots, n).$$

- 1-グループでのモデル:

  $\boldsymbol{x}_i$ ： $\boldsymbol{x}_i = \boldsymbol{r}_i$, $\boldsymbol{r}_i$ は $k \times 1$ 説明変数ベクトル.

  Θ ： $k \times p$ 未知パラメータ行列.

- g-グループでのモデル:

  $\boldsymbol{x}_i$ ： $\boldsymbol{x}_i = \boldsymbol{\delta}_i \otimes \boldsymbol{r}_i$, $\boldsymbol{\delta}_i$ はグループに属するかどうかを決定する $g \times 1$ インディケータベクトル. つまり $\boldsymbol{y}_i$ が l グループに属する場合, $\boldsymbol{\delta}_i$ の第 l 番目の成分が 1, 残りは 0.

  Θ ： $kg \times p$ 未知パラメータ行列.

$X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ とすると, $Y \sim \mathrm{N}_{n \times p}(X\Theta, \Sigma \otimes I_n)$.

使用したモデル

正規多変量線形モデル：

$$\boldsymbol{y}_i \sim \ i.d.\ \mathrm{N}_3(\Theta'\boldsymbol{x}_i, \Sigma), \quad (i = 1, \ldots, n).$$

- 1-グループでのモデル:

  $\boldsymbol{x}_i$ ： $\boldsymbol{x}_i = (1, \boldsymbol{r}_i')'$, $\boldsymbol{r}_i$ は $k \times 1$ 説明変数ベクトル.

  Θ ： $k \times p$ 未知パラメータ行列.

- 2-グループでのモデル:

  $$\boldsymbol{x}_i \ : \ \boldsymbol{x}_i = \begin{cases} (1,\ \boldsymbol{r}_i',\ \mathbf{0}'_{(k+1)\times 1})' & \text{第 1 グループに属するとき} \\ (\mathbf{0}'_{(k+1)\times 1},\ 1,\ \boldsymbol{r}_i')' & \text{第 2 グループに属するとき} \end{cases}$$

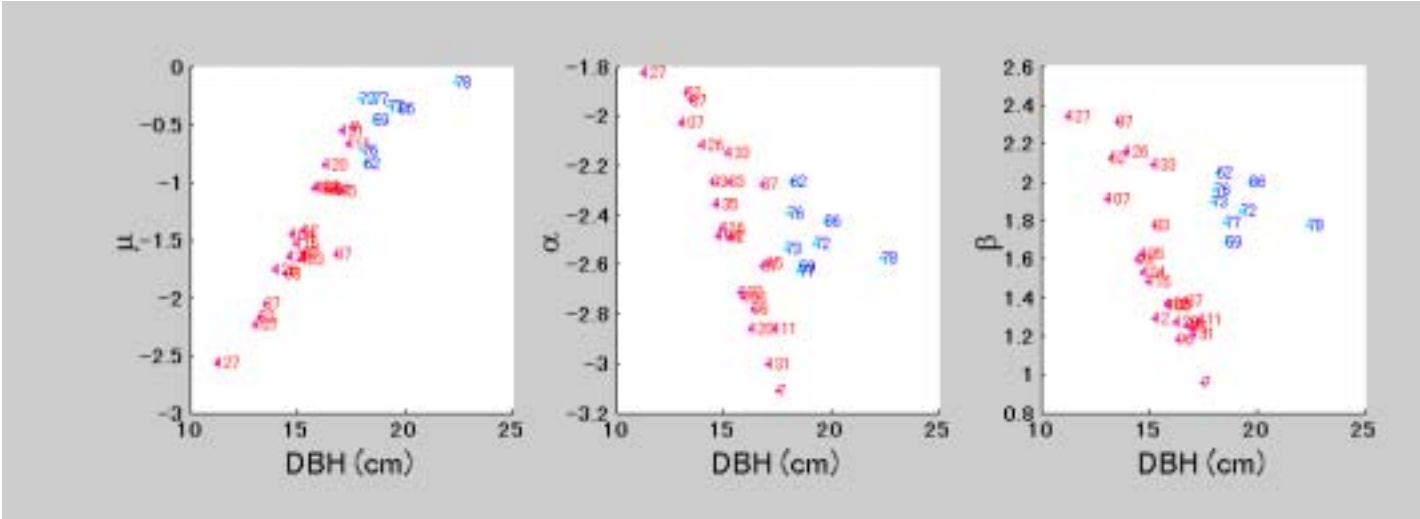
  Θ ： $2(k+1) \times 3$ 未知パラメータ行列.

$X = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ とすると, $Y \sim \mathrm{N}_{n\times 3}(X\Theta, \Sigma \otimes I_n)$.

解析例

- MLE : $\hat{\Theta} = (X'X)^{-1}X'Y$, $\hat{\Sigma} = \frac{1}{n}Y'(I_n - P_X)Y$. ただし P_X は X の列ベクトルである空間に射影する射影行列, $P_X = X(X'X)^{-1}X'$.

- 使用した説明変数 : DBH.

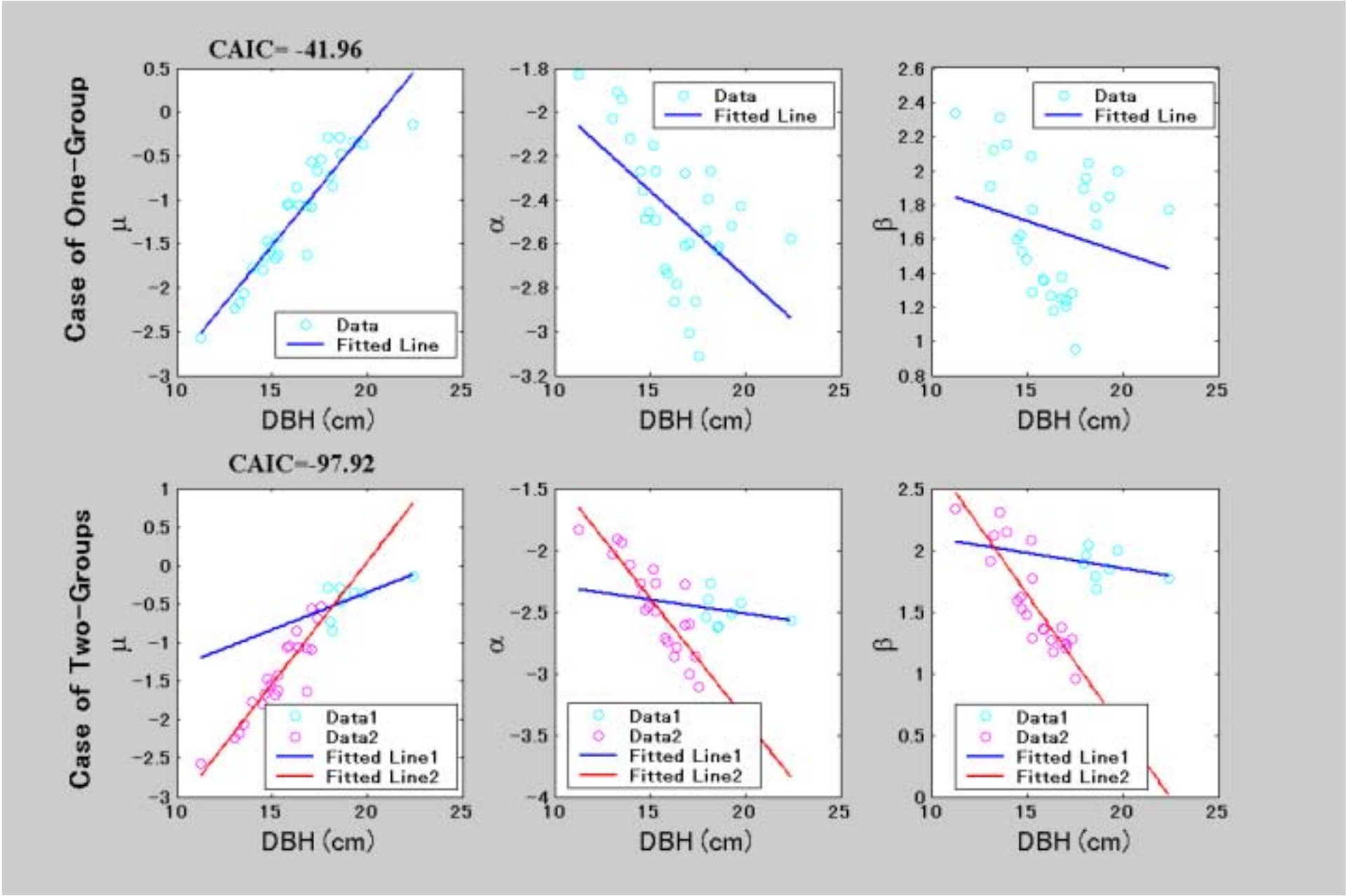


- 使用したモデル評価規準 : CAIC (Fujikoshi and Satoh, 1997).

$$\mathrm{CAIC} = n\log|\hat{\Sigma}| + np\log(2\pi) + \frac{n(n+k_l)p}{n-k_l-p-1}.$$

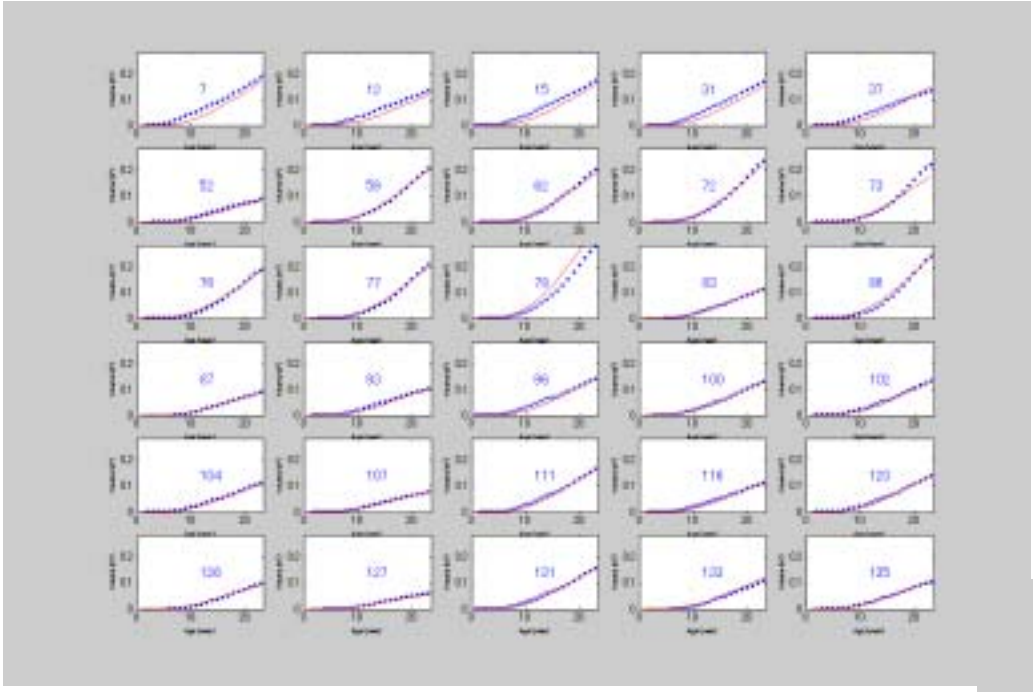
k_l : 説明変数行列の列数 (l=1,2), $l = 1$ のとき $k_1 = 2$, $l = 2$ のとき $k_2 = 4$.

推定結果


CAIC= -41.96
Case of One-Group
μ
α
β
DBH (cm)
Data
Fitted Line
CAIC=-97.92
Case of Two-Groups
Data1
Data2
Fitted Line1
Fitted Line2

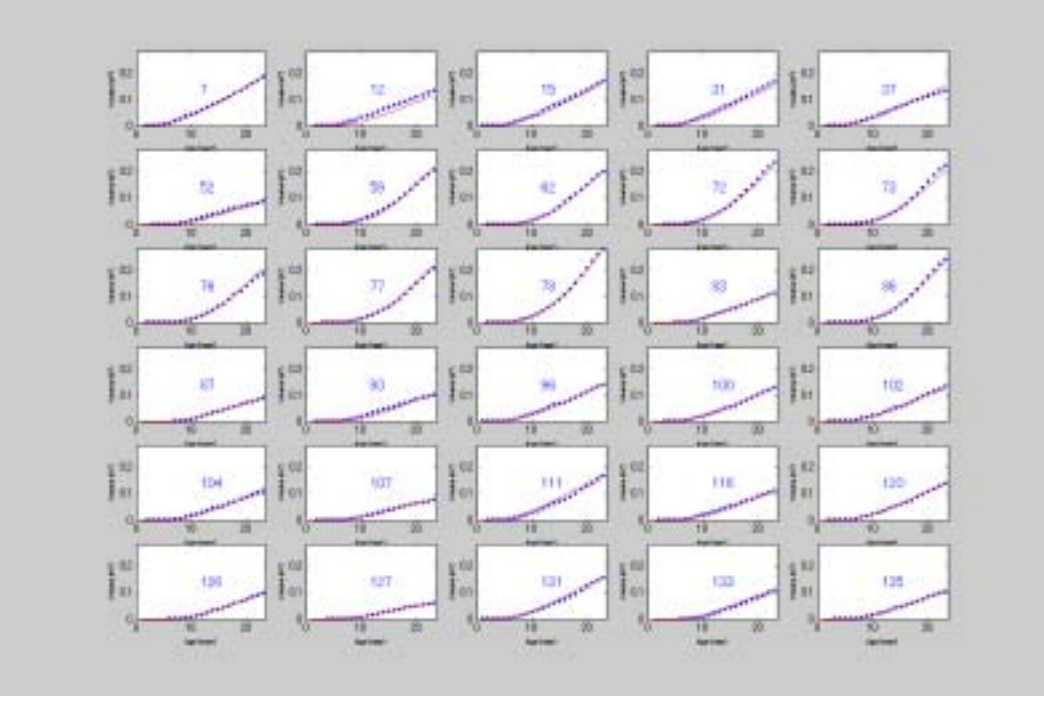
それぞれの推定曲線

成長曲線 $f(t, \boldsymbol{\theta}_i)$ の $\boldsymbol{\theta}_i$ に $\hat{\boldsymbol{\theta}}_i = \hat{\Theta}' \boldsymbol{x}_i$ を入れたときの成長曲線.



Case of one-group

Case of two-groups



４．モデルの外挿とＰＡＩＣを用いた非伐採木の成長パターンのクラスタリング

予測モデル

- 今までは伐採木の成長予測.
- 知りたいのは非伐採木の成長推定

- 実測値モデル：

$$Y \sim \mathrm{N}_{n\times p}(X\Theta, \Sigma \otimes I_n), \quad (Y \text{ は得られたデータ})$$

X ：既知 $n \times k$ 説明変数行列

それぞれ共通

- 予測モデル：

$$Z \sim \mathrm{N}_{m\times p}(W\Theta, \Sigma \otimes I_m), \quad (Z \text{ は得られていないデータ}).$$

W ：既知 $m \times p$ 説明変数行列.

- 多変量正規分布の密度関数：

$$f(Y|X,\Theta,\Sigma) = (2\pi)^{-mp/2}|\Sigma|^{-m/2}\exp\left\{-\frac{1}{2}\mathrm{tr}(Y-X\Theta)'(Y-X\Theta)\right\}.$$

リスクとPAIC

- リスク：

$$\hat{\Theta}_Y = (X'X)^{-1}X'Y, \quad \hat{\Sigma}_Y = \frac{1}{n}Y'(I_n - P_X)Y.$$

Z を $\hat{Z} \sim \mathrm{N}_{m\times p}(W\hat{\Theta}_Y, \hat{\Sigma}_Y \otimes I_m)$ で予測したときのリスク：

$$\begin{aligned} R(W|X) &= -2\mathrm{E}_Y^*\mathrm{E}_Z^*\left[\log\left\{f(Z|W,\hat{\Theta}_Y,\hat{\Sigma}_Y)\right\}\right] \\ &= m\mathrm{E}_Y^*[\log|\hat{\Sigma}_Y|] + mp\log(2\pi) + \frac{n(m+\phi)p}{n-k-p-1}. \end{aligned}$$

ただし $\phi = \mathrm{tr}(W'W(X'X)^{-1})$.

- PAIC : Predictive AIC (Satoh, 1997)

$$\mathrm{PAIC} = m\log|\hat{\Sigma}_Y| + mp\log(2\pi) + \frac{n(m+\phi)p}{n-k-p-1}.$$

PAIC が最小になるモデルを最適なモデルとする！

モデルの内挿と外挿

- モデルの内挿での評価： **同じ観測点でもう一度データを観測したら？**

$$R = -2\mathrm{E}_U^*\mathrm{E}_Y^*\left[\log\left\{f(U|X,\hat{\Theta}_Y,\hat{\Sigma}_Y)\right\}\right]$$
$$= n\mathrm{E}_Y^*[\log|\hat{\Sigma}_Y|] + np\log(2\pi) + \sum_{i=1}^{n}\mathrm{E}_U^*\mathrm{E}_Y^*\left[(\boldsymbol{u}_i - \hat{\Theta}_Y'\boldsymbol{x}_i)'\hat{\Sigma}_Y^{-1}(\boldsymbol{u}_i - \hat{\Theta}_Y'\boldsymbol{x}_i)\right].$$

$U = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)'$: Y と独立に同じ分布に従う.
$U \sim \mathrm{N}_{n\times p}(X\Theta, \Sigma\otimes I_n)$.

- モデルの外挿での評価： **違う観測点でデータを観測したら？**

$$R(W|X) = -2\mathrm{E}_Y^*\mathrm{E}_Z^*\left[\log\left\{f(Z|W,\hat{\Theta}_Y,\hat{\Sigma}_Y)\right\}\right]$$
$$= m\mathrm{E}_Y^*[\log|\hat{\Sigma}_Y|] + mp\log(2\pi) + \sum_{j=1}^{m}\mathrm{E}_Y^*\mathrm{E}_Z^*\left[(\boldsymbol{z}_j - \hat{\Theta}_Y'\boldsymbol{w}_j)'\hat{\Sigma}_Y^{-1}(\boldsymbol{z}_j - \hat{\Theta}_Y'\boldsymbol{w}_j)\right]$$

$Z = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_m)'$, $W = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n)'$.

X と W が同じであれば, 二つのリスクは一致！

クラスタリング & 最適化法

Step 1. 説明変数の組を決定する.

Step 2. $\phi = \sum_{j=1}^{m} \boldsymbol{w}_j'(X'X)^{-1}\boldsymbol{w}_j$ を最小にする分類を行う. つまり第 j 番目の予測データに関して,

説明変数を固定するとPAIC の大きさはφの大きさだけに依存

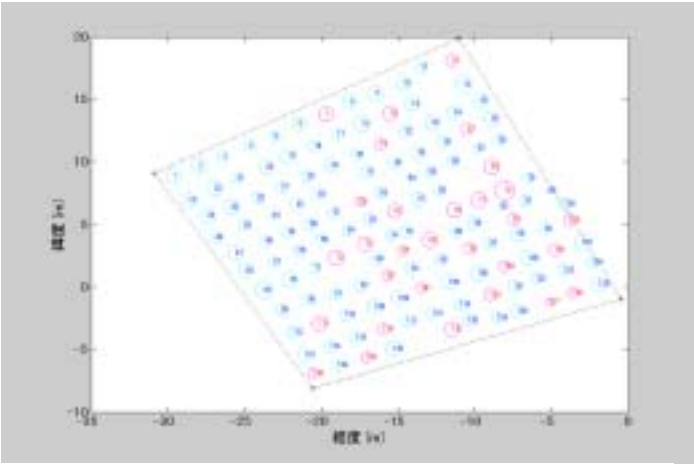
$$d_l = (\boldsymbol{e}_l \otimes \boldsymbol{v}_j)'(X'X)^{-1}(\boldsymbol{e}_l \otimes \boldsymbol{v}_j), \quad (l = 1, \ldots, g),$$

を最小にするグループに分類する ($\boldsymbol{e}_l$: l 番目の成分だけ 1, 残りは 0). ただし, $\boldsymbol{v}_j$ は予測モデルに関する説明変数. その最小になる PAIC をその説明変数の組での PAIC とする.

Step 3. すべての説明変数の組み合わせで *Step 2* を行い, PAIC を最小にするモデルを最適なモデルとする.

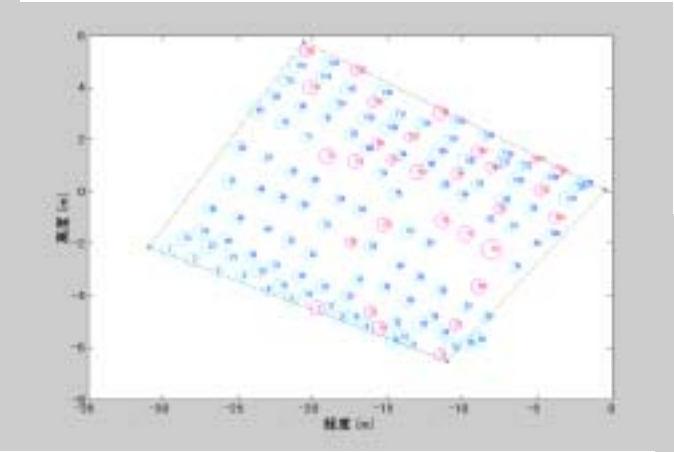
解析の設定

- 観測値の分類は k-平均法で求めたクラスターで固定.
- 使用した説明変数は, 現時点の DBH, 経度, 緯度, 高度.
- PAIC を用いた分類とモデルの最適化を行う.
- 1 グループと2 グループで両方で比較.

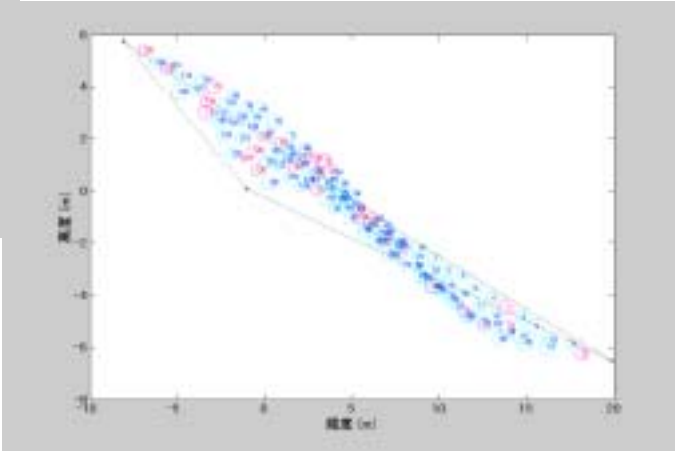


経度 vs 緯度

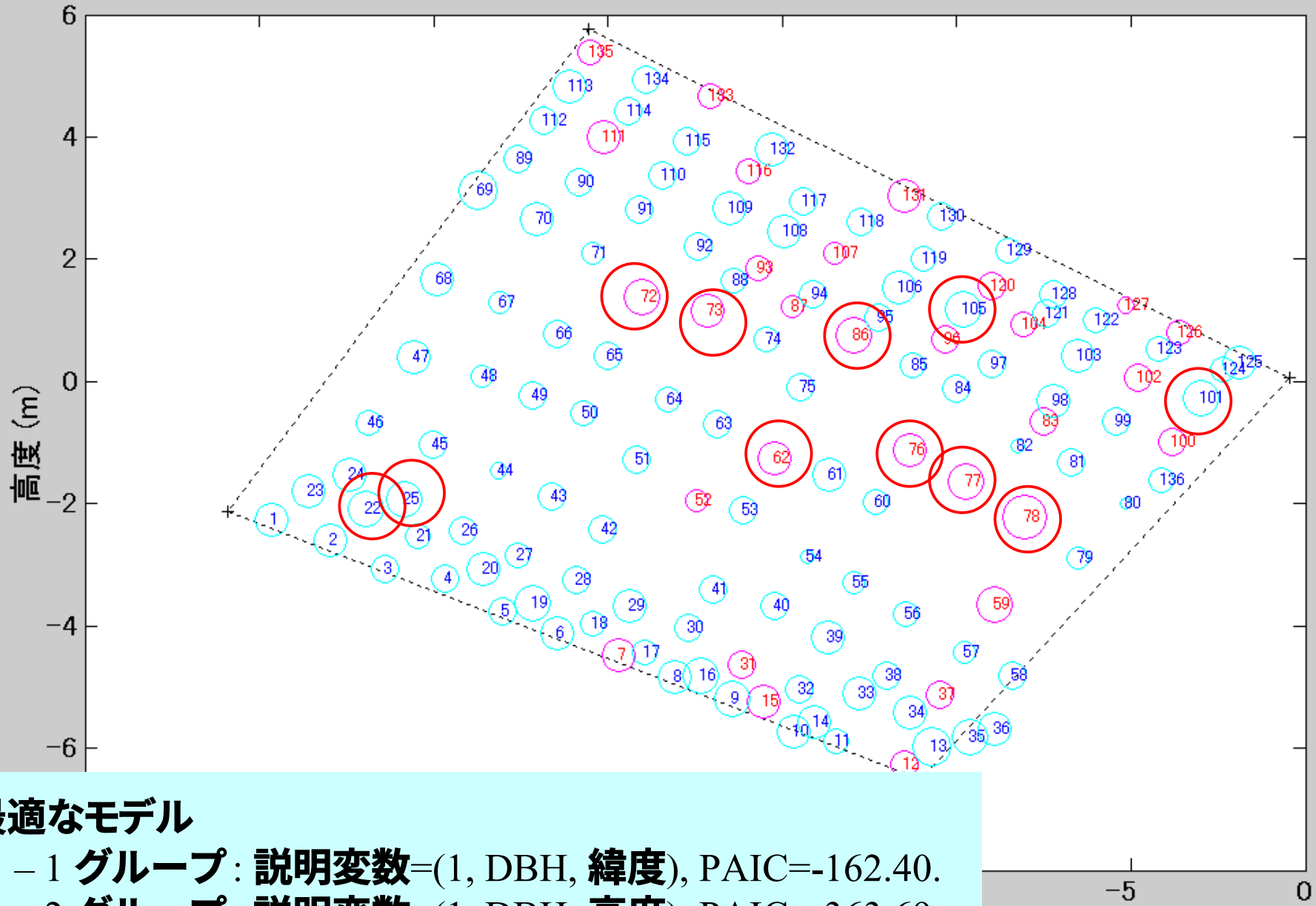
経度 vs 高度



緯度 vs 高度

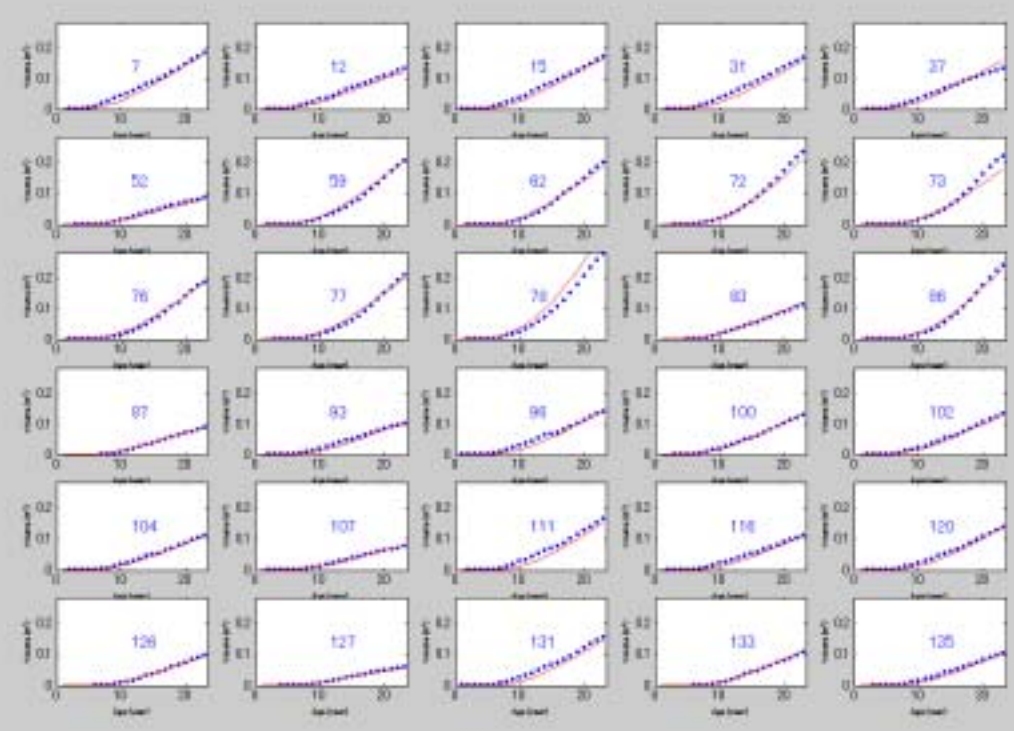


分類結果



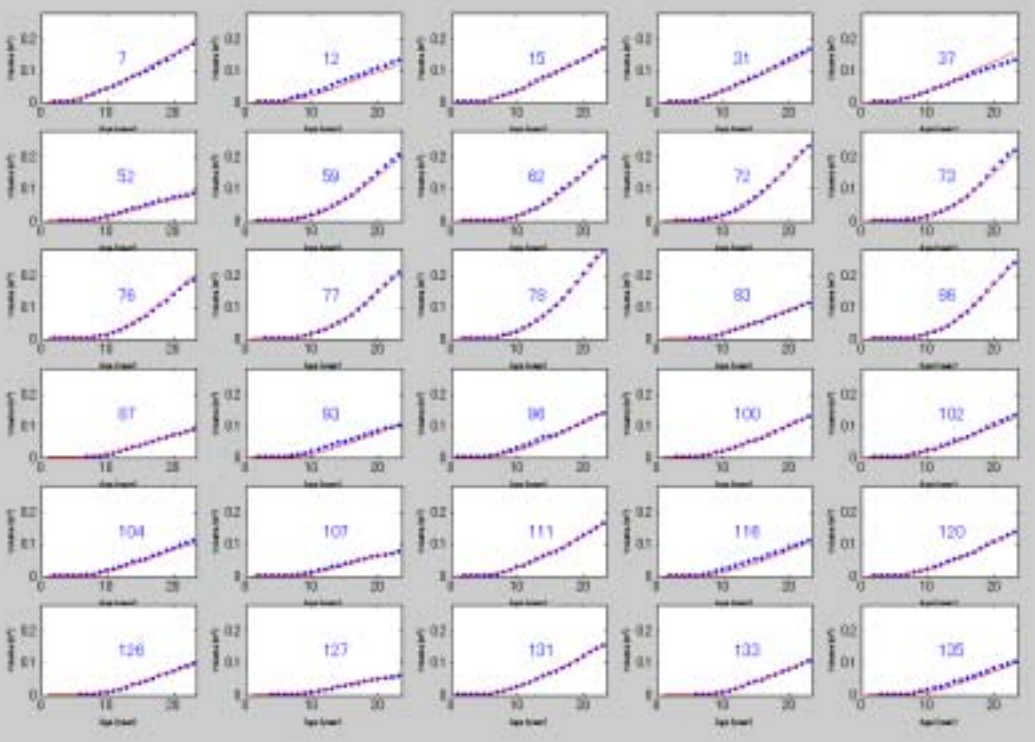
- **最適なモデル**
  - 1 **グループ**: **説明変数**=(1, DBH, **緯度**), PAIC=-162.40.
  - 2 **グループ**: **説明変数**=(1, DBH, **高度**), PAIC=-363.69.

伐採木の推定曲線

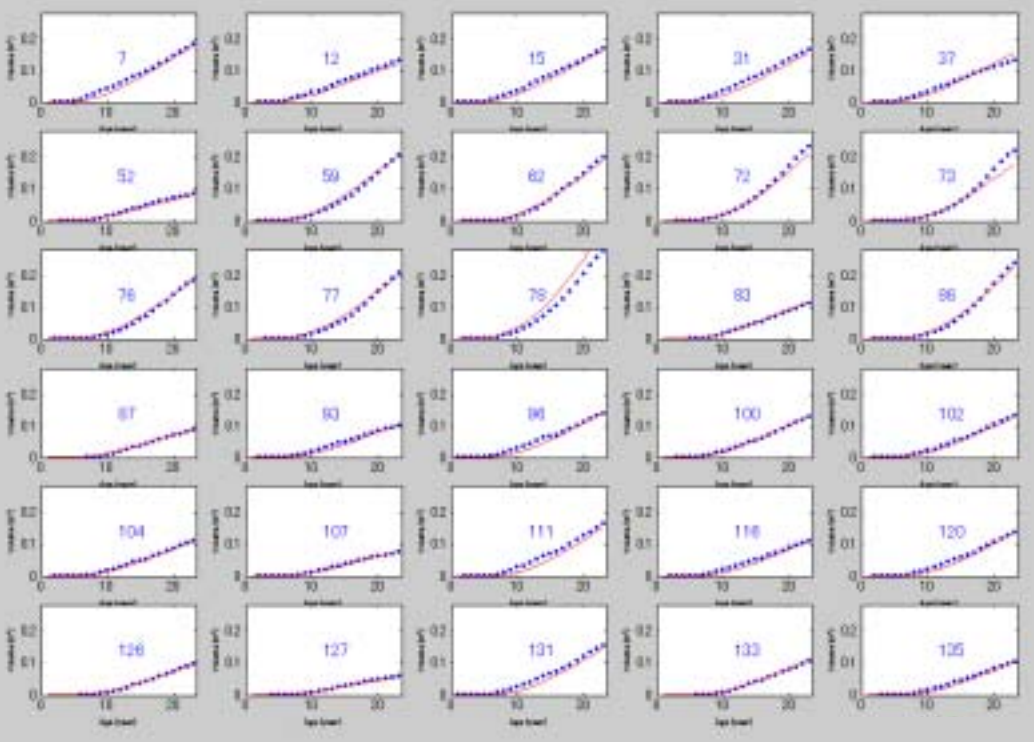


Case of one-group

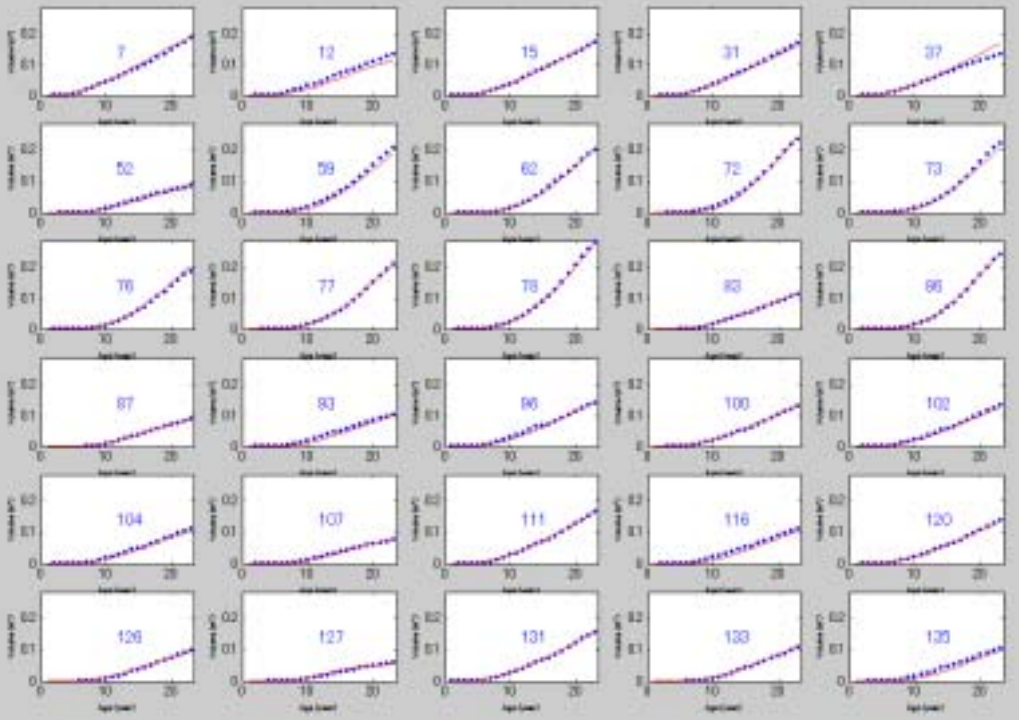
Case of two-groups



伐採木の推定曲線
(Cross Validation)



Case of two-groups



Case of one-group

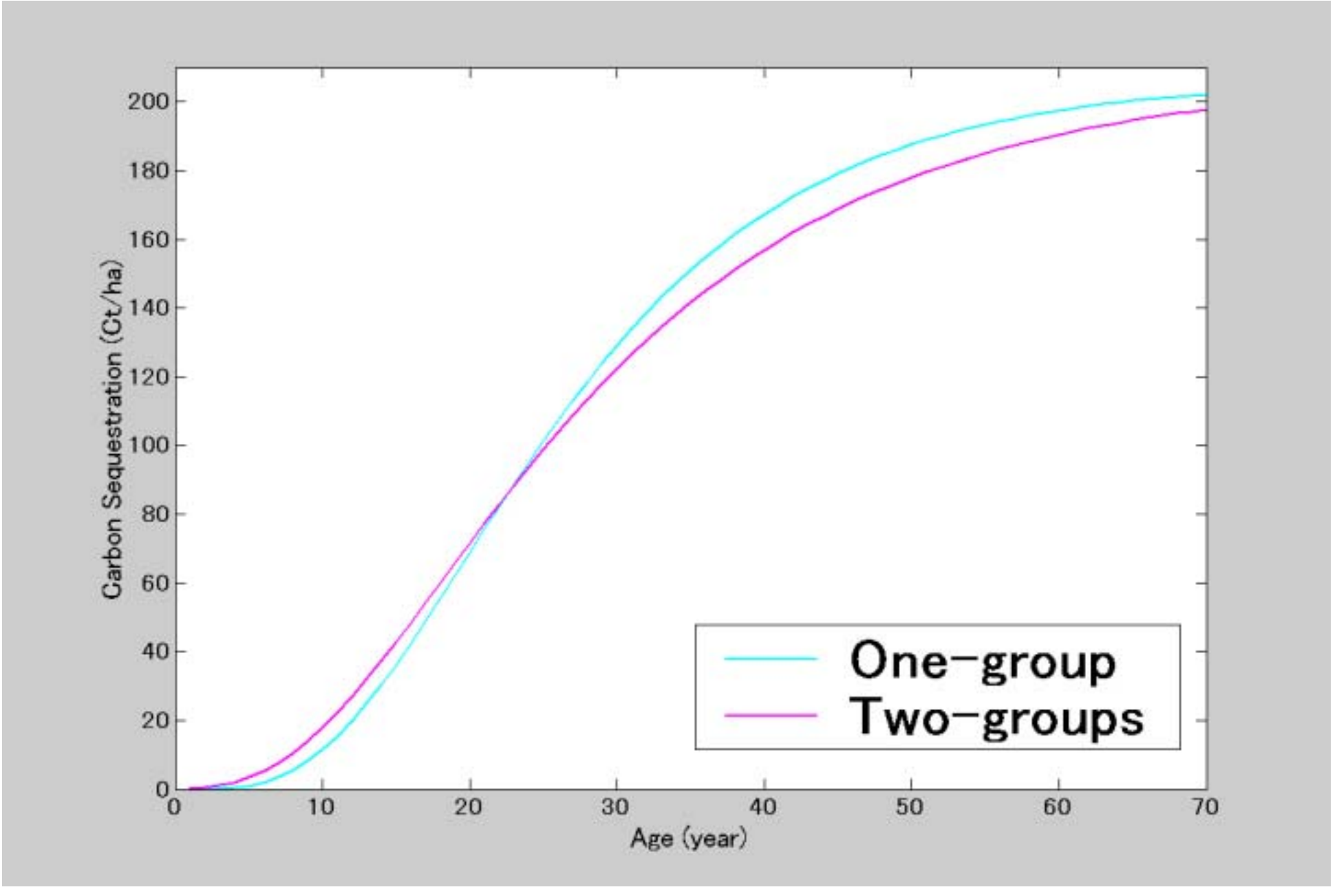
Cross Validation 法 :

i 番目のデータを抜いた $Y_{(-i)}$, $X_{(-i)}$ で予測.

$\hat{\Theta}_{[-i]} = (X'_{(-i)}X_{(-i)})^{-1}X'_{(-i)}Y_{(-i)}$ とすると,

$$\hat{\boldsymbol{y}}_{i[-i]} = \hat{\Theta}'_{[-i]}\boldsymbol{x}_i.$$

総炭素固定量


Carbon Sequestration (Ct/ha)
Age (year)
0
20
40
60
80
100
120
140
160
180
200
0
10
20
30
40
50
60
70
One-group
Two-groups

５．まとめとこれからの課題

まとめ

1. 伐採された林木から成長曲線を推定し, 推定された係数から成長パターンを分類.
   - 分類には k-平均法を用いた.
   - 成長関数には Richards の成長関数を用いた.
2. グループ別けを考慮に入れた多変量線形モデルでの推定.
3. PAIC を用いた非伐採木の成長パターンの分類と最適なモデルの探索.

これらの課題

1. PAIC 使う以外の分類法 (判別等).
2. 正規性の仮定と非正規性の影響.
   - モデルの内挿 : Yanagihara (2004) で非正規性の影響を受けにくいでの規準量を提案. しかし外挿では難しい.
3. その他の説明変数の利用 (占有率等).
4. ランダム効果を伴った非線形多変量回帰モデルへの拡張.
5. とにかく現在わかる情報が少ない！その他のデータをとる (例えば樹高とか).

参考文献 (その 1)

1. Everitt, B. S. (1993). *Cluster Analysis.* 3rd. ed. Edward Arnold.
2. Fujikoshi, Y. and Satoh, K. (1997). Modified AIC and C_p in multivariate linear regression. *Biometrika*, **84**, 707-716.
3. Gordon, A. D. (1981). *Classification*. Chapman and Hall.
4. MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability* (ed. J. Neyman), **1**, 281-298, Berkeley.
5. Ohtaki, M. and Izumi, S. (1999). Globally convergent algorithm without derivatives for maximizing a multivariate function. *Presented at the Symposium on "Exploratory Methods and Analysis for Nonlinear Structures of Data with Random Variation"* in Hiroshima, Japan, January 7 - 11, 1999.

参考文献 (その 2)

6. Richards, F. J. (1958). A flexible growth function to empirical use. *J. Exp. Bot.*, **10**, 290-300.
7. Satoh, K. (1997). AIC-type model selection criterion for multivariate linear regression with a future experiment. *J. Japan Statist. Soc.*, **27**, 135-140.
8. Yanagihara, H. (2004). Corrected version of *AIC* for selecting multivariate normal linear regression models in a general nonnormal case. TR No. 04-04, *Statistical Research Group, Hiroshima University*.
9. Yanagihara, H. and Yoshimoto, A. (2004). Statistical procedure for assessing the amount of carbon sequestrated by sugi (*Cryptomeria japonica*) plantation. *Institute of Policy and Planning Sciences, Discussion Paper Series*, No. 1076.
10. 柳**原 宏和**, **吉本 敦**, **能本 美穂**. (2003). **林分成長分析のための一般化非線形混合効果モデル**. *Institute of Policy and Planning Sciences, Discussion Paper Series*, No. 1071.