分布に対する分類樹・クラタリング手法と
それらを用いた空間の分割

慶應義塾大学理工学部　南美穂子
Inter-American Tropical Tuna Commission　Cleridy Lennert-cody

はじめに 本研究では，分布に対する回帰・分類樹 (CART)，および，クラスタリング手法を考え，さらに，空間内の各地点に対して分布が与えられたときに，分布に対する回帰・分類樹やクラスタリング法を用いて空間を分割する方法を考える．生物の体長などの身体的特性は，生息する場所，季節などの環境要因によって分布が異なる．本研究の目的は，分布を決める要因を解析すること，また，それらの要因によって予測される分布の特定，および，分布のクラスタリングとそれに伴う空間の分割を行うための手法の開発である．応用例として東部太平洋で操業するマグロ漁船で計測されたキハダマグロの体長データの解析を行い，回帰・分類樹を用いた直線による海域の分割，クラスタリングによる柔軟な境界を持つ海域の分割を試みる．

問題設定 ケース i $(i = 1, \cdots, n)$ に対して，分布 p_i，その確信度 m_i, 説明変数ベクトル $\boldsymbol{x}_i$, 隣接情報 S_i が与えられるとする．例えば，キハダマグロ体長データ解析の例では，ケースは地点，分布 p_i はその地点で捕獲されたキハダマグロの体長を区間に分けたときの各区間にはいる個体数の割合で表した分布，確信度はその地点で体長が観測された個体数，説明変数ベクトルは，緯度，経度，捕獲された季節など，隣接情報は，その地点と隣接する地点のケース番号の集合である．

ケース間の距離，不純度，集合間の距離 分布 q から分布 p への Kullback-Leibler (KL) ダイバージェンスを $D(p|q)$ $\left(\equiv \sum_{x\in\Omega_q} p(x)\log\frac{p(x)}{q(x)}\right)$ で表す．分布の集合 $\{p_i; i\in\mathcal{G}\}$ の平均分布を $\overline{p}(p_i; i\in\mathcal{G}) = \frac{1}{\sum_{i\in\mathcal{G}} m_i}\sum_{i\in\mathcal{G}} m_i p_i$ とする．分布間の非類似度（距離）$d(p_i, p_j)$，分布の集合の不純度 $Imp(p_i; i\in\mathcal{G})$, 分布の集合間の距離 $d(\{p_i; i\in\mathcal{G}_L\}, \{p_i; i\in\mathcal{G}_R\})$ を

分布間の非類似度
$$d(p_i, p_j) = m_i D\left(p_i|\overline{p}(p_i, p_j)\right) + m_j D\left(p_j|\overline{p}(p_i, p_j)\right)$$

分布の集合の不純度
$$Imp(p_i; i\in\mathcal{G}) = \sum_{i\in\mathcal{G}} m_i D\left(p_i|\overline{p}(p_i; i\in\mathcal{G})\right)$$

分布の集合間の距離
$$d(\{p_i; i\in\mathcal{G}_L\}, \{p_i; i\in\mathcal{G}_R\}) = Imp(p_i; i\in\mathcal{G}_L\cup\mathcal{G}_R) - Imp(p_i; i\in\mathcal{G}_L) - Imp(p_i; i\in\mathcal{G}_R)$$

で定義する．このとき，分布の集合間の距離，および，クラスターの結合による距離の更新式は，各々の平均分布，および，和集合の平均分布のエントロピーで表すことが示せる．

回帰・分類樹 分布に対する回帰・分類樹では，分割による不純度の改善は，分割された 2 つの子ノードの集合間の距離である．すべてのケースが根ノードに属する状態から始め，説明変数の値に基づく分割候補の中で，分割後の子ノードに属する分布の集合間の距離が最大となる分割を選択する．説明変数として，緯度，経度を含めることにより，得られた回帰樹の条件式に基づいて海域を分割する．

クラスタリング 凝集型の階層的クラスタリングを考える．各分布 1 つをクラスターとする状態から始め，クラスターをなす分布の集合間の距離が最も小さい 2 つのクラスターの結合を全体が 1 つになるまで繰り返す．分布に基づく空間の分割を行う場合は，結合するクラスターを隣接する集合のみに制限してクラスタリングを行う．

参考文献： Cleridy E. Lennert-Cody, Mark N. Maunder, Alexandre Aires-da-Silva, Mihoko Minami (2013) Defining population spatial units: Simultaneous analysis of frequency distributions and time series Fisheries Research 139, 85-92