Non-Hierarchical Clustering for Distribution-Valued Data

Yoshikazu Terada
Graduate School of Culture and Information Science,
Doshisha University.

Hiroshi Yadohisa
Department of Culture and Information Science,
Doshisha University.

➢INDEX

1.1 Symbolic Data Analysis (SDA)

- In recent years,
  - Development of the Internet
  - Improvement of computer performance

1.1 Symbolic Data Analysis (SDA)

- In recent years,
  - Development of the Internet
  - Improvement of computer performance
  - We deal with

    "**Large** data" and "more **Complex** information".

1.1 Symbolic Data Analysis (SDA)

- In recent years,
  - Development of the Internet
  - Improvement of computer performance
  - We deal with

    "**Large** data" and "more **Complex** information".

In some cases,
Difficult to analyze them by using classical methods.

In some cases,
Difficult to describe them by classical data.

1.1 Symbolic Data Analysis (SDA)

- In recent years,
  - Development of the Internet
  - Improvement of computer performance

  - We deal with

    "**Large** data" and "more **Complex** information".

In some cases,
Difficult to analyze them by using classical methods.

In some cases,
Difficult to describe them by classical data.

Some new methods for analyzing them are required.

1.1 Symbolic Data Analysis (SDA)

- In recent years,
  - Development of the Internet
  - Improvement of computer performance
  - We deal with

    "**Large** data" and "more **Complex** information".

In some cases,
Difficult to analyze them by using classical methods.

In some cases,
Difficult to describe them by classical data.

Some new methods for analyzing them are required.

"**Symbolic data analysis**"

1.1 Symbolic Data Analysis (SDA)

- **Symbolic data analysis (SDA)**
  - **A extended classical data analysis** for more complex data table called "**symbolic data table**"

1.1 Symbolic Data Analysis (SDA)

- **Symbolic data analysis (SDA)**
  - **A extended classical data analysis** for more complex data table called "**symbolic data table**"

•A more complex data table
•A cell of that cannot only contain a single quantitative (categorical) value.

1.1 Symbolic Data Analysis (SDA)

- **Symbolic data analysis (SDA)**
  - **A extended classical data analysis** for more complex data table called "**symbolic data table**"

•A more complex data table
•A cell of that cannot only contain a single quantitative (categorical) value.

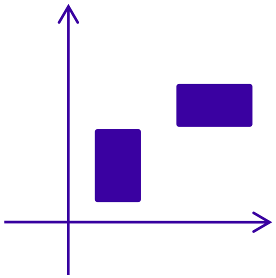
  - SDA has been studied as one of useful methods for analyzing **large** and **complex** datasets.

1.1 Symbolic Data Analysis (SDA)

- **Symbolic data analysis (SDA)**
  - **A extended classical data analysis** for more complex data table called "**symbolic data table**"

•A more complex data table
•A cell of that cannot only contain a single quantitative (categorical) value.

  - SDA has been studied as one of useful methods for analyzing **large** and **complex** datasets.
- **Typical Symbolic data**
  - Interval-valued data



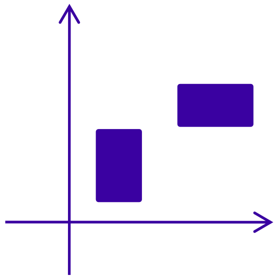
1.1 Symbolic Data Analysis (SDA)

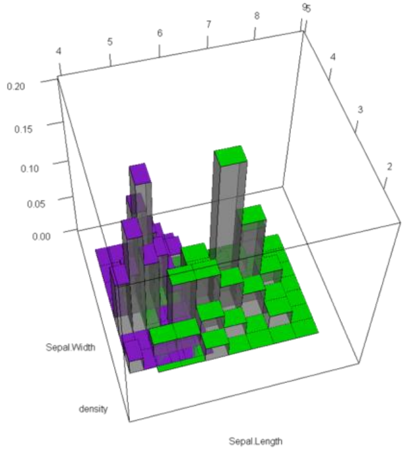
- **Symbolic data analysis (SDA)**
  - **A extended classical data analysis** for more complex data table called "**symbolic data table**"

•**A more complex data table**
•**A cell of that cannot only contain a single quantitative (categorical) value.**

  - SDA has been studied as one of useful methods for analyzing **large** and **complex** datasets.
- **Typical Symbolic data**
  - Interval-valued data
  - Distribution-valued data

⋮





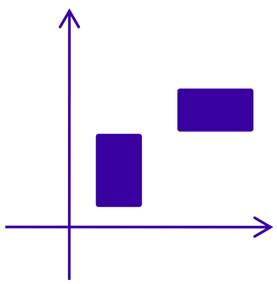
1.1 Symbolic Data Analysis (SDA)

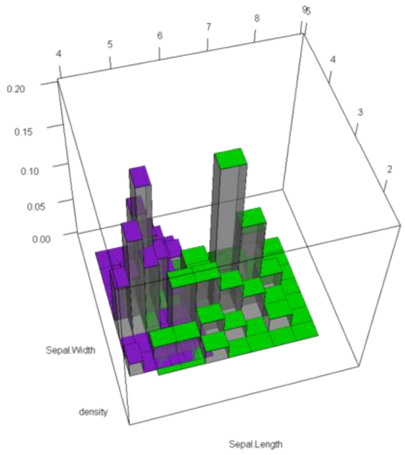
- **Symbolic data analysis (SDA)**
  - **A extended classical data analysis** for more complex data table called "**symbolic data table**"

> •**A more complex data table**
> •**A cell of that cannot only contain a single quantitative (categorical) value.**

  - SDA has been studied as one of useful methods for analyzing **large** and **complex** datasets.
- **Typical Symbolic data**
  - Interval-valued data
  - **Distribution-valued data**

⋮





1.2 Distribution-valued data

- **What is Distribution-valued data?**
  - A cell of such data contains a "**distribution**".

e.g.) **distribution function, density function (histogram) …**

1.2 Distribution-valued data

- **What is Distribution-valued data?**
  - A cell of such data contains a "**distribution**".

e.g.) **distribution function, density function (histogram) …**

Classical data

Variables

$$
\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1k} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}
$$

Entity

1.2 Distribution-valued data

- **What is Distribution-valued data?**
  - A cell of such data contains a "**distribution**".

e.g.) **distribution function, density function (histogram) …**

Classical data

Variables

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1k} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

Entity

Represented by
a single point in $\mathbb{R}^p$

1.2 Distribution-valued data

- **What is Distribution-valued data?**
  - A cell of such data contains a "**distribution**".

e.g.) **distribution function, density function (histogram)** …

Classical data

Variables

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1k} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

Entity

Represented by a single point in $\mathbb{R}^p$

Distribution-valued data

Variables

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1k} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

Entity

1.2 Distribution-valued data

- **What is Distribution-valued data?**
  - A cell of such data contains a "**distribution**".

e.g.) **distribution function, density function (histogram)** …

Classical data

Variables

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1k} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

Entity

Represented by a single point in $\mathbb{R}^p$

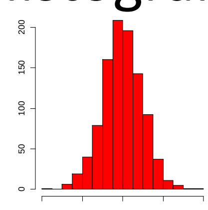
Distribution-valued data

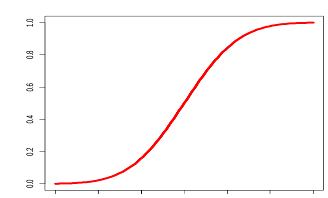
Variables

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & \cdots & x_{1k} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

Entity

e.g.) Histogram, distribution function…





…

1.2 Distribution-valued data

- **When we use distribution-valued data?**
  - Aggregate Large data to more manageable data
  - Describe objects with several values on a one variable

⋮

1.2 Distribution-valued data

- **When we use distribution-valued data?**
  - Aggregate Large data to more manageable data
  - Describe objects with several values on a one variable

⋮

- **Advantages of distribution-valued data**

Initial data

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_1 1} & x_{n_1 2} & \cdots & x_{n_1 k} & \cdots & x_{n_1 p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nk} & \cdots & x_{Np} \end{bmatrix}$$

aggregate

Classical data

$$\boldsymbol{X}^* = \begin{bmatrix} \bar{x}_{11} & \cdots & \bar{x}_{1k} & \cdots & \bar{x}_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{i1} & \cdots & \bar{x}_{ik} & \cdots & \bar{x}_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{n1} & \cdots & \bar{x}_{nk} & \cdots & \bar{x}_{np} \end{bmatrix}$$

1.2 Distribution-valued data

- **When we use distribution-valued data?**
  - Aggregate Large data to more manageable data
  - Describe objects with several values on a one variable

⋮

- **Advantages of distribution-valued dat**

Describe by a **single** representative value (e.g. mean)

Initial data

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_1 1} & x_{n_1 2} & \cdots & x_{n_1 k} & \cdots & x_{n_1 p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nk} & \cdots & x_{Np} \end{bmatrix}$$

aggregate

Classical data

$$X^* = \begin{bmatrix} \bar{x}_{11} & \cdots & \bar{x}_{1k} & \cdots & \bar{x}_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{i1} & \cdots & \bar{x}_{ik} & \cdots & \bar{x}_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{n1} & \cdots & \bar{x}_{nk} & \cdots & \bar{x}_{np} \end{bmatrix}$$

1.2 Distribution-valued data

- **When we use distribution-valued data?**
  - Aggregate Large data to more manageable data
  - Describe objects with several values on a one variable

⋮

- **Advantages of distribution-valued dat**

Initial data

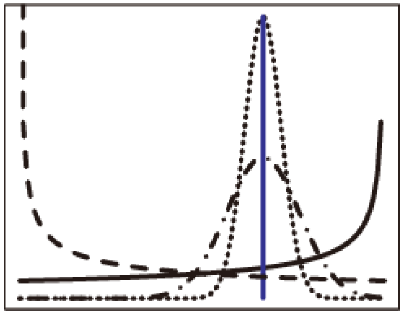
$$
\boldsymbol{X} = \begin{bmatrix}
x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\
x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{n_1 1} & x_{n_1 2} & \cdots & x_{n_1 k} & \cdots & x_{n_1 p} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{ip} \\
\vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\
x_{N1} & x_{N2} & \cdots & x_{Nk} & \cdots & x_{Np}
\end{bmatrix}
$$

aggregate

Classical data

$$
\boldsymbol{X}^* = \begin{bmatrix}
\bar{x}_{11} & \cdots & \bar{x}_{1k} & \cdots & \bar{x}_{1p} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
\bar{x}_{i1} & \cdots & \bar{x}_{ik} & \cdots & \bar{x}_{ip} \\
\vdots & \ddots & \vdots & \ddots & \vdots \\
\bar{x}_{n1} & \cdots & \bar{x}_{nk} & \cdots & \bar{x}_{np}
\end{bmatrix}
$$

Describe by a **single** representative value (e.g. mean)



1.2 Distribution-valued data

- **When we use distribution-valued data?**
  - Aggregate Large data to more manageable data
  - Describe objects with several values on a one variable

⋮

- **Advantages of distribution-valued dat**

Initial data

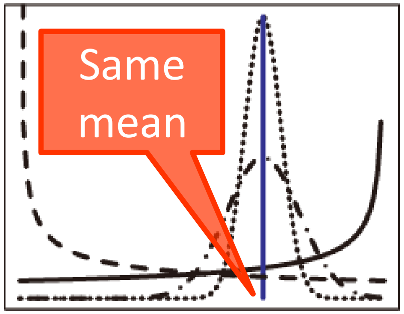
$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_1 1} & x_{n_1 2} & \cdots & x_{n_1 k} & \cdots & x_{n_1 p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nk} & \cdots & x_{Np} \end{bmatrix}$$

aggregate

Classical data

$$\boldsymbol{X}^* = \begin{bmatrix} \bar{x}_{11} & \cdots & \bar{x}_{1k} & \cdots & \bar{x}_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{i1} & \cdots & \bar{x}_{ik} & \cdots & \bar{x}_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{n1} & \cdots & \bar{x}_{nk} & \cdots & \bar{x}_{np} \end{bmatrix}$$

Describe by a **single** representative value (e.g. mean)



1.2 Distribution-valued data

- **When we use distribution-valued data?**
  - Aggregate Large data to more manageable data
  - Describe objects with several values on a one variable

⋮

- **Advantages of distribution-valued dat**

Initial data

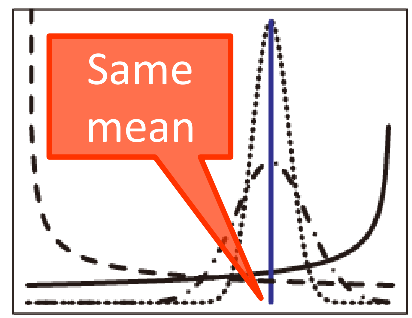
$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_1 1} & x_{n_1 2} & \cdots & x_{n_1 k} & \cdots & x_{n_1 p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nk} & \cdots & x_{Np} \end{bmatrix}$$

aggregate

Classical data

$$\boldsymbol{X}^* = \begin{bmatrix} \bar{x}_{11} & \cdots & \bar{x}_{1k} & \cdots & \bar{x}_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{i1} & \cdots & \bar{x}_{ik} & \cdots & \bar{x}_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \bar{x}_{n1} & \cdots & \bar{x}_{nk} & \cdots & \bar{x}_{np} \end{bmatrix}$$

Describe by a **single** representative value (e.g. mean)



Information loss of **Distribution structure**

1.2 Distribution-valued data

- **<u>When we use distribution-valued data?</u>**
  - Aggregate Large data to more manageable data
  - Describe objects with several values on a one variable

⋮

- **<u>Advantages of distribution-valued data</u>**

Initial data

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_1 1} & x_{n_1 2} & \cdots & x_{n_1 k} & \cdots & x_{n_1 p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nk} & \cdots & x_{Np} \end{bmatrix}$$

aggregate

Distribution-valued data

$$\boldsymbol{X}_{\mathrm{d}} = \begin{bmatrix} \xi_{11} & \cdots & \xi_{1k} & \cdots & \xi_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \xi_{i1} & \cdots & \xi_{ik} & \cdots & \xi_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \xi_{n1} & \cdots & \xi_{nk} & \cdots & \xi_{np} \end{bmatrix}$$

1.2 Distribution-valued data

- **When we use distribution-valued data?**
  - Aggregate Large data to more manageable data
  - Describe objects with several values on a one variable

⋮

- **Advantages of distribution-valued data**

Initial data

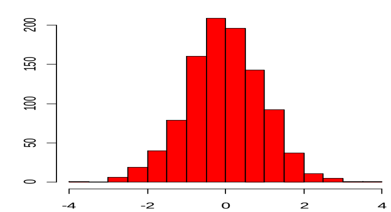
$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_1 1} & x_{n_1 2} & \cdots & x_{n_1 k} & \cdots & x_{n_1 p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nk} & \cdots & x_{Np} \end{bmatrix}$$

aggregate

Distribution-valued data

$$\boldsymbol{X}_{\mathrm{d}} = \begin{bmatrix} \xi_{11} & \cdots & \xi_{1k} & \cdots & \xi_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \xi_{i1} & \cdots & \xi_{ik} & \cdots & \xi_{ip} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \xi_{n1} & \cdots & \xi_{nk} & \cdots & \xi_{np} \end{bmatrix}$$

Distribution-valued



Conservation of Distribution structure

1.2 Distribution-valued data

- **When we use distribution-valued data?**
  - Aggregate Large data to more manageable data
  - Describe objects with several values on a one variable

⋮

- **Advantages of distribution-valued data**

Initial data

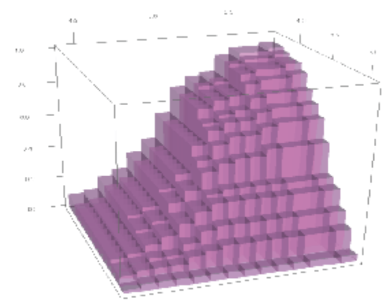
$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_1 1} & x_{n_1 2} & \cdots & x_{n_1 k} & \cdots & x_{n_1 p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nk} & \cdots & x_{Np} \end{bmatrix}$$

aggregate

Distribution-valued data

$$\boldsymbol{X}_{\mathrm{d}} = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_i \\ \vdots \\ \xi_n \end{bmatrix}$$

Joint distribution



1.2 Distribution-valued data

- **When we use distribution-valued data?**
  - Aggregate Large data to more manageable data
  - Describe objects with several values on a one variable

⋮

- **Advantages of distribution-valued data**

Initial data

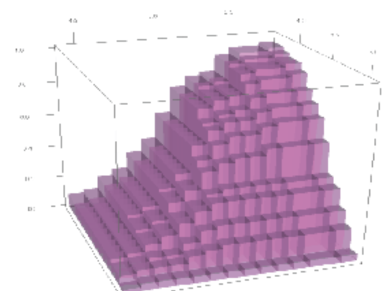
$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_1 1} & x_{n_1 2} & \cdots & x_{n_1 k} & \cdots & x_{n_1 p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ik} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{Nk} & \cdots & x_{Np} \end{bmatrix}$$

aggregate

Distribution-valued data

$$\boldsymbol{X}_{\mathrm{d}} = \begin{bmatrix} \boxed{\xi_1} \\ \vdots \\ \xi_i \\ \vdots \\ \xi_n \end{bmatrix}$$

Joint distribution



We can consider the relationship between variables.

2.1 Dissimilarity measures for distribution-valued data

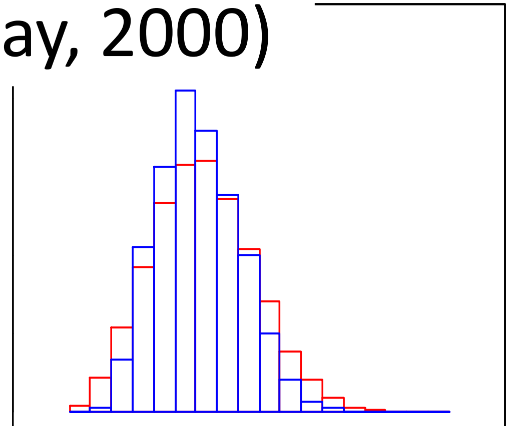
- P, Q : a probability distribution, respectively
- p, q : a density function of P, Q, respectively
- **Dissimilarity measures for density functions**
  - **Kullback-Leibler divergence**
    - Kullback-Leibler information : $I(P \mid Q) = \int \log\left\{\frac{p(x)}{q(x)}\right\} q(x)\, dx$
    - Kullback-Leibler divergence : $J(P, Q) = I(Q \mid P) + I(P \mid Q)$
  - **Minkowski's L^2 distance** (Bock and Diday, 2000)

$$\boxed{d_2(P, Q) = \int (p(x) - q(x))^2\, dx}$$

2.1 Dissimilarity measures for distribution-valued data

- P, Q : a probability distribution, respectively
- p, q : a density function of P, Q, respectively
- **Dissimilarity measures for density functions**
  - **Kullback-Leibler divergence**
    - Kullback-Leibler information : $I(P \mid Q) = \int \log\left\{\frac{p(x)}{q(x)}\right\} q(x)\, dx$
    - Kullback-Leibler divergence : $J(P, Q) = I(Q \mid P) + I(P \mid Q)$
  - **Minkowski's L^2 distance** (Bock and Diday, 2000)

$$\boxed{d_2(P, Q) = \int (p(x) - q(x))^2\, dx}$$

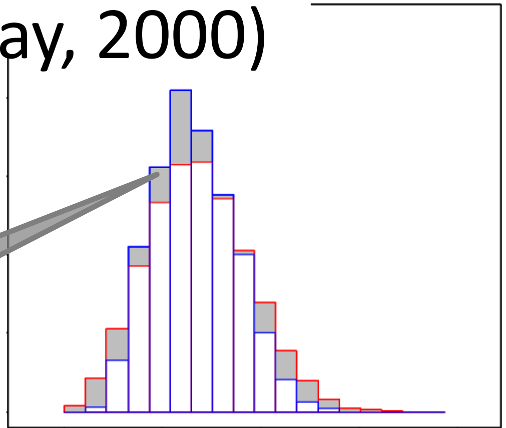


2.1 Dissimilarity measures for distribution-valued data

- P, Q : a probability distribution, respectively
- p, q : a density function of P, Q, respectively
- **Dissimilarity measures for density functions**
  - **Kullback-Leibler divergence**
    - Kullback-Leibler information : $I(P \mid Q) = \int \log\left\{\frac{p(x)}{q(x)}\right\} q(x)\,dx$
    - Kullback-Leibler divergence : $J(P, Q) = I(Q \mid P) + I(P \mid Q)$
  - **Minkowski's L^2 distance** (Bock and Diday, 2000)

$$\boxed{d_2(P, Q) = \int (p(x) - q(x))^2\,dx}$$

The square of this grey area!



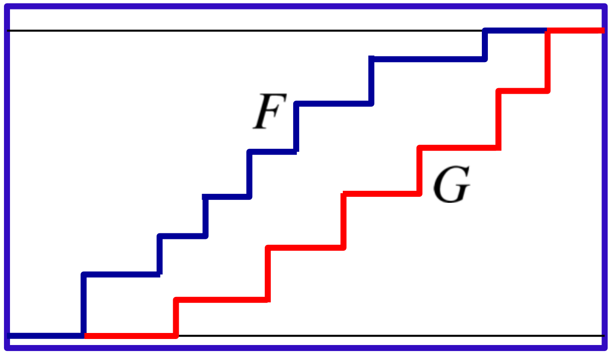
2.1 Dissimilarity measures for distribution-valued data

- P, Q : a probability distribution, respectively
- F, G: a distribution function of P, Q, respectively
- **Dissimilarity measures for distribution functions**
  - **Wasserstein metric**

$$d_W(P, Q) = \int |F(x) - G(x)|\, dx = \int_0^1 |F^{-1}(t) - G^{-1}(t)|\, dt$$

  - **Mallow's distance**

$$d_M(P, Q) = \sqrt{\int_0^1 |F^{-1}(x) - G^{-1}(x)|^2\, dx}$$



2.1 Dissimilarity measures for distribution-valued data

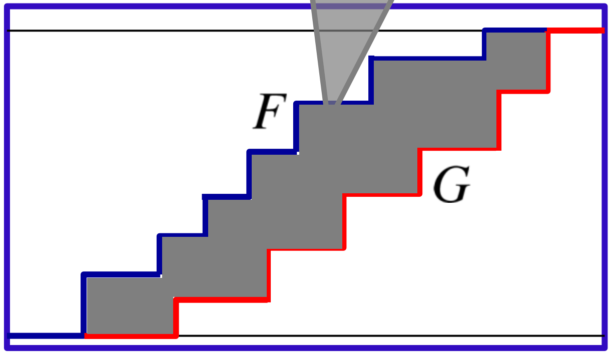
- P, Q : a probability distribution, respectively
- F, G: a distribution function of P, Q, respectively
- **Dissimilarity measures for distribution functions**
  - **Wasserstein metric**

$$d_W(P, Q) = \int |F(x) - G(x)|\, dx = \int_0^1 |F^{-1}(t) - G^{-1}(t)|\, dt$$

This grey area!

  - **Mallow's distance**

$$d_M(P, Q) = \sqrt{\int_0^1 |F^{-1}(x) - G^{-1}(x)|^2\, dx}$$



2.1 Dissimilarity measures for distribution-valued data

- P, Q : a probability distribution, respectively
- F, G: a distribution function of P, Q, respectively
- **Dissimilarity measures for histogram-valued data**

  – **Irpino and Verde (2006) (Irpino et al., 2006)** define a **Wasserstein metric for histogram-valued data**.

  – **Verde and Irpino (2008)** define a **Mahalanobis–Wasserstein distance for histogram–valued data**.

2.2 Clustering for distribution-valued data

- **Irpino and Verde (2006)**
  - **The hierarchical clustering (Ward's method)** of histogram-valued data using **the Wasserstein metric for histogram-valued data**
- **Irpino et al. (2006)**
  - **The dynamic clustering** of histogram-valued data using **the Wasserstein metric for histogram-valued data**
- **Verde and Irpino (2008)**
  - Applying **the Mahalanobis–Wasserstein distance for histogram–valued data** to **the dynamic clustering**

2.2 Clustering for distribution-valued data

- **De Souza et al. (2007)**
  - **Dynamic clustering methods** for **mixed feature-type symbolic data**
  - Pre-processing step for **transforming** mixed feature-type symbolic data **into** **modal symbolic data**
  - Performing the clustering for $\xi = \{(\eta_i, w_i) \mid i = 1, \ldots, n\}$
  - **the transformed data by using the weight vectors**
- **De Carvalho and De Souza (2010)**
  - Unsupervised pattern recognition methods for mixed feature-type symbolic data **using adaptive distances**

2.2 Clustering for distribution-valued data

- **De Souza et al. (2007)**
  - **Dynamic clustering methods** for **mixed feature-type symbolic data**
  - Pre-processing step for **transforming** mixed feature-type symbolic data **into modal symbolic data**
  - Performing the clustering for
  - **the transformed data by using the weight vectors**

$$\xi = \{(\eta_i, w_i) \mid i = 1, \ldots, n\}$$

$$\boldsymbol{w} = (w_1, \ldots, w_n)$$

- **De Carvalho and De Souza (2010)**
  - Unsupervised pattern recognition methods for mixed feature-type symbolic data **using adaptive distances**

2.2 Clustering for distribution-valued data

Here,

- Define **the centroid of a set of distributions**
- Propose **a non-hierarchical clustering** (k-means) **method for more general distribution-valued data by using the centroid dsitribuion**

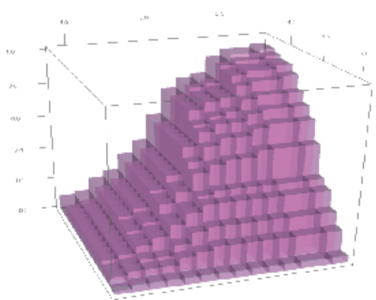
2.2 Clustering for distribution-valued data

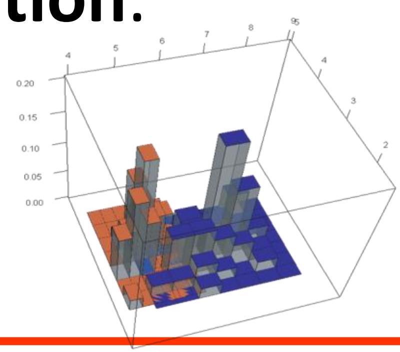
Here,

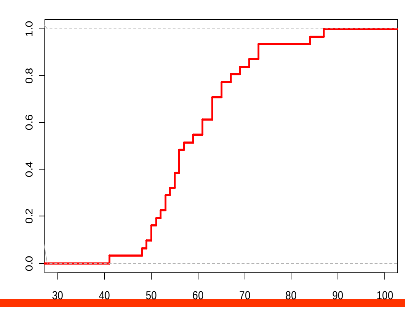
- Define **<u>the centroid of a set of distributions</u>**

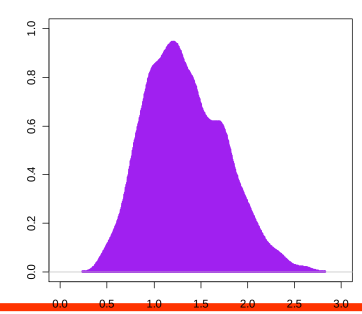
- Propose **a non-hierarchical clustering** (k-means) **method for <u>more general distribution-valued</u>** **<u>data by using the centroid dsitribui</u>**

Represented as joint (or marginal) **distribution function** or **density function**.









3.1 Definition of Centroid distribution

- First,we consider the centroid for distribution-valued data.
  - $\mathscr{P}$: a set of distributions
  - d : a dissimilarity measure on $\mathscr{P}$
  - $P_i \ (i = 1, 2, \ldots, n)$: elements of $\mathscr{P}$
- **Definition of Centroid distribution**

We assume $\mathscr{Q} = \{Q \in \mathscr{P} \mid d(P_i, Q) < \infty \ (i = 1, 2, \ldots, n)\} \neq \emptyset$ and define **the centroid distribution** P_C of distributions P_i, satisfying

$$\boxed{\sum_{i=1}^{n} d^2(P_i, P_C) = \inf_{Q \in \mathscr{P}} \sum_{i=1}^{n} d^2(P_i, Q).}$$

3.1 Definition of Centroid distribution

- First,we consider the centroid for distribution-valued data.
  - $\mathscr{P}$: a set of distributions
  - d : a dissimilarity measure on $\mathscr{P}$
  - $P_i \ (i = 1, 2, \ldots, n)$: elements of $\mathscr{P}$
- **Definition of Centroid distribution**

We assume $\mathscr{Q} = \{Q \in \mathscr{P} \mid d(P_i, Q) < \infty \ (i = 1, 2, \ldots, n)\} \neq \emptyset$ and define **the centroid distribution** P_C of distributions P_i, satisfying

$$\boxed{\sum_{i=1}^{n} d^2(P_i, P_C) = \inf_{Q \in \mathscr{P}} \sum_{i=1}^{n} d^2(P_i, Q).}$$

3.2 Calculation of Centroid distribution

- Here,
  we deal with **Minkowski's L^2 distance**
  for **distribution functions** (or **density functions**).
  - P, Q : a probability distribution, respectively
  - p, q : a density function of P, Q, respectively
  - F, G: a distribution function of P, Q, respectively

$$d_C(P, Q) = \sqrt{\int (F(x) - G(x))^2 \, dx}$$

$$\left(\text{or } d_D(P, Q) = \sqrt{\int (p(x) - q(x))^2 \, dx} \right)$$

3.2 Calculation of Centroid distribution

- Here,
  we deal with **Minkowski's L^2 distance**
  for **distribution functions** (or **density functions**).

– **When we consider marginal distributions,**

  – P, Q : a distribution that has marginal distribution P_j, Q_j , respectively
  – p_j, q_j : a density function of P, Q, respectively
  – F_j, G_j : a distribution function of P, Q, respectively

$$\boxed{d_C(P, Q) = \sqrt{\sum_{j=1}^{r} \int \left(F_j(x) - G_j(x)\right)^2 dx} \left(\text{or } d_D(P, Q) = \sqrt{\sum_{j=1}^{r} \int (p_j(x) - q_j(x))^2 dx} \right)}$$

3.2 Calculation of Centroid distribution

- **The centroid distribution with d_C (or d_D)**
  - $\mathscr{P}_r$: a set of (continuous) distributions on $\mathbb{R}^r$
  - $P_i \ (i = 1,\ 2,\ \ldots,\ n)$: elements of $\mathscr{P}_r$
  - F_i : a distribution function of $P_i \ (i = 1,\ 2,\ \ldots,\ n)$

If $\mathscr{Q} = \{Q \in \mathscr{P} \mid d_C(P_i,\ Q) < \infty \ (i = 1,\ 2,\ \ldots,\ n)\} \neq \emptyset$, then the centroid distribution P_C of P_i is given by the distribution that has the distribution function satisfying

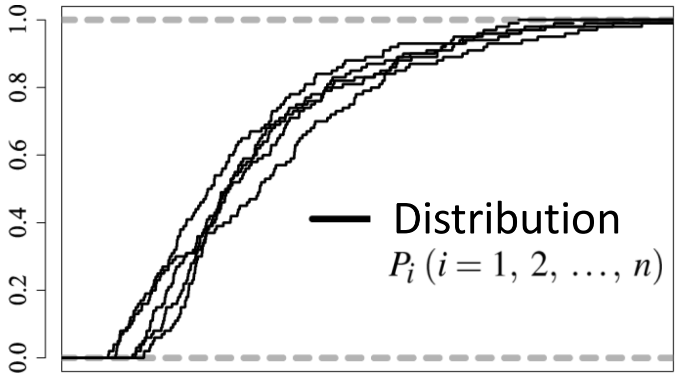
$$\boxed{F_C(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} F_i(\boldsymbol{x}) \qquad (\forall \boldsymbol{x} \in \mathbb{R}^r).}$$

3.2 Calculation of Centroid distribution

- **The centroid distribution with d_C (or d_D)**

If $\mathscr{Q} = \{Q \in \mathscr{P} \mid d_C(P_i, Q) < \infty \ (i = 1, 2, \ldots, n)\} \neq \emptyset$, then the centroid distribution P_C of P_i is given by the distribution that has the distribution function satisfying

$$\boxed{F_C(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} F_i(\boldsymbol{x}) \qquad (\forall \boldsymbol{x} \in \mathbb{R}^r).}$$

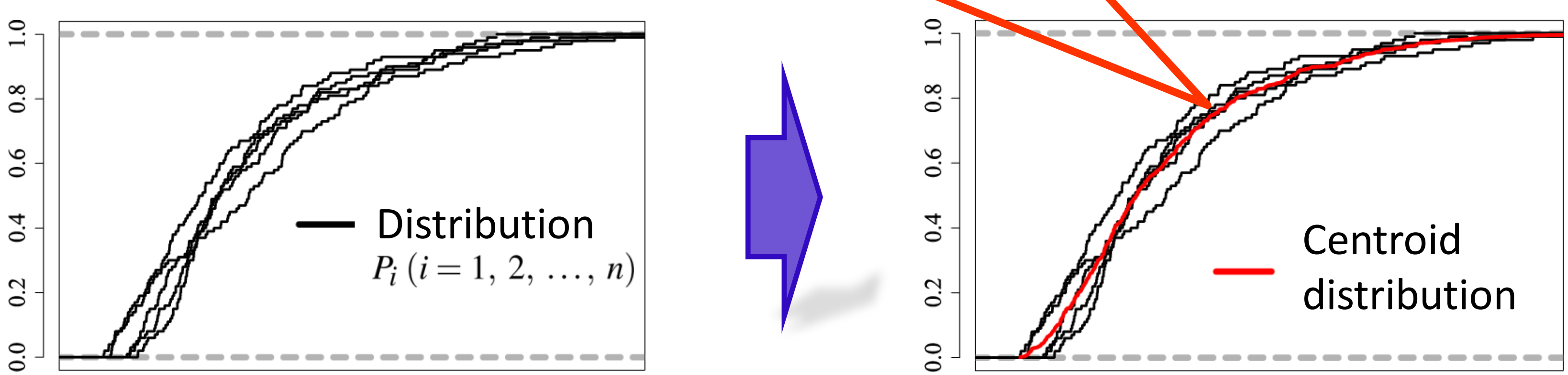


3.2 Calculation of Centroid distribution

- **The centroid distribution with d_C (or d_D)**

If $\mathscr{Q} = \{Q \in \mathscr{P} \mid d_C(P_i, Q) < \infty \ (i = 1, 2, \ldots, n)\} \neq \emptyset$, then the centroid distribution P_C of P_i is given by the distribution that has the distribution function satisfying

$$\boxed{F_C(\boldsymbol{x}) = \frac{1}{n} \sum_{i=1}^{n} F_i(\boldsymbol{x}) \qquad (\forall \boldsymbol{x} \in \mathbb{R}^r).}$$



4.1 Objective function for clustering

- We propose

  "**a non-hierarchical clustering method**" using

  dissimilarity $d_C\,(\mathrm{or}\ d_D)$ and **centroid distribution** P_C **.**

4.1 Objective function for clustering

- We propose

  "**a non-hierarchical clustering method**" using

  dissimilarity d_C (or d_D) and **centroid distribution** P_C **.**

- **Objective function for the clustering**
  - P_j $(j = 1, 2, \ldots, n)$: distributions
  - k : the number of cluster
  - C_i $(i = 1, \ldots, k)$: Clusters constructed by P_j

$$\boxed{Q_C = \sum_{i=1}^{k} \sum_{j \in C_i} d_C^2(P_j, P_{C_i}) \quad \left(\text{or } Q_D = \sum_{i=1}^{k} \sum_{j \in C_i} d_D^2(P_j, P_{C_i})\right).}$$

4.1 Objective function for clustering

- We propose
  "**a non-hierarchical clustering method**" using
  dissimilarity d_C (or d_D) and **centroid distribution P_C .**

- **Objective function for the clustering**
  - $P_j \ (j = 1, 2, \ldots, n)$: distributions
  - k : the number of cluster
  - $C_i \ (i = 1, \ldots, k)$: Clusters constructed by P_j

$$\boxed{Q_C = \sum_{i=1}^{k} \sum_{j \in C_i} d_C^2(P_j, P_{C_i}) \quad \left(\text{or } Q_D = \sum_{i=1}^{k} \sum_{j \in C_i} d_D^2(P_j, P_{C_i}) \right).}$$

4.1 Objective function for clustering

- We propose
  "**a non-hierarchical clustering method**" using
  dissimilarity d_C (or d_D) and **centroid distribution** P_C **.**

- **Objective function for the clustering**
  - P_j $(j = 1, 2, \ldots, n)$: distributions
  - k : the number of cluster
  - C_i $(i = 1, \ldots, k)$: Clusters constructed by P_j

Centroid distribution
of
Cluster C_i $(i = 1, \ldots, k)$

$$Q_C = \sum_{i=1}^{k} \sum_{j \in C_i} d_C^2(P_j, P_{C_i}) \quad \left(\text{or } Q_D = \sum_{i=1}^{k} \sum_{j \in C_i} d_D^2(P_j, P_{C_i}) \right).$$

4.2 Non-hierarchical clustering algorithm

- **Clustering algorithm (*k*-means)**

Step 1: Initial seeds P_{C_j} $(j = 1, 2, \ldots, k)$ are appropriately determined from the objects P_i $(i = 1, 2, \ldots, n)$ described by distributions (e.g. by using random numbers).

4.2 Non-hierarchical clustering algorithm

- **Clustering algorithm (*k*-means)**

Step 1: Initial seeds P_{C_j} $(j = 1, 2, \ldots, k)$ are appropriately determined from the objects P_i $(i = 1, 2, \ldots, n)$ described by distributions (e.g. by using random numbers).

Step 2: Dissimilarity $d_C(P_i, P_{C_j})$(or $d_D(P_i, P_{C_j})$) from seed P_{C_j} to object P_i is evaluated for $i = 1, 2, \cdots, n$; $j = 1, 2, \cdots, k$.

4.2 Non-hierarchical clustering algorithm

- **Clustering algorithm (*k*-means)**

Step 1: Initial seeds P_{C_j} $(j = 1, 2, \ldots, k)$ are appropriately determined from the objects P_i $(i = 1, 2, \ldots, n)$ described by distributions (e.g. by using random numbers).

Step 2: Dissimilarity $d_C(P_i, P_{C_j})$(or $d_D(P_i, P_{C_j})$) from seed P_{C_j} to object P_i is evaluated for $i = 1, 2, \cdots, n$; $j = 1, 2, \cdots, k$.

Step 3: The centroid distribution P_{C_j} of each cluster C_j $(j = 1, 2, \ldots, k)$ is decided as a new seed.

4.2 Non-hierarchical clustering algorithm

- **Clustering algorithm (*k*-means)**

Step 1: Initial seeds P_{C_j} $(j = 1, 2, \ldots, k)$ are appropriately determined from the objects P_i $(i = 1, 2, \ldots, n)$ described by distributions (e.g. by using random numbers).

Step 2: Dissimilarity $d_C(P_i, P_{C_j})$(or $d_D(P_i, P_{C_j})$) from seed P_{C_j} to object P_i is evaluated for $i = 1, 2, \cdots, n$; $j = 1, 2, \cdots, k$.

Step 3: The centroid distribution P_{C_j} of each cluster C_j $(j = 1, 2, \ldots, k)$ is decided as a new seed.

Step 4: Each object is assigned to the nearest seed.

4.2 Non-hierarchical clustering algorithm

- **Clustering algorithm (*k*-means)**

Step 1: Initial seeds P_{C_j} $(j = 1, 2, \ldots, k)$ are appropriately determined from the objects P_i $(i = 1, 2, \ldots, n)$ described by distributions (e.g. by using random numbers).

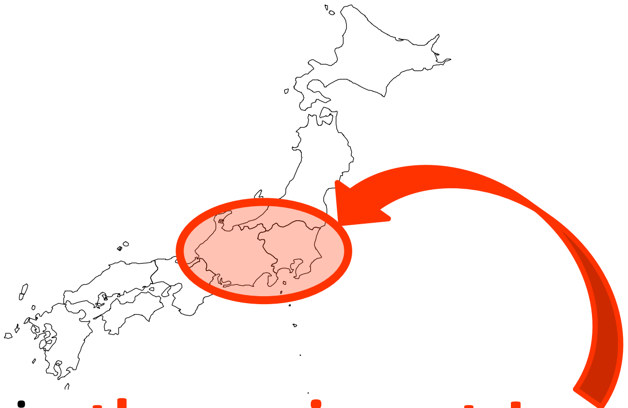
Step 2: Dissimilarity $d_C(P_i, P_{C_j})$(or $d_D(P_i, P_{C_j})$) from seed P_{C_j} to object P_i is evaluated for $i = 1, 2, \cdots, n$; $j = 1, 2, \cdots, k$.

Step 3: The centroid distribution P_{C_j} of each cluster C_j $(j = 1, 2, \ldots, k)$ is decided as a new seed.

Step 4: Each object is assigned to the nearest seed.

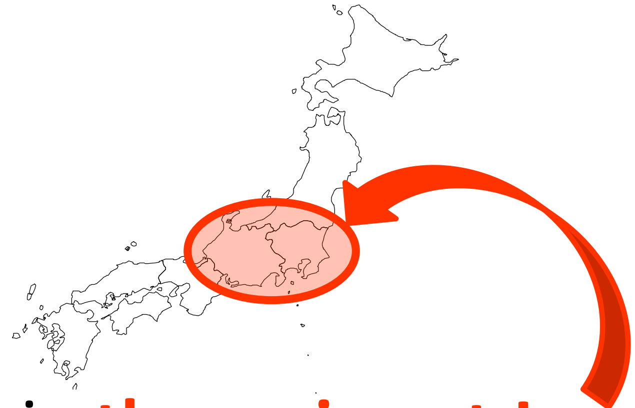
Step 5: If it satisfies a stopping rule (e.g. pre-determined maximum iteration number) then stop, else go to Step 2.

5.1 Applying for the weather data at Japan



- **<u>The weather data</u>**
  - Date : 1 to 31 March 2009
  - Observation points : meteorological observatories in **the region at Japan**
  - Variables : average temperature and humidity (per day)

5.1 Applying for the weather data at Japan

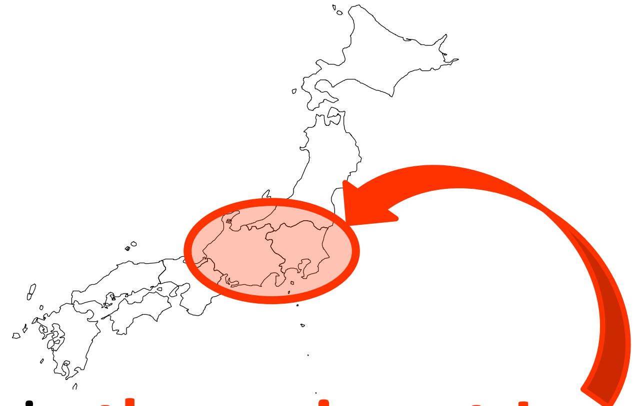


- **The weather data**
  - Date : 1 to 31 March 2009
  - Observation points : meteorological observatories in **the region at Japan**
  - Variables : average temperature and humidity (per day)
- **Transformed distribution-valued data**

Initial data

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n_1 1} & x_{n_1 2} \\ \vdots & \vdots \\ x_{N1} & x_{N2} \end{bmatrix}$$

5.1 Applying for the weather data at Japan



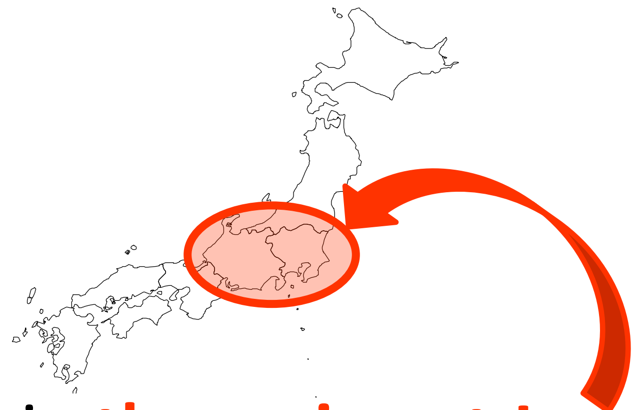
- **The weather data**
  - Date : 1 to 31 March 2009
  - Observation points : meteorological observatories in **the region at Japan**
  - Variables : average temperature and humidity (per day)
- **Transformed distribution-valued data**

Initial data

Observatory A1

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n_1 1} & x_{n_1 2} \\ \vdots & \vdots \\ x_{N1} & x_{N2} \end{bmatrix} \xrightarrow{\text{aggregate}} \boldsymbol{X}_d = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_i \\ \vdots \\ \xi_m \end{bmatrix}$$

5.1 Applying for the weather data at Japan



- **<u>The weather data</u>**
  - Date : 1 to 31 March 2009
  - Observation points :
    meteorological observatories in **the region at Japan**
  - Variables : average temperature and humidity (per day)
- **<u>Transformed distribution-valued data</u>**

Initial data

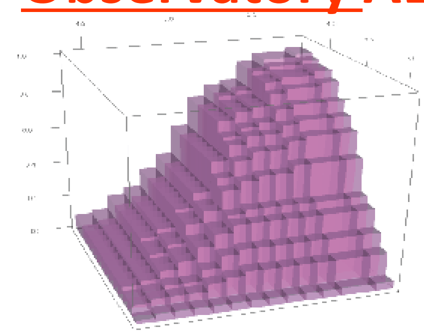
<u>Observatory A1</u>

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n_1 1} & x_{n_1 2} \\ \vdots & \vdots \\ x_{N1} & x_{N2} \end{bmatrix}$$

aggregate

$$\boldsymbol{X}_d = \begin{bmatrix} \xi_1 \\ \vdots \\ \xi_i \\ \vdots \\ \xi_m \end{bmatrix}$$

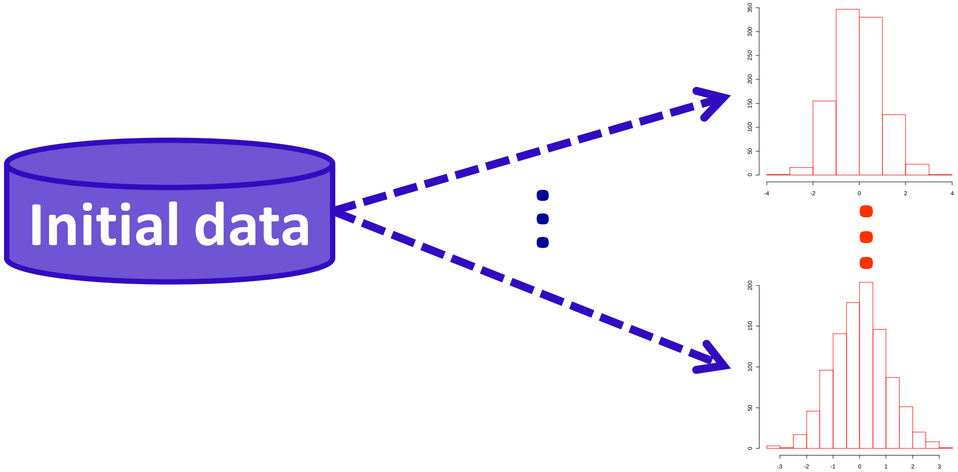
<u>Observatory A1</u>



<u>(joint) Empirical distribution function</u>

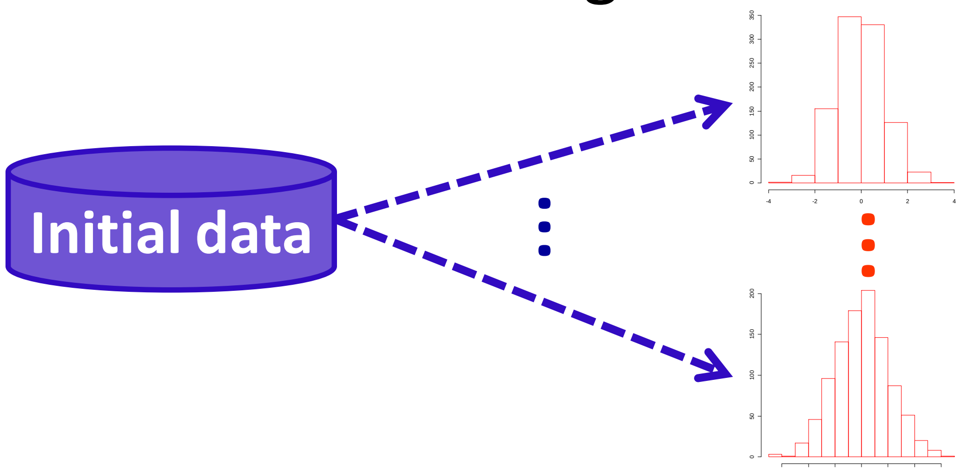
5.1 Applying for the weather data at Japan

- **Transformed distribution-valued data**
  - If we use histograms,
    the number of bins or range of bins affect the result.

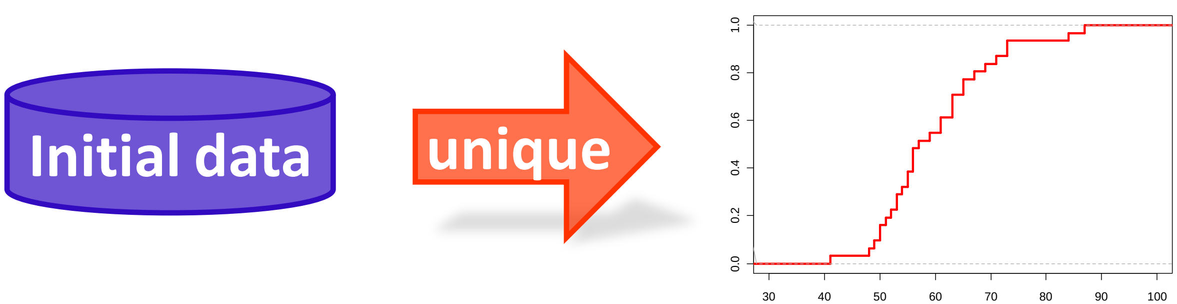


5.1 Applying for the weather data at Japan

- **Transformed distribution-valued data**
  - If we use histograms,
    the number of bins or range of bins affect the result.



  - Here, we use **the empirical distribution function**.



5.1 Applying for the weather data at Japan

- We also apply the classical (k-means) method for the following classical data transformed **by using means**.

Initial data

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n_1 1} & x_{n_1 2} \\ \vdots & \vdots \\ x_{N1} & x_{N2} \end{bmatrix}$$

aggregate

classical data

$$\boldsymbol{X}^* = \begin{bmatrix} \bar{x}_{11} & \bar{x}_{12} \\ \vdots & \vdots \\ \bar{x}_{i1} & \bar{x}_{i2} \\ \vdots & \vdots \\ \bar{x}_{m1} & \bar{x}_{m2} \end{bmatrix}$$

5.1 Applying for the weather data at Japan

- We also apply the classical (k-means) method for the following classical data transformed **by using means**.

Initial data

Observatory A1

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n_1 1} & x_{n_1 2} \\ \vdots & \vdots \\ x_{N1} & x_{N2} \end{bmatrix}$$

aggregate

classical data

Observatory A1

$$\boldsymbol{X}^* = \begin{bmatrix} \bar{x}_{11} & \bar{x}_{12} \\ \vdots & \vdots \\ \bar{x}_{i1} & \bar{x}_{i2} \\ \vdots & \vdots \\ \bar{x}_{m1} & \bar{x}_{m2} \end{bmatrix}$$

5.1 Applying for the weather data at Japan

- We also apply the classical (k-means) method for the following classical data transformed **by using means**.

Initial data

Observatory A1

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n_1 1} & x_{n_1 2} \\ \vdots & \vdots \\ x_{N1} & x_{N2} \end{bmatrix}$$

aggregate

classical data

Observatory A1

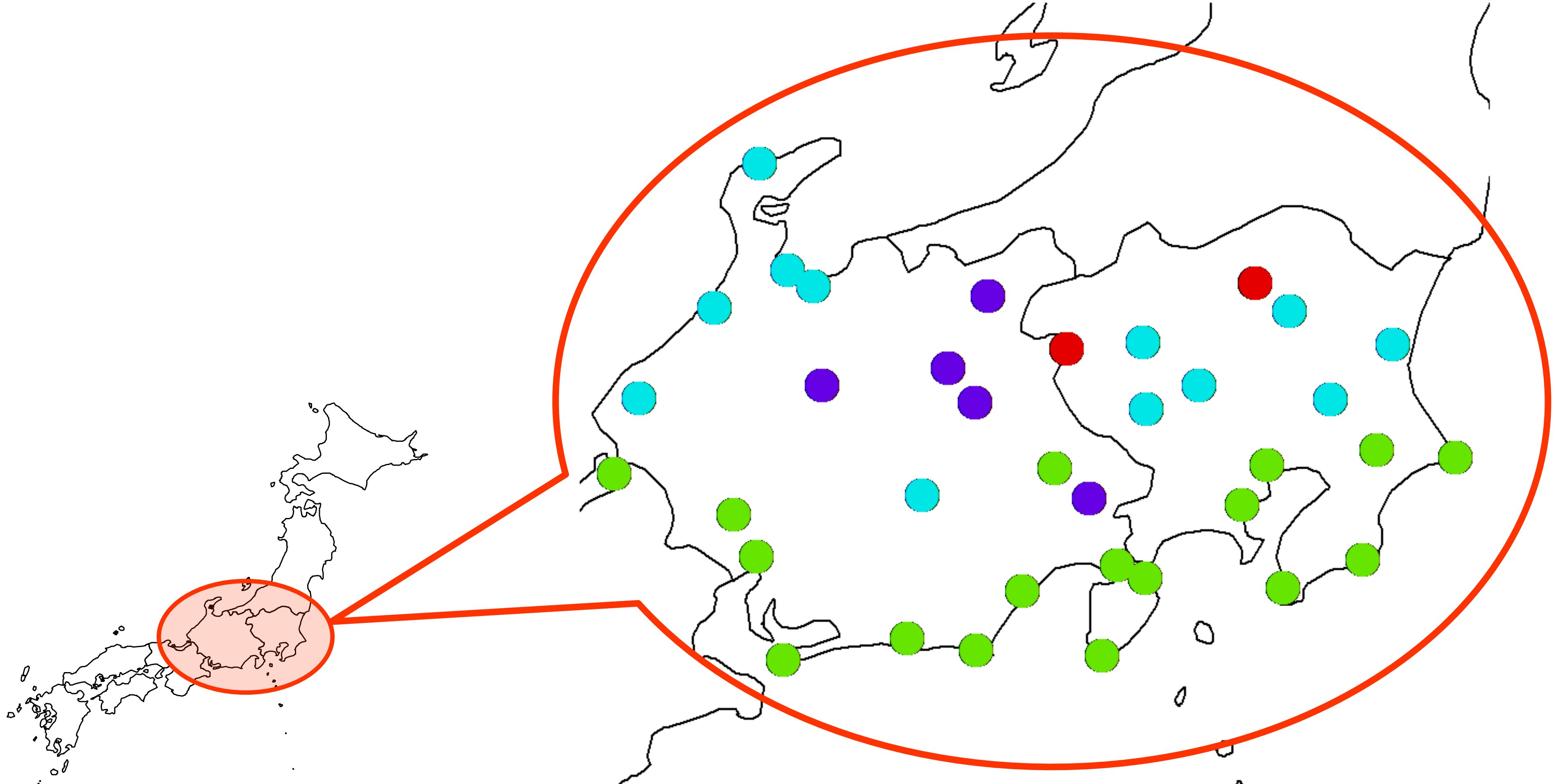
$$\boldsymbol{X}^* = \begin{bmatrix} \bar{x}_{11} & \bar{x}_{12} \\ \vdots & \vdots \\ \bar{x}_{i1} & \bar{x}_{i2} \\ \vdots & \vdots \\ \bar{x}_{m1} & \bar{x}_{m2} \end{bmatrix}$$

- We compare the result of the classical method and the proposal method.

5.1 Applying for the weather data at Japan

- <u>**Result of the classical method (*k*-means)**</u>

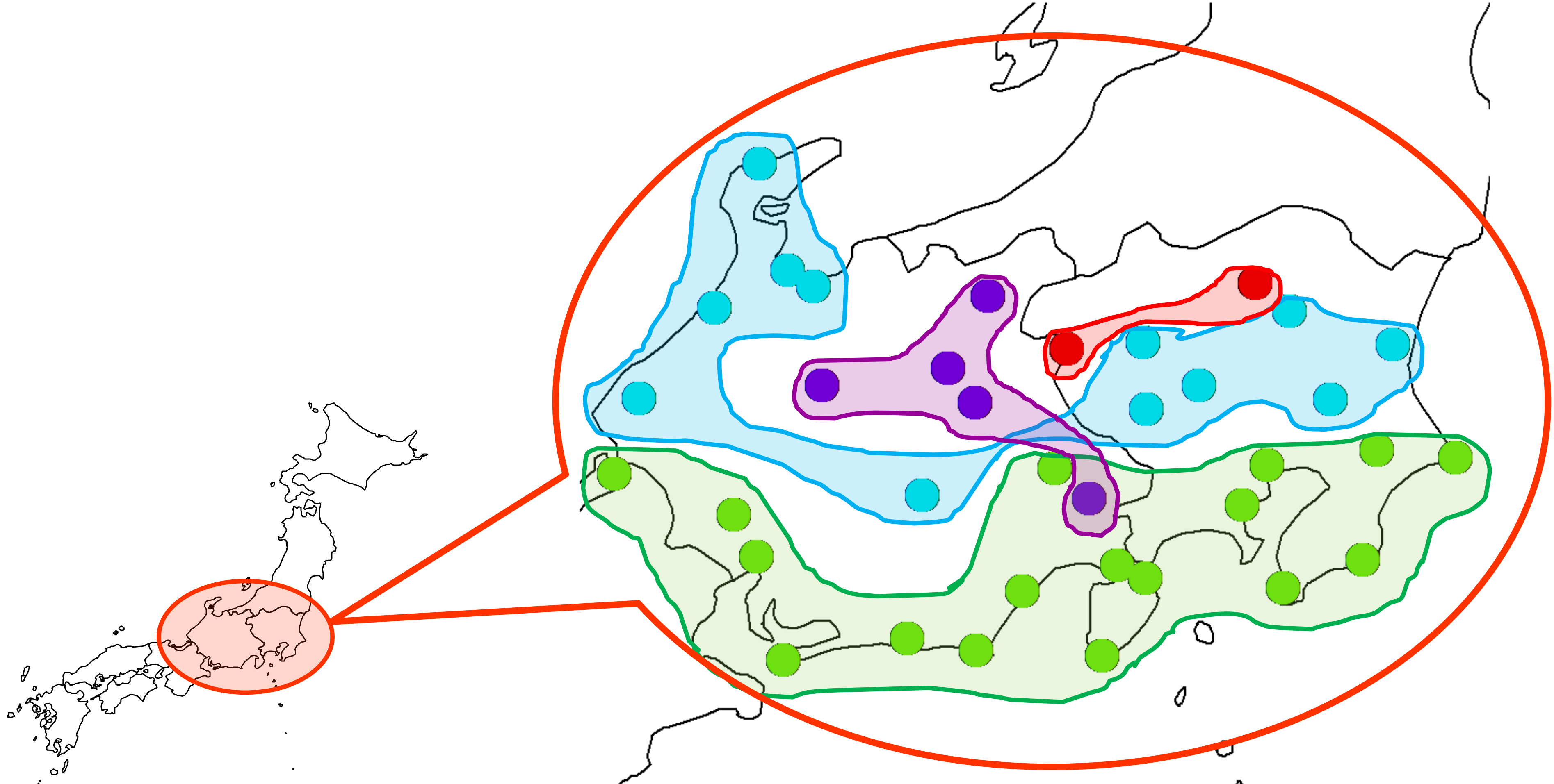
➢ Classified by <u>**a degree of latitude**</u> and <u>**altitude**</u>



5.1 Applying for the weather data at Japan

- **Result of the classical method (*k*-means)**

➢ Classified by **a degree of latitude** and **altitude**

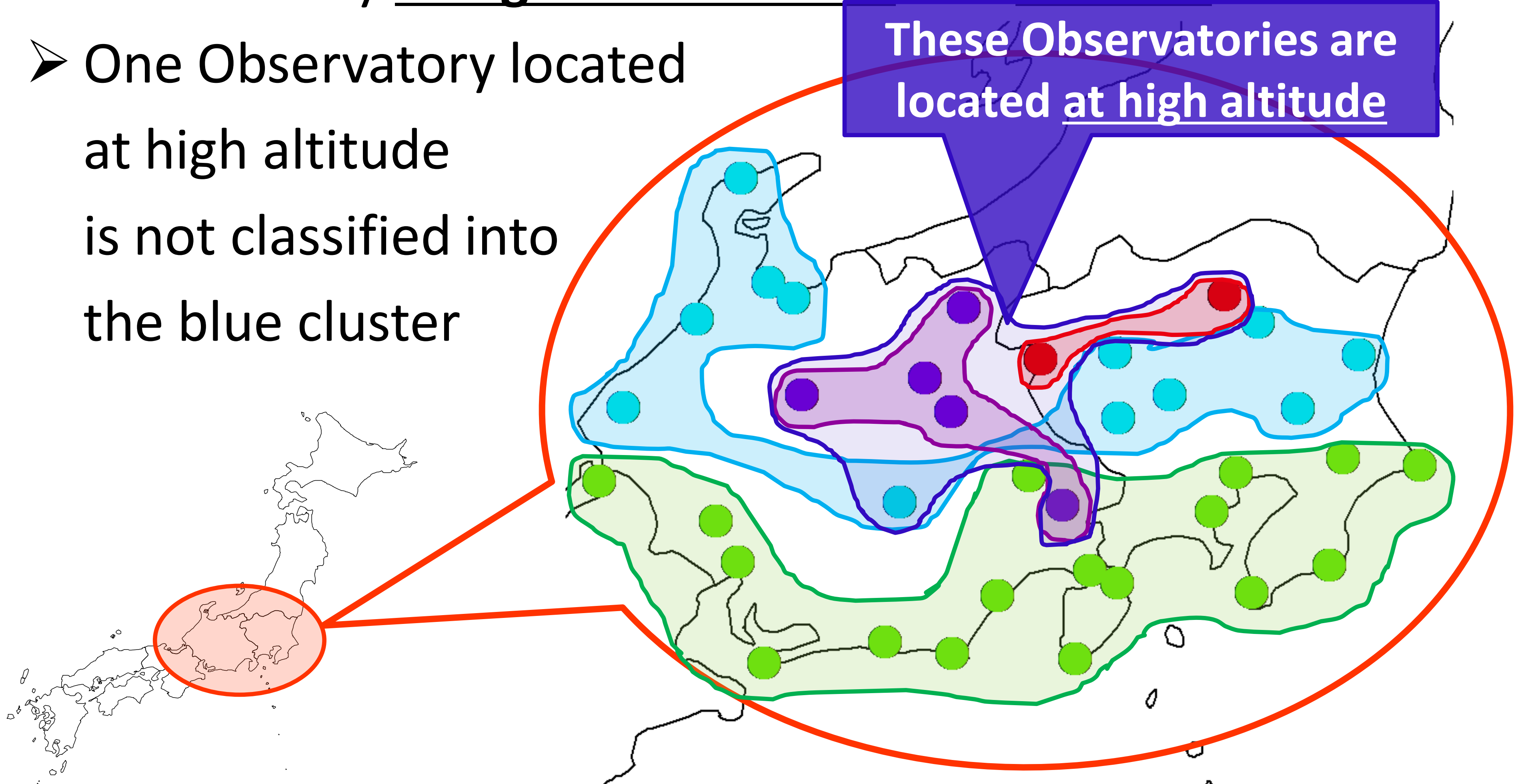


5.1 Applying for the weather data at Japan

- **Result of the classical method (*k*-means)**

➢ Classified by **a degree of latitude** and **altitude**

➢ One Observatory located at high altitude is not classified into the blue cluster

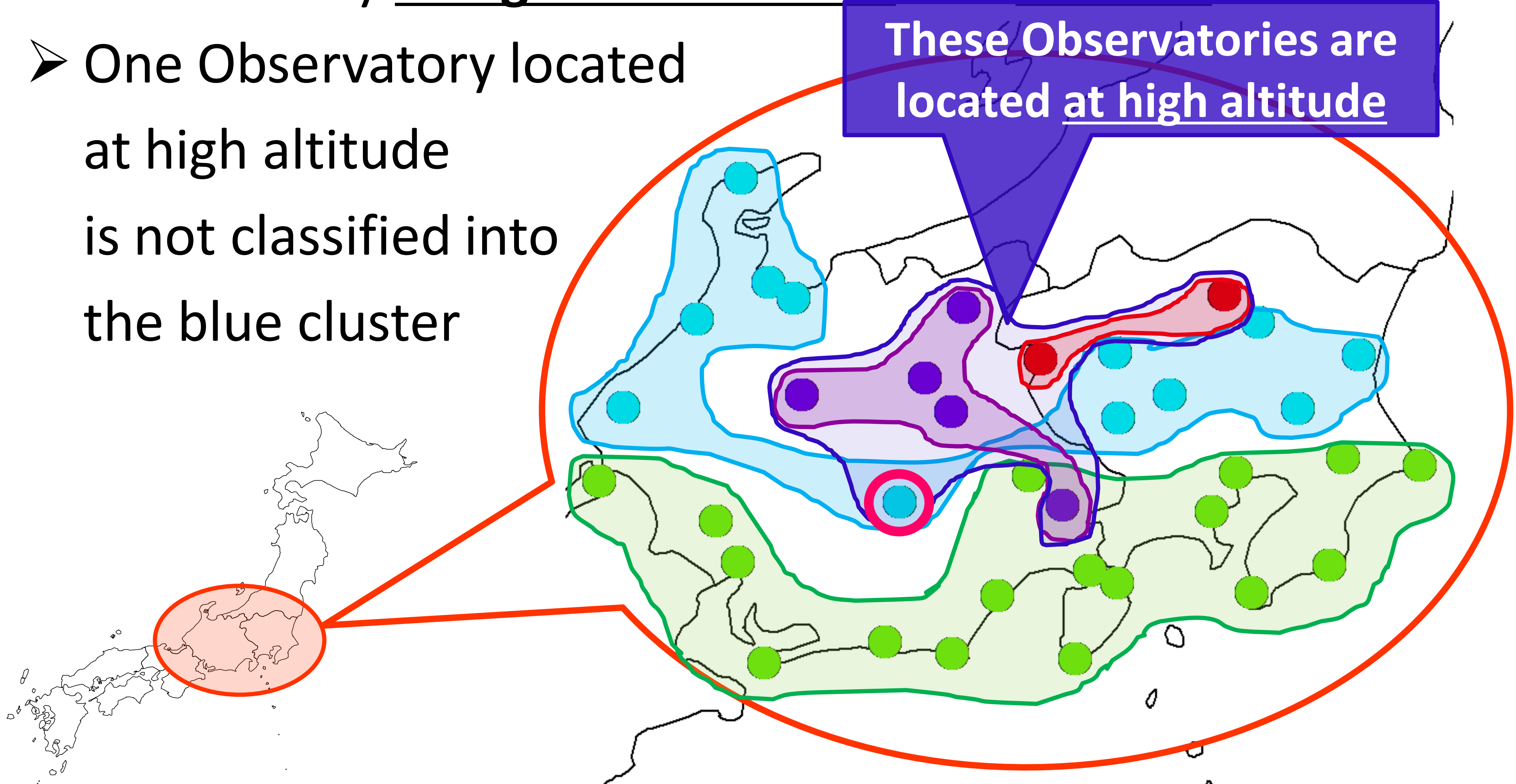


5.1 Applying for the weather data at Japan

- **Result of the classical method (*k*-means)**

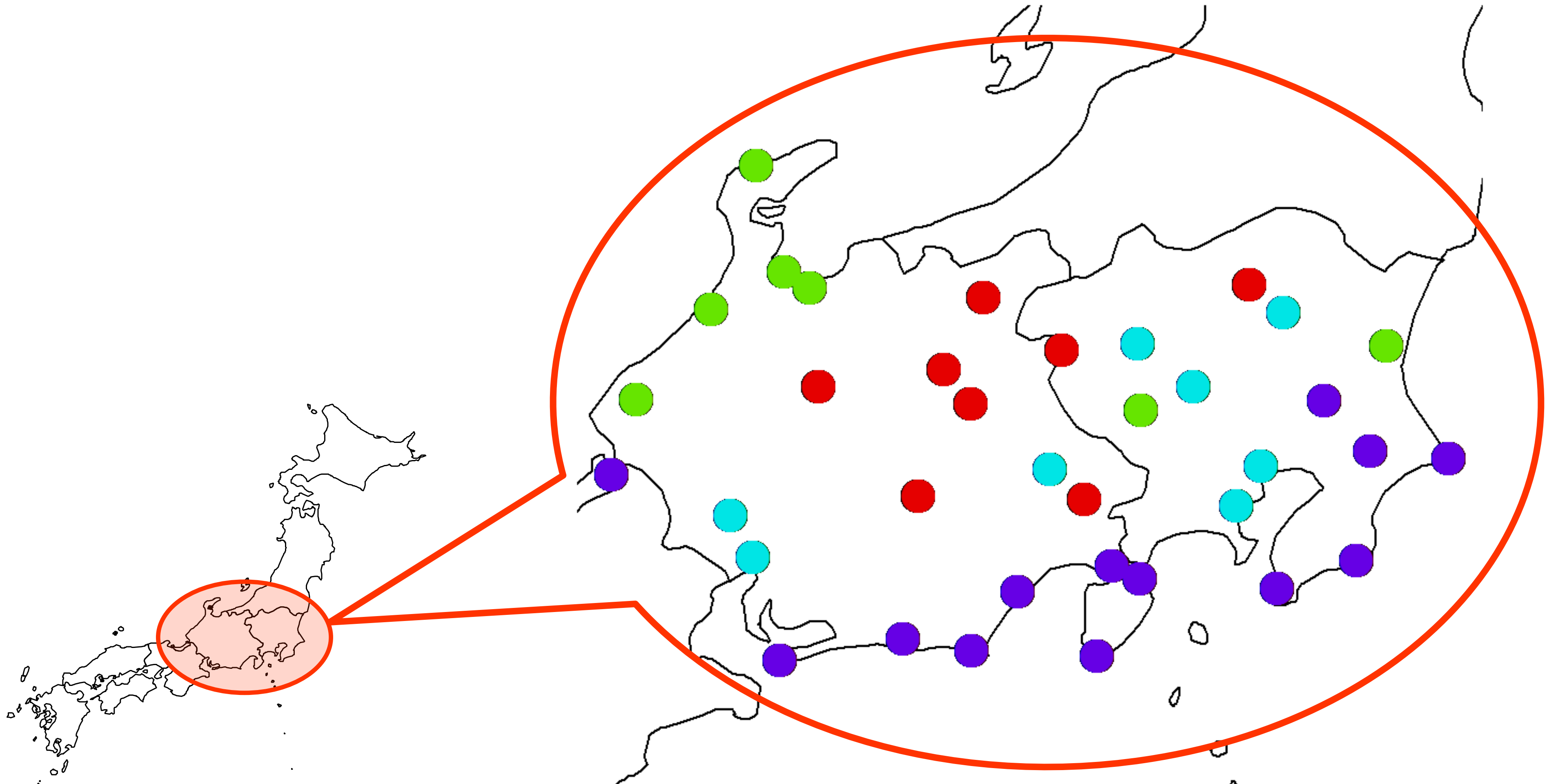
➤ Classified by **a degree of latitude** and **altitude**

➤ One Observatory located at high altitude is not classified into the blue cluster



5.1 Applying for the weather data at Japan

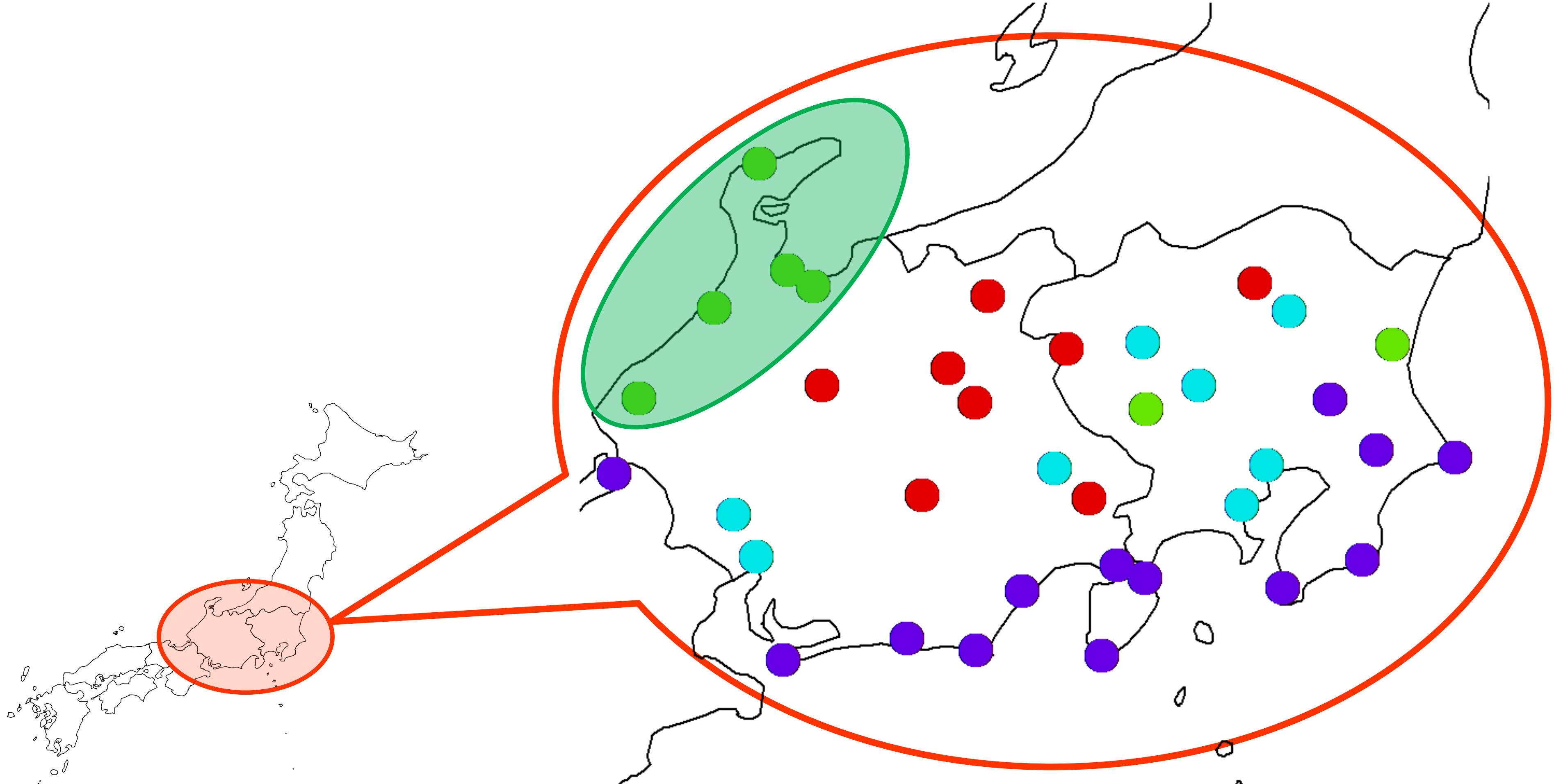
- **Result of the proposal method**



5.1 Applying for the weather data at Japan

- **Result of the proposal method**

➢ The green cluster --- On the East sea side, mainly



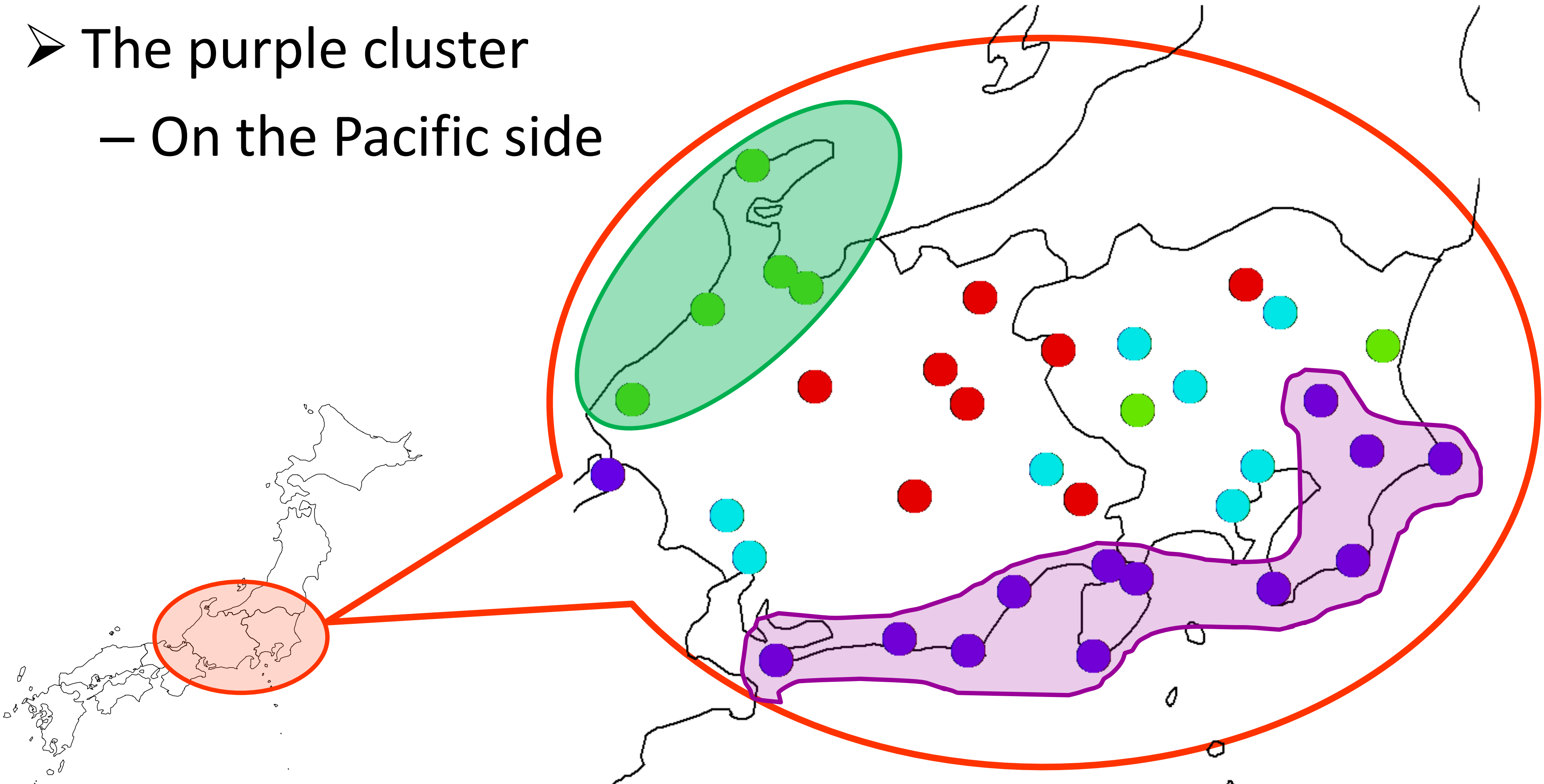
5.1 Applying for the weather data at Japan

- **Result of the proposal method**

➢ The green cluster --- On the East sea side, mainly

➢ The purple cluster

  – On the Pacific side



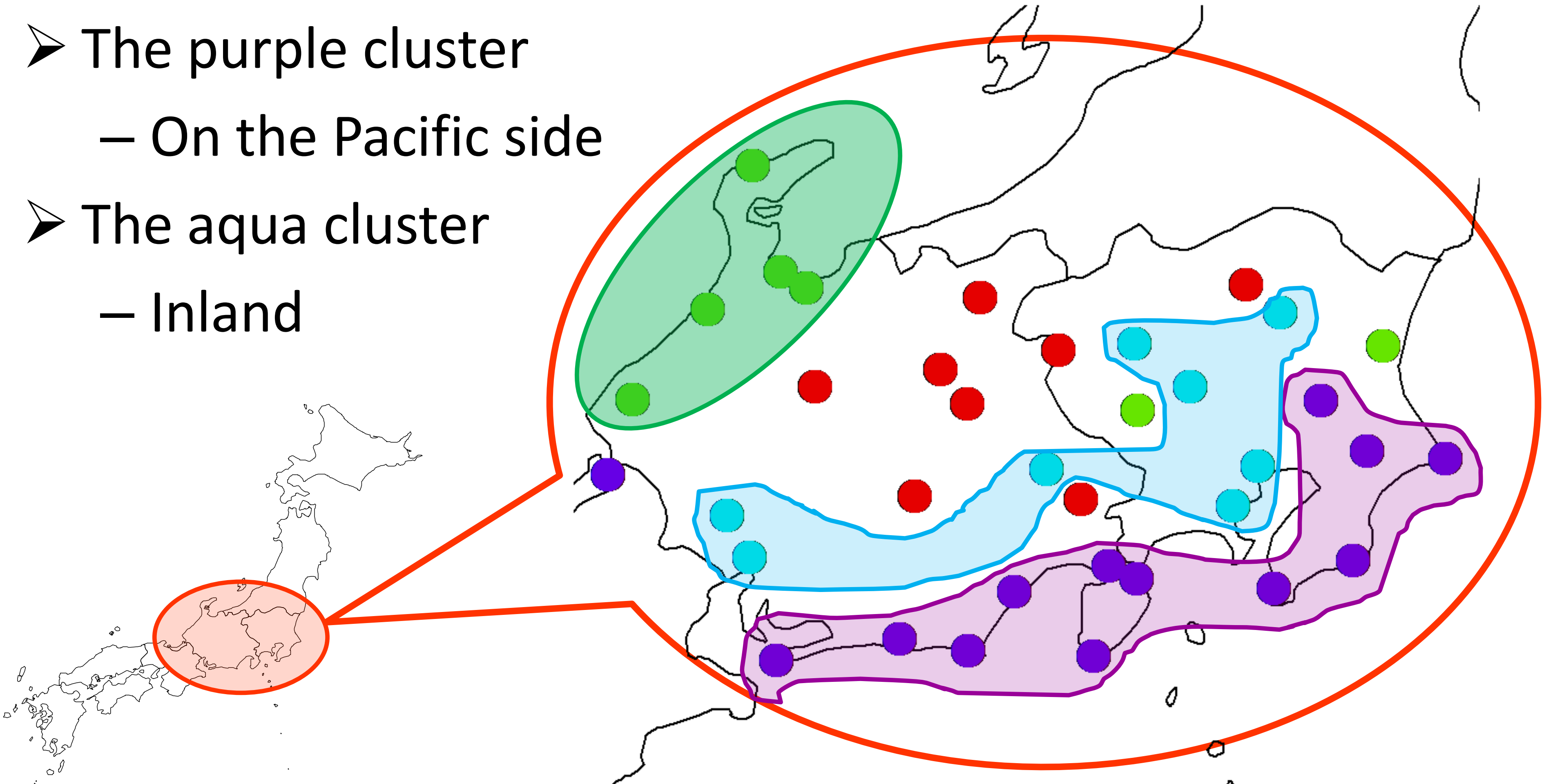
5.1 Applying for the weather data at Japan

- **Result of the proposal method**

➢ The green cluster --- On the East sea side, mainly

➢ The purple cluster
  – On the Pacific side

➢ The aqua cluster
  – Inland



5.1 Applying for the weather data at Japan

- **Result of the proposal method**

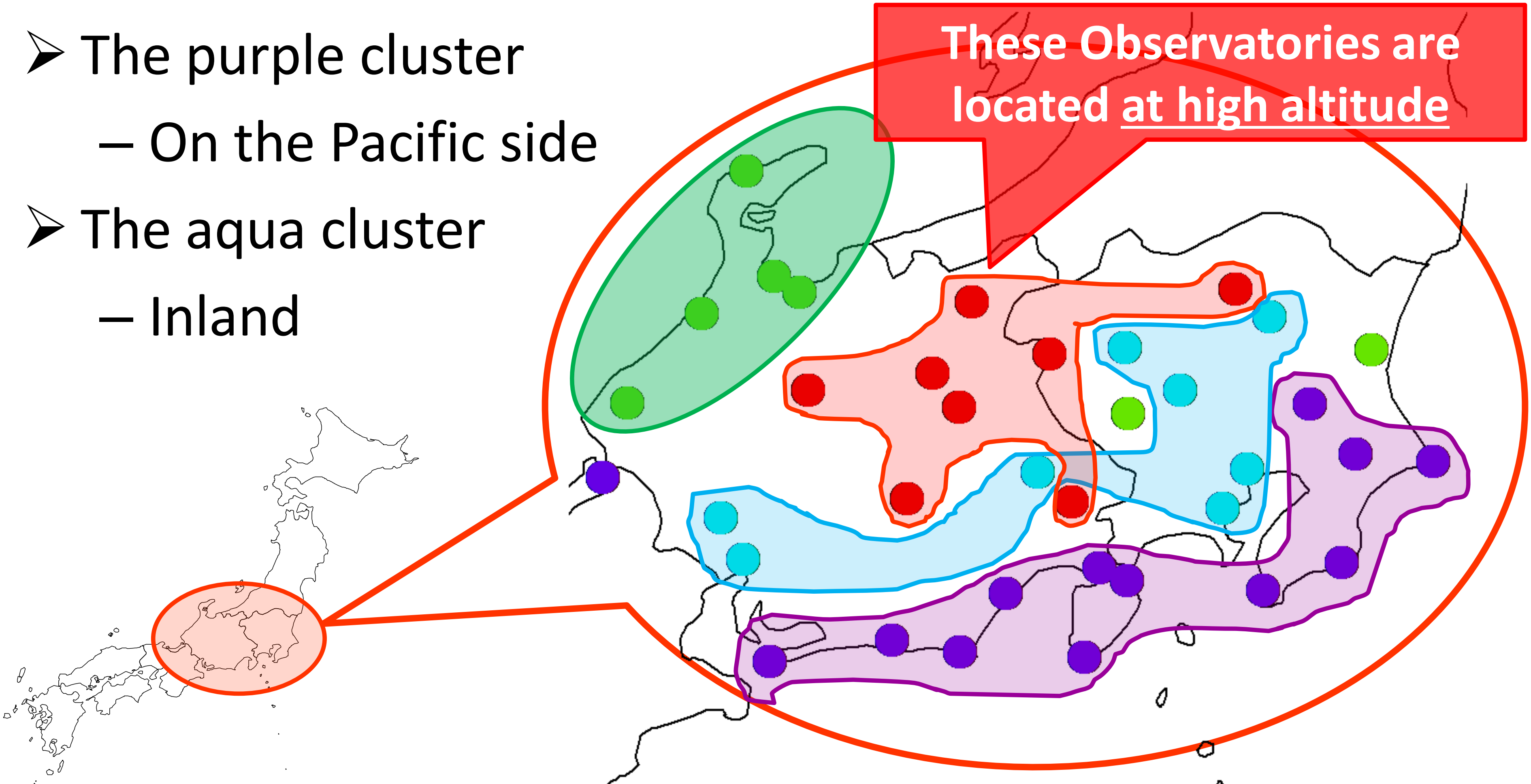
➢ The green cluster --- On the East sea side, mainly

➢ The purple cluster

  – On the Pacific side

➢ The aqua cluster

  – Inland



6. Conclusion

- In this presentation,
  - We define **the centroid dsitribution**
  - and proposed a **non-hierarchical clustering method for more general distribution-valued data** by using the centroid dsitribuion
- Possibility that new classification structures are found by using proposal method.
- For the future study,
  - Calculation of the centroid dsitribution on the other dissimilarity measure.

6. Conclusion

- In this presentation,
  - We define **the centroid dsitribution**
  - and proposed a **non-hierarchical clustering method for** **more general distribution-valued data** by using the centroid dsitribuion
- Possibility that new classification structures are found by using proposal method.
- For the future study,
  - Calculation of the centroid dsitribution on the other dissimilarity measure.

Thank you very much for your attention!

Appendix. Altitude of observatories

- **Altitude of observatories**