CÁC TIÊU CHUẨN CẦN QUAN TÂM ĐỂ ĐÁNH GIÁ KẾT QUẢ PHÂN NHÓM DỮ LIỆU KHI XÂY DỰNG BẢN ĐỒ CHUYÊN ĐỀ
REQUIRED CRITERION FOR DATA CLASSIFICATION IN THEMATIC MAPPING

Lê Minh Vĩnh, Lê Văn Trung(*) và Trần Tấn Lộc(*)

Khoa Địa lý, Đại học Khoa học Xã hội và Nhân Văn, Tp. Hồ Chí Minh, Việt Nam
(*)Bộ môn Địa Tin học, Khoa Kỹ thuật Xây dựng, Đại học Bách Khoa, Tp. Hồ Chí Minh, Việt nam

BẢN TÓM TẮT

Trong khi sử dụng một số phương pháp thể hiện nội dung bản đồ chuyên đề như phân vùng số lượng, biểu đồ, đồ giải..., người làm bản đồ phải thực hiện việc phân nhóm dữ liệu. Cho đến nay, công việc này vẫn được xem là vấn đề của kinh nghiệm. Bài báo đề cập đến các tiêu chuẩn cần quan tâm để đánh giá kết quả một phép phân nhóm dữ liệu nhằm đảm bảo chất lượng truyền thông của các bản đồ chuyên đề .

ABSTRACT

Data classification is one of problems that map makers often confront with when they create thematic maps, especially cloropleth maps… in which map makers have to classify data. Up to now, this has been solved only thanks to experience. This paper concerns about criterion required for data classification to get a suitable result map.

1. ĐẶT VẤN ĐỀ

Khác với việc xây dựng bản đồ địa hình, trong đó nội dung và cách thể hiện đã được quy định chi tiết và cụ thể; khi xây dựng bản đồ chuyên đề, có nhiều vấn đề người làm bản đồ cần giải quyết, trong đó có thể kể đến việc lựa chọn giải pháp thể hiện sao cho hiệu quả.

Một trong những vấn đề của bài toán lựa chọn giải pháp thể hiện nội dung bản đồ chuyên đề là *phân nhóm dữ liệu*: khi làm việc với dữ liệu định lượng (thang khoảng cách và thang hữu tỉ), trong một số phương pháp thể hiện nội dung, ta có nhu cầu phân nhóm các dữ liệu này. Ví dụ:

- Trong phương pháp ký hiệu theo điểm, ký hiệu theo theo tuyến hay phương pháp biểu đồ, kích thước các ký hiệu có thể tỉ lệ với giá trị cụ thể của hiện tượng (thang liên tục) nhưng cũng có thể chỉ thể hiện giá trị đã phân nhóm (thang gián đoạn).
- Trong phương pháp phân vùng số lượng, đồ giải… ta không thể hiện từng giá trị của mỗi vùng mà sẽ gom chúng thành một số nhóm nhất định, tức là phân dữ liệu thành các nhóm để bản đồ nhìn đơn giản và trực quan hơn.

Phân nhóm dữ liệu là việc chia dữ liệu ta cần thể hiện thành từng nhóm và sau đó, tất cả những đối tượng nằm trong một nhóm sẽ chỉ mang một giá trị chung. Việc phân nhóm như vậy nếu với một cách tiếp cận đúng, sẽ không làm sai lệch nội dung mà làm cho bản đồ trở nên đơn giản, dễ nhìn, dễ xử lý hơn. Có thể xem như phân nhóm dữ liệu chính là **một hình thức cuả khái quát hoá** bản đồ mà ***nếu thực hiện đúng*** sẽ đem lại hiệu quả cao trong khai thác bản đồ.

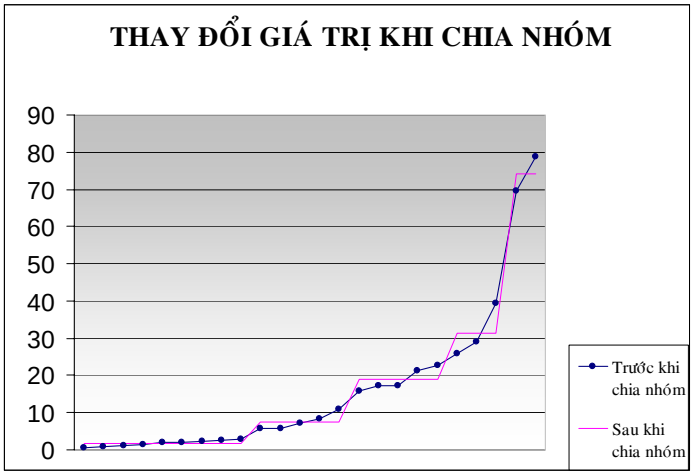
Và ở đây, điều quan trọng là "***nếu thực hiện đúng***"…

Vấn đề phân nhóm dữ liệu đã là mối băn khoăn của những người làm bản đồ từ rất lâu. Người ta đã xây dựng rất nhiều phương pháp chia nhóm: chia đều khoảng, dùng cấp số cộng, chia đều số lượng, dùng độ lệch chuẩn, dùng độ chênh lệch giá trị (natural break), chia tối ưu theo cách của G. Jenk... Với sự ra đời và phát triển công nghệ bản đồ số, vấn đề tính

toán để thực hiện các phép chia không còn là một công việc nặng nhọc, khó khăn nữa. Đa số các phần mềm làm bản đồ đã đưa ra được những phần hỗ trợ, cho phép lựa chọn các phép chia dữ liệu khác nhau và xử lý khá nhanh chóng. Vấn đề còn lại là phải làm rõ, ***thế nào là phép phân nhóm tốt***, bởi với cùng một tập dữ liệu, ta có thể có nhiều cách chia, mỗi cách chia sẽ *vẽ nên một bức tranh rất khác nhau*: chỉ cần thay đổi biên của nhóm, chúng ta sẽ làm cho hai đối tượng được đánh đồng (xếp cùng nhóm) hay làm cho chúng trở thành hai đẳng cấp khác nhau! Ta cần có cơ sở, tiêu chuẩn khách quan cho việc lựa chọn cách phân nhóm cho bản đồ

Trước hết, cần thống nhất rằng không có một kết quả tốt nhất duy nhất đúng cho mọi trường hợp. "*Tốt nhất*" có thể hiểu là "*phù hợp nhất*" với yêu cầu, đặc điểm của bản đồ, đối tượng sử dụng... mà các yếu tố này luôn thay đổi trong mỗi trường hợp cụ thể. Ở đây, ta tạm không xét đến những yêu cầu rất đặc biệt (ví dụ, muốn làm nổi bật một nhóm đối tượng nào đó, muốn phân chia đều số lượng để tiện quản lý…- và khi đó ta sẽ có các phương pháp phân chia riêng) mà thử đưa ra một quan điểm về phép phân nhóm tốt nhất trong điều kiện chung.

Một cách chung nhất, phép phân nhóm tốt là phép phân nhóm sao cho phản ánh sát *thực tế* nhất. *Thực tế* là mỗi đối tượng có một *giá trị riêng*, sau khi phân nhóm, ta đã gán cho những đối tượng trong cùng nhóm một *giá trị chung* (thường được chọn một giá trị có tính đại diện cho cả nhóm), nghĩa là đã có ***sự biến đổi, làm lệch giá trị.không tránh khỏi được***



Hình 1 Giá trị ban đầu bị thay đổi khi phân nhóm

Các nhà làm bản đồ thường cân nhắc rất kỹ và vận dụng các kinh nghiệm cũng như kiến thức chuyên môn khi khi chọn một phương thức phân nhóm dữ liệu nào đó. Kết quả một phép phân nhóm, do đó, thường phản ánh được quan điểm của người làm bản đồ và dễ dàng được chấp nhận. Tuy nhiên, đối với người làm bản đồ không chuyên thì các lựa chọn và kết quả thường chỉ được tiếp nhận một cách *cảm tính* (cảm thấy được) mà không có tiêu chí nào để đánh giá một cách khách quan.

Vì vậy, để đảm bảo chất lượng sản phẩm bản đồ, ta cần có những cách đánh giá kết quả phép phân nhóm khách quan hơn, bằng những tiêu chí cụ thể hơn.

2, ĐÁNH GIÁ MỨC ĐỘ SÁT THỰC TẾ CỦA KẾT QUẢ PHÂN NHÓM

Một cách tổng quát, phép phân nhóm dữ liệu tốt nhất là phép phân chia cho kết quả gần với số liệu ban đầu nhất, phản ánh phân bố thực của dữ liệu tốt nhất. Theo lý thuyết phân tích cụm, điều đó được diễn đạt như sau: ***các đối tượng trong cùng một nhóm càng giống nhau càng tốt và các đối tượng thuộc hai nhóm khác nhau càng khác nhau càng tốt***. Chúng ta sẽ xem xét một số tiêu chuẩn có thể được sử dụng

2.1. Hệ số tương quan

Gọi $x_1, x_2, x_3..., x_N$ là tập hợp các giá trị ban đầu

Sau khi chia thành k nhóm, mỗi nhóm có n_k đối tượng, các đối tượng thuộc cùng một nhóm sẽ được gán cùng một giá trị, tức là ta có một tập mới với các đối tượng

$$\underbrace{x_1, \ldots, x_{n1}}_{\text{Nhóm 1}}, \underbrace{x_{n1+1}, \ldots, x_{n1+n2}}_{\text{Nhóm 2}}, \ldots\ldots\ldots, \underbrace{x_{n1+n2+\ +nk-1} \ldots . x_{n1+n2+\ +nk}}_{\text{Nhóm k}} (= x_N).$$

Gọi $x_{\text{nhóm k}}$ là giá trị dùng gán cho các đối tượng trong nhóm k. Tập dữ liệu mới sẽ có giá trị

$$\underbrace{x_{\text{nhóm 1}}, x_{\text{nhóm 1}}, \ldots, x_{\text{nhóm 1}}}_{n_1 \text{ lần}}, \underbrace{x_{\text{nhóm 2}}, \ldots, x_{\text{nhóm 2}}}_{n_2 \text{ lần}}, \ldots\ldots, \underbrace{x_{\text{nhóm k}}, \ldots, x_{\text{nhóm k}}}_{n_k \text{ lần}}$$

Gọi tập dữ liệu mới này là $Y = \{y_1, y_2, \ldots\ldots, y_N\}$, ta có thể xét "*sự gần giống nhau* " của hai tập dữ liệu X và Y bằng *hệ số tương quan*:

$$r_{xy} = \frac{\sum_{i=1}^{N}(x_i - \bar{x}) * (y_i - \bar{y})}{\sqrt{\sum_{i=1}^{N}(x_i - \bar{x})^2 * \sum_{i=1}^{N}(y_i - \bar{y})^2}}$$

Với $\bar{x}$, $\bar{y}$ lần lượt là giá trị trung bình của tập X và Y

Hệ số tương quan có giá trị $-1 =< r_{xy} =<1$. Giá trị r_{xy} càng gần 1 thì tương quan giữa hai tập càng chặt chẽ (càng giống nhau) tức là phép phân nhóm có kết quả càng sát thực tế

Hệ số này rất quen thuộc và dễ sử dụng, thường được đưa vào trong hầu hết các phần mềm thống kê. Tuy nhiên, hệ số tương quan chưa phản ánh được đầy đủ hiệu quả của phép chia: nó chỉ kiểm tra ***mức độ giống nhau*** của hai bộ dữ liệu (trước và sau khi phân nhóm) nhưng không kiểm tra khía cạnh phân nhóm: ***các đối tượng đã được gom lại một cách hợp lý hay chưa ?***

Vì vậy, ta chỉ nên dùng hệ số này như một chỉ số để **kiểm tra bổ sung**. Nếu r_{xy} càng gần 1 thì xem như kết quả phân nhóm dữ liệu sẽ càng tốt.

2.2. Tiêu chuẩn đánh giá chung của nhà bản đồ học George Jenk

Năm 1971 nhà bản đồ học Jenk G. đã đưa ra các tiêu chuẩn sau [5, trang 71]:

- *Độ sai lệch* của giá trị thể hiện trên bản đồ so với giá trị thật của đối tượng *tại một vị trí cụ thể*. (độ sai lệch bảng-chi tiết)
- *Độ sai lệch "toàn cảnh"* (overview error): độ sai lệch giữa kết quả phản ánh chung trên bản đồ so với giá trị thực
- *Độ sai lệch* giữa giá trị chênh lệch của các đối tượng sau khi đã phân nhóm với độ chênh lệnh giá trị thô giữa các đối tượng

Tiêu chuẩn này khá toàn diện: vừa xét "tính giống nhau" vừa xét cả tính hợp lý trong phân nhóm. Tuy nhiên, các tiêu chuẩn này khó sử dụng vì chỉ ở mức mô tả.

2.3. Tiêu chuẩn đánh giá theo nhà bản đồ học Terry Slocum

Slocum [5, trang 71], đề nghị sử dụng bản đồ dựng 3D (Prism) để minh họa các tiêu chuẩn của Jenk đã đưa ra trên như sau:

- *Độ chênh lệch chiều cao* của các khối lăng trụ trong bản đồ không phân nhóm (unclassed) so với bản đồ đã phân nhóm.
- *Độ chênh lệch về thể tích* (có quan tâm đến phân bố không gian của đối tượng) của các khối lăng trụ trong bản đồ không phân nhóm (unclassed) so với bản đồ đã phân nhóm.

- *Độ chênh lệch* (lấy tỉ số) giữa sai biệt của n cặp có sai biệt cao nhất trong bản đồ không phân nhóm với n cặp hình lăng trụ tương ứng trong bản đồ phân nhóm.

Các tiêu chuẩn này đã cụ thể hơn, khá toàn diện nhưng vẫn rất khó đưa vào tính toán vì các công thức chưa được làm rõ

2.4. Tiêu chí đánh giá theo chỉ số độ lệch của George Jenk

Để có thể tính toán cụ thể các chỉ số, Jenk G. (1974) đã đưa ra chỉ số "*độ phù hợp theo độ lệch tuyệt đối*" (Goodness of absolute deviation fit -GADF) và sau đó là Robinson (1984) đưa ra chỉ số "*độ phù hợp theo phương sai*" (Goodness of variance fit GVF) được tính như sau:

$$GADF = 1 - \frac{\text{Tổng độ lệch tuyệt đối của các đối tượng với trung bình của từng nhóm}}{\text{Tổng độ lệch tuyệt đối của các đối tượng với trung bình của toàn tập dữ liệu}}$$

Tương tự,

$$GVF = 1 - \frac{\text{Tổng phương sai của các đối tượng với trung bình của từng nhóm}}{\text{Tổng phương sai của toàn tập dữ liệu}}$$

Chúng ta vận dụng lý thuyết thống kê để làm rõ bản chất ý nghĩa các chỉ số này và đưa thành công thức để có thể áp dụng tính toán:
Xem một tập dữ liệu có N giá trị (của N đối tượng) là x_1, x_2, x_3..., x_N, để xây dựng bản đồ, ta phân dữ liệu ra thành k nhóm, mỗi nhóm có n_k đối tượng.

Nhóm thứ nhất có các đối tượng: x_{11}, x_{12}, x_{13},...., x_{1n_1} và $\overline{x_1}$ là giá trị trung bình của nhóm thứ nhất.

Nhóm thứ j có các đối tượng: $x_{j1}, x_{j2}, x_{j3}, ..., x_{jn_j}$ và $\overline{x_j}$ là giá trị trung bình của nhóm j

- Theo lý thuyết thống kê, khi các đối tượng trong một tập Y càng giống nhau và gần với giá trị trung bình thì tổng độ lệch bình phương của từng đối tượng (y_i) trong nhóm với trung bình cộng $\overline{y}$ của chúng sẽ càng nhỏ, nghĩa là $\sum_{i=1}^{n}(y_i - \overline{y})^2$ rất nhỏ.
- Vậy, điều kiện "*các đối tượng trong cùng một nhóm càng giống nhau càng tốt"* sẽ được hiểu là tổng độ lệch bình phương trong từng nhóm nhỏ, và từ đó tổng tất cả các độ lệch bình phương trong từng nhóm của k nhóm cũng sẽ nhỏ. Nghĩa là:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2 \text{ -> min}$$

- Một mặt khác, điều kiện "*các đối tượng thuộc hai nhóm khác nhau càng khác nhau càng tốt*" sẽ là tổng bình phương độ lệch giữa trung bình mỗi nhóm với trung bình chung của toàn bộ càng lớn càng tốt

$$\sum_{i=1}^{k}(\overline{x}_i - \overline{X})^2 \text{ -> max}$$

$\overline{X}$ là trung bình của toàn bộ dữ liệu

- Vậy, điều kiện "*các đối tượng trong cùng một nhóm càng giống nhau càng tốt và các đối tượng thuộc hai nhóm khác nhau càng khác nhau càng tốt"*có thể được diễn giải là tỉ số

$$\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \overline{x}_i)^2}{\sum_{i=1}^{k}(\overline{x}_i - \overline{X})^2} \text{ -> min}$$

Tỉ số này sẽ bị rơi vào trường hợp vô nghĩa khi ta phân thành 1 nhóm (mẫu số =0). Vì vậy, ta sử dụng thêm chỉ số xác định *đặc điểm tập trung của toàn bộ dữ liệu* là tổng bình phương độ lệch của mỗi đối tượng với trung bình chung của toàn bộ dữ liệu

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{X})^2$$

- Các phép biến đổi toán học cho ta:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{X})^2 = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2 + \sum_{i=1}^{k}(\overline{x}_i-\overline{X})^2$$

Như vậy, ta dùng có tỉ số đánh giá độ thích hợp là:

$$\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2}{\sum_{i=1}^{k}(\overline{x}_i-\overline{X})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{X})^2}$$

Để có những giá trị quen thuộc với người sử dụng, Jenk G. đề nghị sử dụng hiệu số của 1 với tỉ số nêu trên làm chỉ số "độ phù hợp". Khi đó, ta có :

- Chỉ số bằng 1 khi không hề có sự nhóm gộp nào tức là số nhóm bằng số đối tượng (trường hợp tốt tuyệt đối)
- Chỉ số =0 khi tất cả gộp trong một nhóm (k=1) (trường hợp xấu nhất)
- Chỉ số càng gần 1 thể hiện phép phân nhóm càng tốt

$$\mathbf{1-}\ \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2}{\sum_{i=1}^{k}(\overline{x}_i-\overline{X})^2+\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2} = \mathbf{1}\ -\ \frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{x}_i)^2}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij}-\overline{X})^2}$$

Đây chính là **chỉ số GFV "độ phù hợp theo phương sai".**
Nếu ta dùng "**độ phù hợp theo độ lệch tuyệt đối**" **GADF** thì công thức là**:**

$$1-\frac{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left|x_{ij}-\overline{x}_i\right|}{\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left|x_{ij}-\overline{x}\right|}$$

Như vậy, chỉ số GFV hay GADF đều là các chỉ số có thể dùng để đánh giá kết quả một phép phân nhóm theo yêu cầu " *các đối tượng trong cùng một nhóm càng giống nhau càng tốt và các đối tượng thuộc hai nhóm khác nhau càng khác nhau càng tốt"*. Với các công thức cụ thể và với sự hỗ trợ của máy tính, ta có thể tính các chỉ số này một cách dễ dàng và dựa vào nó để đánh giá kết quả phân nhóm một cách *định lượng, khách quan*

3. CÁC GIÁ TRỊ ĐẶC BIỆT KHI PHÂN NHÓM

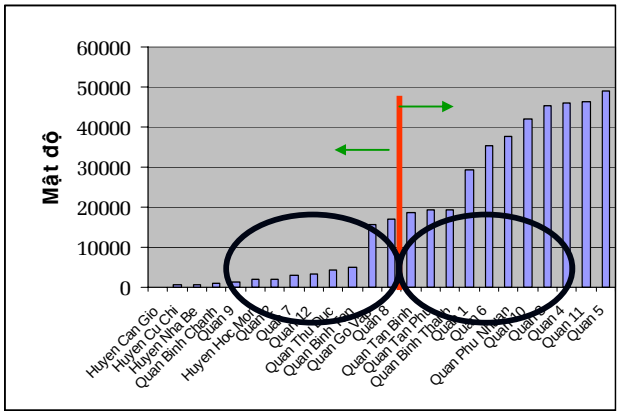
Tuy nhiên, trong thực tế, bên cạnh yêu cầu "sát thực tế", người ta còn phải lưu ý đến những giá trị đặc biệt, có ý nghĩa không chỉ về *lượng* mà còn về *chất*. Đó là những giá trị như chuẩn phấn đấu, ngưỡng phân loại, phân hạng, các giá trị đại diện… Những con số đặc biệt này *cần được phản ảnh trên bản đồ mà trước hết là trên kết quả phân nhóm dữ liệu*

Ví dụ,

- Khi nhóm các giá trị độ cao, những giá trị 0m, 200m, 500m, 1000m… phải là những giá trị được phản ánh vì nó không chỉ là độ cao mà còn đại diện cho các giới hạn về địa hình: đồng bằng, cao nguyên, núi thấp…
- Khi xây dựng bản đồ thể hiện tình trạng nghèo trong thành phố với chỉ số là tỉ lệ hộ nghèo trong quận: nếu phân nhóm dữ liệu về tỉ lệ hộ nghèo của các quận huyện và sử dụng phương pháp *phân nhóm theo độ chênh lệch thực tế* thì hai quận. huyện có tỉ lệ hộ nghèo 4,95% và 5,05% (giá trị chênh lệch rất ít) có thể sẽ được xếp vào cùng một nhóm. Tuy nhiên, nếu chuẩn để xếp loại một quận là nghèo do thành phố quy định là có tỉ lệ hộ nghèo >5% thì ta không thể gộp hai quận này vào cùng nhóm được vì khi đó chúng đã phải thuộc về hai nhóm khác nhau (*nghèo* và *đã thoát nghèo*).

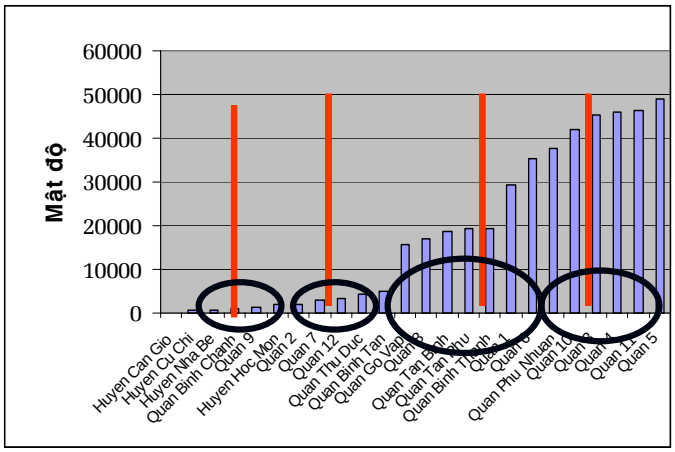
Như vậy, ngay cả khi ta thực hiện phép phân chia có "*độ phù hợp"* tốt thì kết quả cũng vẫn chưa phải là chấp nhận được nếu những khoảng chia ấy *không phù hợp với thực tế, tức là không phản ánh được các giá trị có ý nghĩa*. Các giá trị cần quan tâm ấy là:

- ***Ngưỡng*:** *giá trị có ý nghĩa, được quy định bởi các văn bản hay các chỉ tiêu phấn đấu, giá trị trung bình giai đoạn trước* hay *giá trị trung bình toàn thể khu vực*...Các ngưỡng này được dùng để so sánh sự "vượt qua", "trên mức", "chưa đạt chuẩn".....
  Nếu các giá trị dưới ngưỡng / trên ngưỡng *vẫn còn cần phân nhóm nhỏ hơn* để đánh giá tiếp "đạt rất tốt", "đạt tốt". "chưa đạt", "rất kém"..., ta có **ngưỡng phân cấp.**
  Nếu các giá trị dưới ngưỡng sẽ không cần xét phân nhỏ hơn, ta có **ngưỡng dưới (ngưỡng min);** ngược lại, *nếu giá trị trên ngưỡng sẽ không cần xét phân nhỏ hơn*, ta có **ngưỡng trên (ngưỡng max*).***
  Như vậy, ***giá trị ngưỡng luôn luôn phải là biên của phép phânchia*** dù các giá trị có thể tập trung rất gần xung quanh nó.



Hình2. Giá trị ngưỡng: biên bắt buộc của phép phân nhóm

- ***Giá trị đại diện***: là các giá trị có tính chất phổ biến, đặc trưng cho một ý nghĩa nào đó.
  Như vậy, giá trị đại diện phải là "tâm" của các cụm, nhóm của phép phân chia và đảm bảo là đối tượng nằm trong nhóm của giá trị đại diện nào thì sẽ gần giá trị đại diện ấy nhất.



Hình3. Giá trị đại diện: tâm của các nhóm sau khi phân

4. KẾT LUẬN

Để đánh giá một phép phân nhóm dữ liệu, phải lưu ý đến **cả các giá trị đặc biệt lẫn chỉ số đánh giá**.

Phép phân nhóm tốt nhất là phép phân chia đảm bảo các ngưỡng có ý nghĩa sẽ là biên của nhóm; các đối tượng trong nhóm sẽ tập trung đều quanh giá trị đại diện (nếu có) và đồng thời, cho ra chỉ số độ phù hợp cao nhất có thể trong một số giới hạn về số nhóm (từ 4-9) .

Ngoài ra, còn có một số yếu tố nữa như *mức độ dễ hiểu*, *bảng chú giải rõ ràng, dễ tính toán* ... đôi khi cũng cần được quan tâm thêm và đưa vào như ***yếu tố phụ*** khi cần cân nhắc để lựa chọn giữa các phép phân chia có cùng chỉ số phù hợp.

Xác định cụ thể và rõ ràng về các tiêu chuẩn cần quan tâm khi đánh giá kết quả phân nhóm như vậy sẽ là cơ sở giúp ta xây dựng các lý luận trong việc phân nhóm dữ liệu để đảm bảo tính khách quan và chất lượng của các bản đồ kết quả.

TÀI LIỆU THAM KHẢO

1. Ed Madej: Cartographic design - Using ArcView GIS, On World Press (2001).
2. Kraak M. J., Ormeling F.J.: Cartography-visualization of spatial data, Longman (1995).
3. Mark Monmonier: How to lie with maps, TheUniversity of Chicago Press (1996).
4. Nguyễn Công Khanh: Ứng dụng SPSS for Windows xử lý và phân tích dữ liệu trong các nghiên cứu về giáo dục, y tế, tâm lý và xã hội, NXB Đại học Quốc Gia Hà Nội (2001).
5. Terry A. Slocum: Thematic cartography and visualization, Prentice Hall (1999).
6. Trần Tấn Lộc, Lê Tiến Thuần: Bản đồ học chuyên đề, NXB Đại học Quốc Gia TP. HCM (2004).
7. Võ Văn Huy, Võ Thị Lan, Hoàng Trọng: Ứng dụng SPSS for Windows để xử lý và phân tích dữ kiện nghiên cứu, NXB Khoa học và Kỹ thuật, (1997).