

# THỂ HIỆN VÀ PHÂN NHÓM SỐ LIỆU THỐNG KÊ VỚI WEBGIS

TRẦN TRỌNG ĐỨC

Trường Đại học Bách Khoa – Đại học Quốc gia Thành phố Hồ Chí Minh

## Tóm tắt:

Các số liệu thống kê về dân số, giáo dục, y tế, ... thường được tổng kết và phổ biến trên internet dưới dạng bảng dữ liệu. Bằng cách cho phép chọn lọc, tích hợp và biểu diễn dữ liệu cũng như thực hiện các hoạt động phân nhóm đơn biến hoặc đa biến trên nền bản đồ web, người sử dụng không cần có một phần mềm chuyên dụng vẫn sẽ có một cái nhìn sâu hơn về sự phân bố của dữ liệu theo không gian, cũng như sẽ tìm thấy các nhóm dữ liệu có đặc tính tương đồng. Để minh họa cho ý tưởng này, một hệ thống WebGIS đã được xây dựng trên nền các phần mềm mã nguồn mở GeoServer, OpenLayers. Dữ liệu đưa vào hệ thống là dữ liệu thống kê từ Cục Thống kê Thành phố Hồ Chí Minh và lớp dữ liệu ranh giới hành chính của Tp. Hồ Chí Minh. Các dữ liệu này được tổ chức và lưu trong cơ sở dữ liệu PostgreSQL-PostGIS. Thông qua hệ thống WebGIS này người sử dụng với những hiểu biết nhất định về phân tích dữ liệu không gian hoàn toàn có thể thực hiện hoạt động phân tích dữ liệu thống kê và biểu diễn chúng trên nền bản đồ theo mục đích của mình.

## 1. Đặt vấn đề

Các số liệu thống kê về dân số, giáo dục, y tế,... của các quận huyện thuộc thành phố Hồ Chí Minh có thể tìm thấy trong Niên giám thống kê hàng năm được biên tập và xuất bản bởi Cục Thống kê Thành phố Hồ Chí Minh (2015) hoặc được phổ biến tại website của Cục Thống kê Thành phố Hồ Chí Minh (<http://www.pso.hochiminhcity.gov.vn/web/guest/niengiamthongke-nam2015>). Các số liệu thống kê được tổ chức theo chủ đề - ví dụ *Dân số và lao động*, *Giáo dục*, *Y tế*,... - và thể hiện dưới dạng bảng dữ liệu. Mỗi dòng là số liệu của một quận/huyện. Mỗi cột thể hiện 1 chỉ tiêu đã đo lường được của quận đó, ví dụ dân số. Mặc dầu các số liệu này rất có ích trong cung cấp cho người đọc thông tin về đặc tính hoặc sự phát triển kinh tế xã hội của từng quận huyện. Nhưng sự thể hiện dạng bảng này hoàn toàn không cho thấy kiểu mẫu không gian của sự phát triển của một đặc tính kinh tế xã hội nào đó hoặc không cho thấy sự tương tác theo không gian giữa các đặc tính kinh tế xã hội giữa các quận huyện. Trong trường hợp đơn giản, người sử dụng có thể sẽ rất muốn nhìn thấy một bản đồ hành chính trên đó cho phép thể hiện đặc tính thống kê của

quận huyện theo màu hoặc cho phép phân nhóm các quận huyện theo một đặc tính thống kê nào đó, ví dụ theo số lượng cơ sở y tế. Thông qua bản đồ phân nhóm này, người sử dụng có thể nhận ra được nhóm quận nào có số lượng cơ sở y tế nhiều, trung bình hoặc ít dưới dạng màu sắc, ký hiệu khác nhau một cách trực quan... Trong trường hợp phức tạp hơn, các phương pháp phân nhóm đa biến, ví dụ *k-means*, có thể được sử dụng để so sánh sự phát triển kinh tế - xã hội của các đơn vị hành chính và nhận dạng các biến chung ảnh hưởng đến sự phát triển của các đơn vị hành chính này (Ieva Brauksa 2013).

Việc phân chia các nhóm dữ liệu cũng nên do chính người xem quyết định. Việc tạo ra các bản đồ chuyên đề từ số liệu điều tra kinh tế xã hội thường được thực hiện bởi các chuyên gia sử dụng *Hệ thống thông tin địa lý nền Desktop*. Tuy nhiên sự phát triển của Internet cùng với công nghệ Web đã cho phép người sử dụng dễ dàng tiếp cận đến các số liệu thống kê kinh tế xã hội trên nền bản đồ thông qua các hệ thống WebGIS. Một trong số các hệ thống WebGIS có thể kể đến là DataShine Census (Oliver O'Brien và James Cheshire 2016). DataShine Census cho phép người sử dụng tự chọn để xem các bản đồ

Ngày nhận bài: 23/5/2018, ngày chuyên phân biên: 25/5/2018, ngày chấp nhận phân biên: 04/6/2018, ngày chấp nhận đăng: 08/6/2018

chuyên đề xây dựng sẵn thể hiện các số liệu kinh tế xã hội của các thành phố thuộc các nước thuộc Liên hiệp Anh. Cho mỗi bản đồ chuyên đề đã thể hiện, hệ thống tự động phân nhóm dữ liệu ra làm 8 nhóm. Hạn chế của hệ thống WebGIS này là người sử dụng không có chọn lựa về số nhóm cũng như lựa chọn phương pháp phân nhóm dữ liệu. Nhằm cho phép người sử dụng linh động hơn trong việc khảo sát và phân tích dữ liệu thống kê, hệ thống WebGIS thực nghiệm đã được xây dựng và trình bày trong bài báo này. Hệ thống này cho phép người sử dụng tự khảo sát và phân tích dữ liệu thống kê căn cứ trên các chỉ tiêu thống kê đã công bố bởi Cục Thống kê Tp. Hồ Chí Minh. Người sử dụng tự chọn năm công bố, chọn chủ đề và chọn chỉ tiêu thống kê để biểu diễn trên nền bản đồ hành chính. Người sử dụng tự chọn phương pháp phân nhóm *đơn biến* hay *đa biến* để tiến hành phân tích dữ liệu.

## 2. Thiết kế hệ thống WebGIS

Với mục đích thử nghiệm khả năng thể hiện và phân tích dữ liệu thống kê trên nền bản đồ thông qua internet, hệ thống WebGIS đã được thiết kế với các thành phần được mô tả dưới đây.

### 2.1. Kiến trúc tổng quát của hệ thống WebGIS

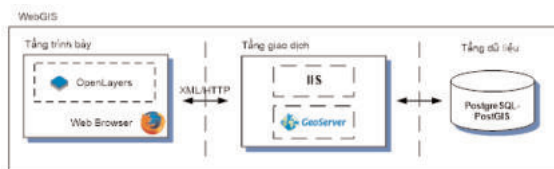
Hệ thống WebGIS xây dựng được thiết kế theo mô hình kiến trúc 3 tầng như hình 1:

■ **Tầng trình bày:** đơn thuần là một trình duyệt web, ví dụ Mozilla Firefox, Internet Explorer... để mở ứng dụng web có thiết kế định sẵn. Ứng dụng Web được viết bằng công nghệ chuẩn, sử dụng thư viện mã nguồn mở Javascript OpenLayers trong truy vấn và hiển thị thông tin bản đồ theo các chuẩn định dạng WMS/WFS. Bên cạnh các chức năng cơ bản thường gặp khi làm việc với bản đồ, như phóng to, thu nhỏ, rê bản đồ, truy vấn thông tin... còn có các chức năng cho phép chọn biểu diễn và phân nhóm dữ liệu thống kê. Thiết kế chi tiết của chức năng phân nhóm dữ liệu sẽ được bàn chi tiết ở phần sau.

■ **Tầng giao dịch:** sử dụng Internet

Information Services (IIS) Web Server tạo bởi Microsoft và phần mềm mã nguồn mở GeoServer có chức năng như một Map Server để cung cấp các dữ liệu và dịch vụ liên quan đến bản đồ theo các chuẩn dịch vụ WMS/WFS. GeoServer được xây dựng dựa trên công nghệ Java Servlet cũng như Spring FrameWork và chạy trên nền mặc định Jetty.

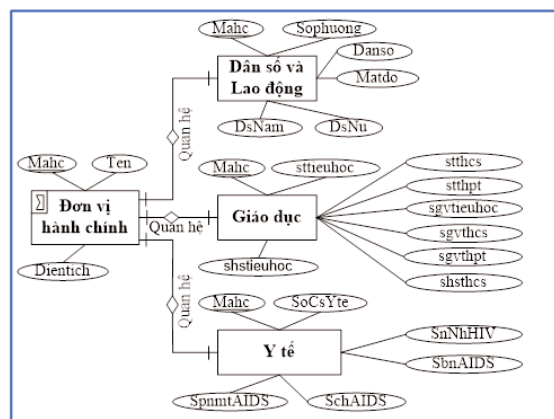
■ **Tầng dữ liệu:** Cho mục đích của nghiên cứu này, dữ liệu được đưa vào hệ thống bao gồm dữ liệu ranh giới hành chính của Tp. Hồ Chí Minh và dữ liệu thống kê lấy từ Niên giám thống kê của Cục Thống kê Thành phố Hồ Chí Minh (2015). Dữ liệu được lưu trong cơ sở dữ liệu PostgreSQL-PostGIS.



Hình 1: Kiến trúc của hệ thống WebGIS

### 2.2. Thiết kế cơ sở dữ liệu

Trên cơ sở i) phân tích nhu cầu thể hiện dữ liệu thống kê trên nền bản đồ và ii) cách thức tổ chức dữ liệu thống kê theo năm, chủ đề, và theo dạng bảng dữ liệu được xuất bản trong niên giám thống kê, các đối tượng cần quản lý và quan hệ giữa các đối tượng được thiết kế và xây dựng. Do khuôn khổ bài báo có giới hạn, nên chỉ một phần cơ sở dữ liệu được minh họa trong hình 2.



Hình 2: Lược đồ quan hệ giữa 1 số đối tượng trong hệ thống

Trong cơ sở dữ liệu, đối tượng “đơn vị hành chính” được tổ chức thành lớp dữ liệu dạng vùng với các thuộc tính: mã đơn vị hành chính (Mahc), tên (Ten), và diện tích (Dientich).

Số liệu thống kê được tổ chức dưới dạng bảng độc lập cho từng năm thống kê, bao gồm:

- Bảng dữ liệu “Dân số và Lao động” với các thuộc tính: mã hành chính (Mahc), Số phường (sophuong), dân số trung bình (Danso), mật độ dân số (Matdo), dân số nam (DsNam), dân số nữ (DsNu).

- Bảng dữ liệu “Giáo dục” với các thuộc tính: mã hành chính (Mahc), số trường tiểu học (sttieuhoc), số trường trung học cơ sở (stthcs), số trường trung học phổ thông (stthpt) số giáo viên tiểu học (sgvtieuhoc), số giáo viên trung học cơ sở (sgvthcs), số giáo viên trung học phổ thông (sgvthpt), số học sinh tiểu học (shstieuhoc), số học sinh trung học cơ sở (shstthcs), số học sinh trung học phổ thông (shstthpt) và

- Bảng dữ liệu “y tế” với các thuộc tính: mã hành chính (Mahc), số cơ sở y tế (SoCsYte), số người nhiễm HIV (SnNhHIV), số bệnh nhân AIDS (SbnAIDS), số chết do AIDS (SchAIDS), số phụ nữ mang thai từ 15 – 25 tuổi có HIV (SpnmtAIDS).

Quan hệ giữa các đối tượng thuộc lớp dữ liệu hình thể “đơn vị hành chính” và các đối tượng trong bảng dữ liệu độc lập là quan hệ 1-1 và dựa vào thuộc tính chung là mã đơn vị hành chính.

### 2.3. Thiết kế chức năng

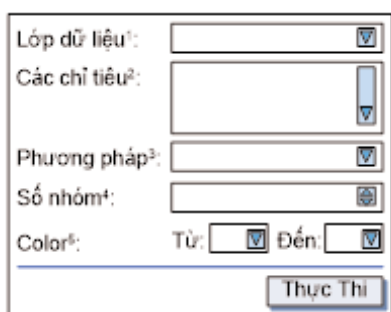
Trong một hệ thống WebGIS sẽ có nhiều nhóm chức năng khác nhau. Trong phạm vi của nghiên cứu này, chỉ nhóm chức năng phân nhóm dữ liệu được thảo luận chi tiết.

- Chức năng phân nhóm dữ liệu thống kê dựa trên một chỉ tiêu hoặc một biến dữ liệu được thiết kế với giao diện như hình 3.

Hình 3: Giao diện công cụ phân nhóm dữ liệu đơn biến

Người sử dụng có thể chọn năm<sup>1</sup> cần biểu diễn, chuyên đề<sup>2</sup> để biểu diễn. Các chuyên đề có thể chọn bao gồm *Dân số và Lao động*, *Giáo dục*, *Y tế*,... Trong mỗi chuyên đề được chọn người sử dụng được cung cấp một danh sách các chỉ tiêu<sup>3</sup>. Danh sách các chỉ tiêu chính là tên của các trường thuộc tính có trong bảng dữ liệu (chuyên đề) đã chọn. Tại phương pháp<sup>4</sup> người sử dụng có thể chọn một trong các phương pháp phân nhóm sau: phương pháp phân nhóm *Equal Interval* (khoảng bằng nhau), *Quantile* (tần số bằng nhau), *Standard Deviation* (Độ lệch chuẩn), *Arithmetic Progression*, *Geometric Progression*, *Jenks Natural Break*. Người sử dụng có thể nhập vào giá trị số nhóm<sup>5</sup>. Quá trình phân nhóm dữ liệu được thực hiện có điều chỉnh cho phù hợp dựa trên thư viện Javascript viết bởi Simon Georget (2011). Người sử dụng chọn màu bắt đầu và màu kết thúc tại color<sup>6</sup>. Màu sẽ được nội suy tuyến tính giữa hai màu đã chọn. Quá trình tạo ra các nhóm màu sẽ được thực hiện trên cơ sở sử dụng thư viện Javascript về màu viết bởi Gregor Aisch (2017). Khi người sử dụng đã nhập đầy đủ các giá trị yêu cầu và nhấn nút “thực thi”. Quá trình phân nhóm được tiến hành và người sử dụng sẽ nhìn thấy bản đồ đơn vị hành chính thể hiện kết quả phân nhóm dữ liệu theo chuyên đề và theo chỉ tiêu đã chọn.

- Chức năng phân nhóm k-means dựa trên nhiều chỉ tiêu hoặc nhiều biến dữ liệu do người sử dụng chọn được thiết kế với giao diện như hình 4.





Hình 4: Giao diện công cụ phân nhóm dữ liệu đa biến



Người sử dụng có thể: 1) chọn lớp dữ liệu<sup>1</sup> mà kết quả phân nhóm sẽ được thể hiện trên lớp dữ liệu này; 2) Chọn các chỉ tiêu - trong danh sách các chỉ tiêu<sup>2</sup> - làm cơ sở để tiến hành phân nhóm. Các chỉ tiêu nào được đưa vào danh sách này là do người sử dụng quyết định và đã được upload từ cơ sở dữ liệu đặt tại server sử dụng các công cụ trợ giúp khác có trong hệ thống; 3) Chọn phương pháp<sup>3</sup> phân nhóm đa biến, ví dụ k-means; 4) Chọn số nhóm<sup>4</sup> và 5) Chọn màu bắt đầu và màu kết thúc tại color<sup>5</sup>. Quá trình phân nhóm dữ liệu *k-means* được thực hiện có điều chỉnh dựa trên thư viện Javascript viết bởi Shudima (2015) và Burak Kanber (2012)

### 3. Thực nghiệm phân nhóm dữ liệu trong hệ thống WebGIS

Hình 5 minh họa giao diện của hệ thống WebGIS, với các controls (nút bấm) đã được xây dựng và dữ liệu bản đồ hành chính của khu vực nghiên cứu được hiển thị. Tương tác với bản đồ tạo bởi OpenLayers được thực hiện thông qua các Controls. Có hai nhóm controls: a) Nhóm thứ nhất gồm các controls




được sử dụng để tương tác với bản đồ, như rê bản đồ, phóng to, thu nhỏ, đo lường, chọn đối tượng, tra xét, tìm kiếm thông tin; b) Nhóm thứ hai gồm các controls  được sử dụng để nạp dữ liệu từ cơ sở dữ liệu, hiển thị và phân nhóm dữ liệu. Khi người sử dụng muốn làm việc với dữ liệu, ví dụ y tế, thì chỉ cần nhấn vào control  một giao diện như hình 6 sẽ hiện ra, cho phép người sử dụng thực hiện việc nạp

dữ liệu lấy từ cơ sở dữ liệu của hệ thống, ví dụ nạp dữ liệu *số cơ sở y tế* từ lớp dữ liệu chuyên đề y tế năm 2015. Người sử dụng có thể xem dữ liệu đã nạp lên trang webGIS dưới dạng bảng dữ liệu bằng cách nhấn vào control . Hình 7 minh họa quá trình và kết quả thực hiện phân nhóm dữ liệu *đơn biến*. Khi người sử dụng đưa ra yêu cầu thể hiện và phân nhóm dữ liệu bằng cách bấm vào nút bấm quy định, . Một giao diện giống như hình 3 sẽ xuất hiện. Nếu người sử dụng muốn phân các quận/huyện ra làm bốn nhóm theo số lượng *cơ sở y tế* có trong mỗi quận/huyện vào năm 2015 theo phương pháp *khoảng bằng nhau* (Equal Interval), thì kết quả nhập sẽ giống như giao diện ở hình 7. Khi nhấn nút thực thi, một câu lệnh SQL sẽ được tạo, có dạng “select mahc, SoCsYte from Health2015” và sẽ được gửi đến Web server. Dữ liệu trả về từ Web server nếu có sẽ có dạng, ví dụ ([[760,13],[769,2],[770,9],[773,1],[774,17],[775,1],[778,4],[776,4],[763,2],[771,12],[772,1],[761,2],[764,4],[766,8],[767,3],[765,4],[768,7],[762,2],[777,5],[783,2],[784,1],[785,1],[786,1],[787,1]]). Một cột thuộc tính có tên “SoCsYte” sẽ được tạo thêm vào trong lớp dữ liệu đối tượng “đơn vị hành chính” với giá trị đưa vào cho mỗi đơn vị hành chính sẽ lệ thuộc vào mã hành chính chung giữa các đối tượng thuộc lớp dữ liệu và dữ liệu trả về. Dựa vào các giá trị này, và dựa vào các thông số phân nhóm đã cung cấp, quá trình phân nhóm được thực hiện, màu được gán cho các đơn vị hành chính dựa vào nhóm mà đơn vị hành chính thuộc về.

Hình 8 minh họa quá trình và kết quả thực hiện phân nhóm dữ liệu *đa biến*. Ví dụ nếu người sử dụng muốn phân tích xem các quận huyện nào trong thành phố có sự tương đồng về số lượng các bệnh viện và số lượng các trường trung học cơ sở trong năm 2016, trước tiên người sử dụng đưa ra yêu cầu nạp dữ liệu về số lượng các bệnh viện (select mahc, benhvien from Yte2016) và yêu cầu nạp dữ liệu về số lượng các trường trung học cơ sở (select mahc, stthpt from Giaoduc2016). Hai cột thuộc tính có tên “stthpt” và “benhvien” sẽ được tạo thêm vào trong lớp dữ

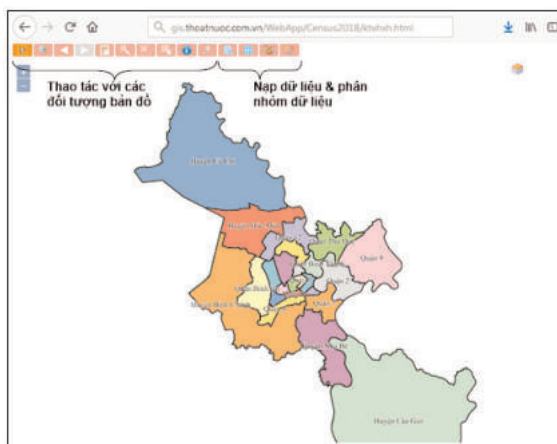


liệu đối tượng “đơn vị hành chính” với giá trị đưa vào cho mỗi đơn vị hành chính sẽ lệ thuộc vào mã hành chính chung giữa các đối tượng thuộc lớp dữ liệu và dữ liệu trả về. Khi người sử dụng đưa ra yêu cầu thể hiện và phân nhóm đa biến dữ liệu bằng cách bấm vào một nút bấm quy định, . Một giao diện giống như hình 4 sẽ xuất hiện với các thuộc tính đã đưa vào “stthpt” và “benhvien”. Người sử dụng chọn các thuộc tính này, chọn số nhóm<sup>4</sup>, chọn thang màu biểu diễn<sup>5</sup> và nhấn nút “thực thi”. Thuật toán *K-means* sẽ được gọi để tiến hành hoạt động phân nhóm. Quá trình thực hiện và kết quả sẽ giống như hình 8. Người sử dụng nhìn vào bản đồ sẽ nhìn thấy được các quận nào có sự tương đồng (thuộc về cùng một nhóm màu) về số lượng các bệnh viện và số lượng các trường trung học cơ sở.

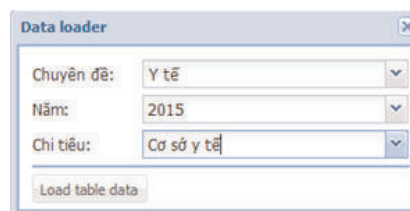
#### 4. Kết luận

Các số liệu thống kê về *Dân số và Lao động, Giáo dục, Y tế, ...* của các quận huyện Tp. Hồ Chí Minh được cung cấp tại trang Web của Cục Thống kê Thành phố Hồ Chí Minh ở dạng các bảng dữ liệu. Mặc dầu các giá trị này giúp người đọc thấy được hiện trạng phát triển tại các quận huyện nhưng nó không cho thấy kiểu mẫu của sự phát triển theo không gian vị trí. Bài báo này trình bày về 1 hệ thống WebGIS có thể được xây dựng, và thông qua hệ thống WebGIS này, người

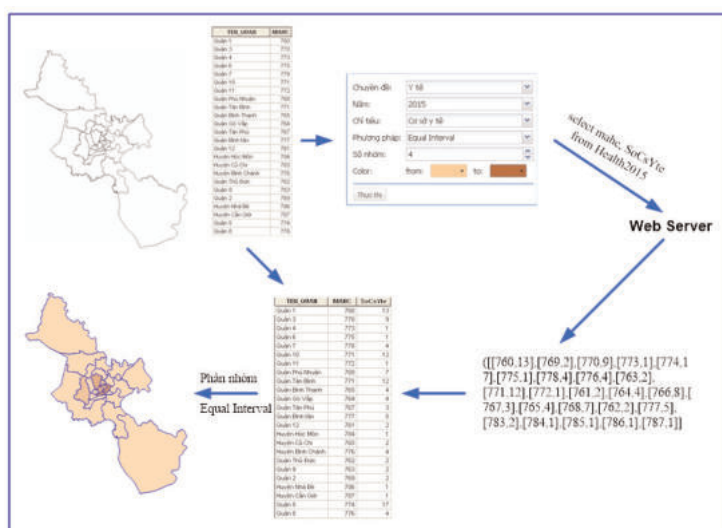
sử dụng có thể tự mình khảo sát đánh giá, phân nhóm sự phát triển của các quận huyện cũng như sự tương đồng theo các chuyên đề và chỉ tiêu khác nhau theo thời gian và theo không gian. Kết quả thực nghiệm chứng tỏ tính hữu ích của hệ thống WebGIS này. ○



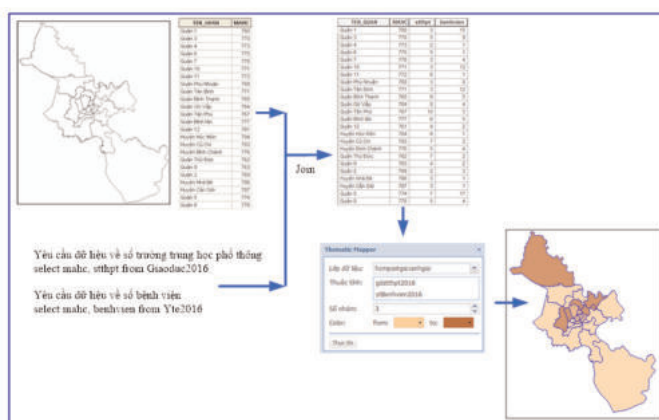
Hình 5: Giao diện của trang webGIS



Hình 6: Giao diện nạp dữ liệu



Hình 7: Thể hiện quá trình và kết quả phân nhóm đơn biến dữ liệu



Hình 8: Thể hiện quá trình và kết quả phân nhóm đa biến dữ liệu

### Tài liệu tham khảo

[1]. Cục Thống Kê Thành phố Hồ Chí Minh, 2015. Niên giám thống kê: Statistical Yearbook of Hochiminh City, Xí nghiệp in thống kê.

[2]. Oliver O'Brien & James Cheshire, 2016. Interactive mapping for large, open demographic data sets using familiar geographical features, Journal of Maps, Taylor & Francis.

[3]. Simon Georget, 2011. Geostats - Javascript classification library. <http://www.wintermezzo-coop.eu/mapping/geostats/>.

[4]. Gregor Aisch, 2017. Chroma - JavaScript

library for all kinds of color manipulations. <http://gka.github.io/chroma.js>

[5]. Ieva Brauksa, 2013. Use of Cluster Analysis in Exploring Economic Indicator Differences among Regions: The Case of Latvia, Journal of Economics, Business and Management, Vol. 1, No. 1, February 2013

[6]. Shudima, 2015. dimas-kmeans. <https://github.com/shudima/dimas-kmeans>

[7]. Burak Kanber, 2012. Machine Learning: k-Means Clustering Algorithm in Javascript. <https://www.burakkanber.com/blog/machine-learning-k-means-clustering-in-javascript-part-1/>

### Summary

#### Display and classify statistical data on WebGIS

Tran Trong Duc - Ho Chi Minh city University of Technology, VNU – HCM

Statistical data on population, education, health, ... are summarized and distributed in tabular form. By allowing users to select, integrate and display these data on map, as well as perform single-variable or multivariate classification without needs of professional software, they can have a deeper understanding of spatial distribution of statistical value as well as can find out groups of data with similar characteristics. To illustrate this idea, a WebGIS is developed based on open source GeoServer, OpenLayers softwares. Data used for the illustration are statistical data from statistical office in Ho Chi Minh city and administrative boundary map. These data are organized and stored in PostgreSQL-PostGIS database. Through this WebGIS system, users with certain knowledge of data analysis can completely perform statistical analysis and display them on an internet-based map for their own purposes. ○