

# GIỚI THIỆU CÁC LOẠI ĐỒ THỊ THƯỜNG ÁP DỤNG TRONG THỐNG KÊ MÔ TẢ (Tiếp theo kỳ trước và hết)

Dominique Haughton - Nguyễn Phong

<http://win-www.uia.ac.be/u/statis/index.html>.

## 4. Đồ thị hình hộp hai chiều: Đồ thị hình túi (bagplots)

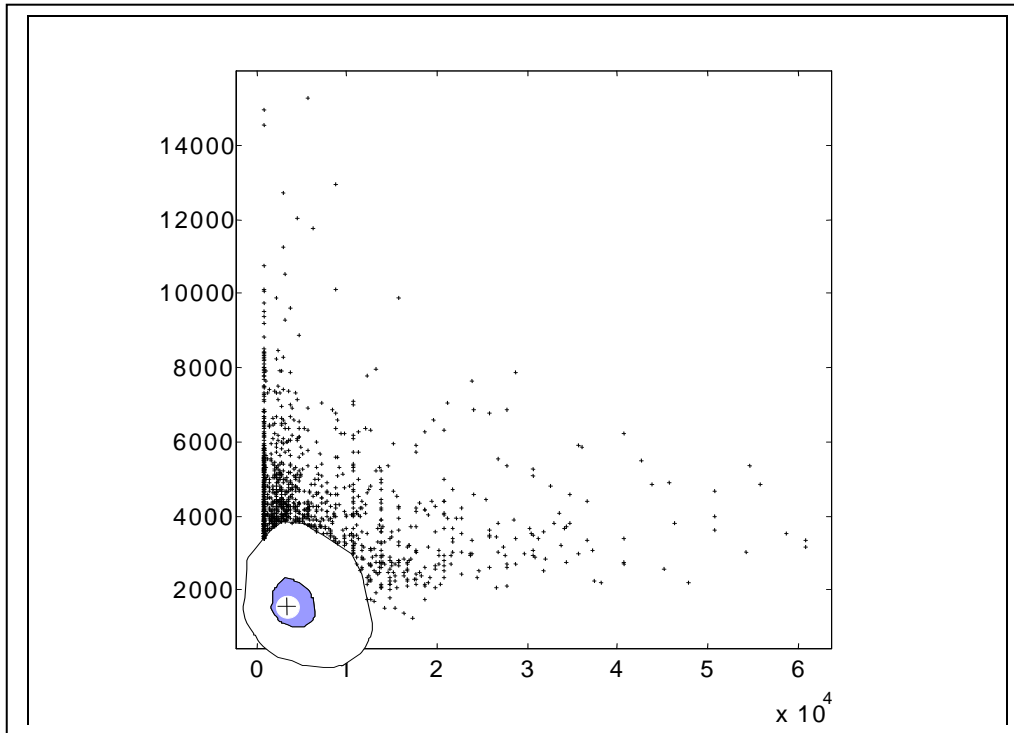
Đồ thị hình túi được Rousseeuw, Rut và Tukey giới thiệu năm 1999 là phiên bản hai chiều của đồ thị hình hộp. Loại đồ thị này giúp chúng ta nhìn được vị trí, sự dàn trải, độ lệch và các giá trị ngoại lai của hai biến số cùng một lúc. Trong Hình 8, chúng tôi trình bày đồ thị túi đối với chi tiêu thực tế bình quân đầu người ( $y$ ) và đất canh tác hàng năm ( $x$ , đơn vị mét vuông) của 4.320 hộ gia đình nông thôn có đất trồng cây hàng năm. Ngoài ra, chúng tôi cũng loại bỏ những hộ gia đình có diện tích loại đất này lớn hơn 60.000 mét vuông hoặc chi tiêu thực tế bình quân đầu người lớn hơn 15.000.000 đồng. Dấu “+” ở phần dưới bên trái của đồ thị là số trung vị chiều sâu (*depth median*). Đây là số trung vị hai chiều, tương tự như số trung vị một chiều, được định nghĩa là độ sâu cao nhất ở giữa (*highest halfspace location depth*) của bộ số liệu (xem Rousseeuw, Ruts and Tukey (1999)). Túi bên trong (*inner bag*) (phần tô đậm) chứa khoảng 50% số liệu và tương ứng với hình hộp trên đồ thị hình hộp. Túi bên ngoài (*outer bag*) (được viền bởi đường cong gián đoạn) được tạo nên bằng cách mở rộng túi bên trong lên 3 lần số trung vị chiều sâu và nó tương ứng với những những đường biên trong đồ thị hình hộp. Những quan sát nằm ngoài túi bên ngoài này là những giá trị ngoại lai. Đồ thị túi có thể tạo được bằng chương trình Fortran và Matlab do Rousseeuw và các tác giả cung cấp (1999); những chương trình này có sẵn trên trang web

Chúng ta có thể thấy rằng biến hai chiều (diện tích đất trồng cây hàng năm và chi tiêu thực tế bình quân đầu người) lệch phải với 50% quan sát có diện tích đất trồng cây hàng năm và chi tiêu thực tế bình quân đầu người thấp. Đồng thời chúng ta có thể thấy các điểm số liệu trên đồ thị phân bố theo dạng hình phễu. Điều này chứng tỏ mối quan hệ giữa chi tiêu thực tế bình quân đầu người và diện tích đất trồng cây hàng năm có dạng phân tán (*heteroskedastic*).

## 5. Đồ thị hình hộp kết hợp với đồ thị mật độ: Đồ thị hình cây đàn violông

Gần đây đã có nhiều nỗ lực trong việc kết hợp đồ thị hình hộp với những dạng đồ thị mật độ. Đồ thị hình cây đàn violông do Hintze và Nelson phát hiện vào năm 1998 là một sự kết hợp giữa một đồ thị hình hộp (chỉ bao gồm hình hộp và hai râu, không có các giá trị ngoại lai) và một đồ thị ước lượng mật độ đơn giản, được gọi là đồ thị thể hiện dấu hiệu của mật độ (*density trace*) như được trình bày ở đầu mục 3. Có thể tạo ra đồ thị hình cây đàn violông bằng phần mềm NCSS (*Number Cruncher Statistical System*). Phần mềm này có thể cài đặt và sử dụng miễn phí từ [www.ncss.com](http://www.ncss.com) trong 30 ngày và có thể dùng cho tối đa 100 quan sát. Hình 9 dưới đây trình bày đồ thị hình cây đàn violông đối với tỉ lệ tốt nghiệp cấp trung học phổ thông ở 61 tỉnh/thành phố của Việt Nam (TCTK2001).

HÌNH 8. ĐỒ THỊ HÌNH TỬ CHI TIÊU THỰC TẾ BÌNH QUÂN ĐẦU NGƯỜI VÀ DIỆN TÍCH ĐẤT TRỒNG CÂY HÀNG NĂM

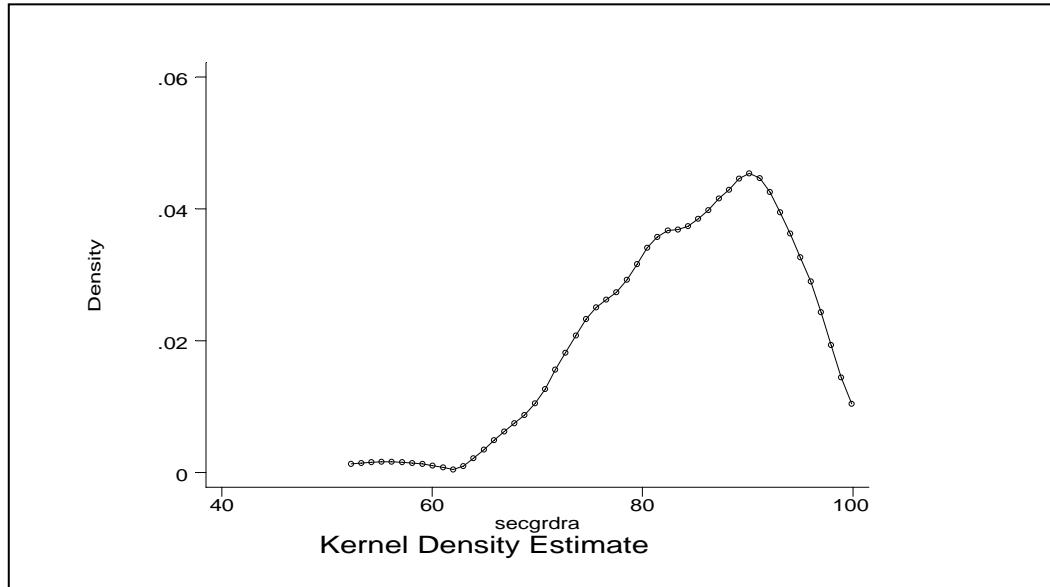


HÌNH 9: ĐỒ THỊ HÌNH CÂY ĐÀN VIÔLÔNG ĐỐI VỚI TỈ LỆ TỐT NGHIỆP TRUNG HỌC PHỔ THÔNG Ở 61 TỈNH/THÀNH PHỐ CỦA VIỆT NAM (TCTK2001)

Tất nhiên, vấn đề nổi lên là sự lựa chọn h khi vẽ đồ thị mật độ. Phần mềm NCSS ngầm định 40% miền số liệu bằng 2 lần h, nhưng theo “quy tắc giai đoạn 0” trình bày ở mục 2 lại khuyến nghị khoảng 20% miền số liệu bằng 2 lần h. Hình 9 sử dụng độ rộng h theo cách thứ 2. Trong đồ thị, có ít nhất 3 nhóm tỉnh/thành phố được hình thành, thể hiện

ở những chỗ phình ra và phần đuôi dài.. Ước lượng mật độ Kernel tạo được theo ngầm định của phần mềm Stata (Hình 10) dường như khẳng định những đặc điểm của đồ thị này. Lưu ý rằng trong đồ thị hình cây đàn violông, hình mật độ được vẽ 2 hai lần bằng 2 đường đối xứng nhau để có thể nhìn rõ hơn độ rộng của những phần phình ra.

HÌNH 10 - ƯỚC LƯỢNG MẬT ĐỘ KERNEL ĐỐI VỚI TỈ LỆ TỐT NGHIỆP TRUNG HỌC PHỔ THÔNG Ở 61 TỈNH/THÀNH PHỐ CỦA VIỆT NAM



### Tài liệu tham khảo

Chiu, S.T. "Lựa chọn độ rộng cho ước lượng mật độ Kernel", Biên niên sử thống kê, 19, 1883-1905 (1991).

Hintze, J. and R.D. Nelson. "Đồ thị viôlông: Sự kết hợp giữa đồ thị hình hộp và đồ thị mật độ", Tạp chí nhà thống kê Mỹ, 52, 181-184 (1998).

Tổng cục Thống kê. Niên giám thống kê. Hà nội: Nhà xuất bản thống kê (2001).

Rousseeuw, P., I. Ruts and J. Tukey. "Đồ thị hình túi: Đồ thị hình hộp biến thiên hai chiều", Tạp chí nhà thống kê Mỹ, 53, 382-387 (1999).

Scott, D.W. "Về đồ thị hình cột tối ưu và đồ thị hình cột dựa trên số liệu thực tế", Biometrika, 66, 605-610 (1979).

Silverman, B. Ước lượng mật độ cho số liệu thống kê và phân tích số liệu. London: Chapman and Hall (1986).

Công ty Stata. "Lệnh kdensity", Sách hướng dẫn Stata cho phiên bản 6, tập 2, 144-151 (1999).

Tukey, J. Phân tích số liệu thăm dò. Reading, MA: Addison-Wesley (1977).

Wand, M.P. "Lựa chọn số cột cho đồ thị hình cột dựa trên các bộ số liệu thực tế", Tạp chí nhà Thống kê, 51, 59-64 (1997).