ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC& KỸ THUẬT MÁY TÍNH

CẤU TRÚC RỜI RẠC CHO KHMT (CO1007)

Ứng dụng thống kê
khảo sát hành trình di chuyển của phương tiện vận tải công cộng

GVHD: Huỳnh Tường Nguyên
Võ Thanh Hùng
Nguyễn Đức Hiệp
Ngô Minh Quốc Hưng
SV thực hiện: Nguyễn Văn A – 22102134
Trần Văn B – 88471475
Lê Thị C – 36811334
Phạm Ngọc D – 97501334
Kiều Thị E – 12341334

Tp. Hồ Chí Minh, Tháng 10/2017

Mục lục

1 Động cơ nghiên cứu

Trên thế giới hiện này, nhằm hỗ trợ tốt nhất cho người dân sử dụng phương tiện công cộng thì chính quyền luôn cố gắng để có được một hệ thống giao thông phù hợp. Tuy nhiên, do một số lý do khách quan lẫn chủ quan, gây ra hiện tượng kẹt xe dẫn tới lãng phí thời gian và chậm trễ công việc của hành khách. Ngoài ra, các trạm xe buýt không được khai thác triệt để khiến cho lãng phí tiền bạc.

Một số thuật toán và giải pháp đã được đề xuất và áp dụng trên thế giới. Tuy nhiên, các giải pháp này chỉ tập trung vào giải quyết một vùng cụ thể được nghiên cứu. Do đó, khi thay đổi ngữ cảnh – tập dữ liệu khác, các thuật toán này không còn đạt được hiệu quả như mong muốn.

Và vì vậy, để áp dụng vào trong nước, cụ thể là thành phố Hồ Chí Minh, chúng ta cần thu thập thông tin cụ thể của thành Phố. Dựa vào những thông tin đó để xây dựng các giải pháp thuật toán phù hợp để có thể giải quyết hiệu quả những vấn đề nêu trên.

Hiện nay trong nước, đặt biệt là thành phố Hồ Chí Minh. Lượng người sử dụng phương tiện công cộng, cụ thể là xe buýt ngày một tăng cao. Chính vì thế, để đáp ứng nhu cầu người dân, số lượng xe buýt đã được gia tăng đáng kể, đồng thời kéo số lượng trạm xe buýt tăng theo.

Nhưng xe buýt là một trong những nguyên nhân gây kẹt xe hàng đầu: đón trả khách, kích thước lớn, tần suất cao,... Những năm gần đây, Chính Phủ đang cố gắng giảm thiểu xe máy, điều này cũng khiến cho lượng xe buýt ngày một nhiều. Vì thế vấn đề kẹt xe hay hiệu suất sử dụng trạm trở nên đáng quan tâm hơn bao giờ hết.

Các giải pháp hiện nay được đưa ra chỉ là tăng cường phát triển dịch vụ đường sắt, phân luồng giao thông, tăng giá xe hay giải quyết vấn đề lấn chiếm lòng lề đường, v.v... Các giải pháp này chỉ mang tính tạm thời, giảm quyết được phần nào vấn nạn kẹt xe cục bộ hoặc là cần có thời gian triển khai lâu dài, không thể áp dụng trong tương lai gần.

Vì thế, cần có một giải pháp để có thể dự đoán điểm kẹt xe để thông báo kịp thời nhằm điều tiết giao thông hiệu quả. Ngoài ra, việc này cũng chỉ ra được các trạm ít sử dụng để đưa ra đề xuất cắt bỏ để giảm chi phí.

2 Mục tiêu

Trong bài tập lớn này, sinh viên được yêu cầu làm một số phép tính toán, thống kê dựa trên dữ liệu thực về xe buýt tại TP. HCM. Qua đó, sinh viên được rèn luyện một số kỹ năng, kiến thức về thống kê. Ngoài ra, sinh viên cũng được luyện tập với lập trình trên ngôn ngữ lập trình hỗ trợ mạnh cho thống kê là **R**.

Bài tập lớn cũng là bước đầu tiên cho phép sinh viên làm quen với kỹ năng giải quyết vấn đề, và hướng tới mục tiêu cao hơn là khai phá dữ liệu, mô hình hóa các bài toán để áp dụng trong thực tiễn.

Qua bài tập lớn này, sinh viên rèn luyện những kỹ năng sau đây:

- Kỹ năng giải quyết vấn đề
- Thống kê căn bản
- Lập trình với ngôn ngữ R phục vụ cho thống kê
- Viết, trình bày báo cáo bằng LaTeX
- Kỹ năng làm việc nhóm

3 Mô tả dữ liệu

Đính kèm đề bài tập lớn là file **data.zip** trong đó chứa thông tin về các tuyến, trạm, hành trình của từng tuyến cũng như thông tin trích xuất hành trình của một số xe buýt.

Sinh viên xem file đặc tả **spec.docx** ở trong file zip để biết định dạng, thuộc tính dữ liệu. Phần rút gọn mô tả như ở bên dưới:

Dữ liệu hành trình một ngày của xe buýt:

Dữ liệu mô tả sự di chuyển trong một ngày của xe buýt TP HCM (tạm gọi là hành trình). Hành trình được biểu diễn dưới dạng tập hợp các điểm GPS theo tọa độ Latitude và Longitude.

Các điểm này được hộp đen xe buýt ghi lại. Dữ liệu này bao gồm ba cột:

- *Lat*: Vĩ độ

- *Long*: Kinh độ
- *Receiving time*: Thời gian nhận được tín hiệu GPS mà hộp đen xe buýt gửi lên, tính theo hệ **Unix Epoch**.

Bảng 1 là ví dụ thông tin gửi GPS của xe tại một thời điểm.

Bảng 1: *Thông tin gửi GPS của xe*

Lat	Long	Receiving time	Mô tả
10.82935238	106.7039795	1472749215	Xe buýt đã đi qua điểm có: vĩ độ:10.82935238, kinh độ: 106.7039795, và hệ thống ghi nhận vào thời gian 1472749215

Dữ liệu tuyến xe buýt:

Dữ liệu mô tả các tuyến xe buýt của TPHCM. Các tuyến được biểu diễn bởi tập hợp các trạm, mỗi trạm được miêu tả bằng một điểm GPS theo tọa độ Latitude và Longitude.

Một dữ liệu tuyến bao gồm các cột sau:

- *Route_Id*: Số hiệu của tuyến
- *Station_Id*: Số hiệu của trạm (trong toàn bộ các trạm TPHCM)
- *Station_Code*: Mã trạm
- *Station_Direction*: Hướng trạm (0 – Trạm nằm trên chiều đi, 1 – Tram nằm trên chiều về)
- *Station_Order*: Thứ tự của trạm trong mỗi chiều của tuyến
- *Station_Name*: Tên trạm
- *Station_Address*: Địa chỉ trạm
- *Lat*: Vĩ độ
- *Lng*: Kinh độ
- *Polyline*: Tập hợp các điểm GPS biểu diễn lộ trình di chuyển từ trạm liền trước tới trạm hiện tại.
- *Distance*: Khoảng cách từ trạm hiện tại tới trạm liền trước (tính theo mét)

Các thư mục **Route*** chứa các thông tin về tất cả các tuyến xe buýt ở TP. HCM.

Các thư mục **Journey*** chứa các thông tin về hành trình của một số xe buýt tại một ngày.

Các thư mục ***GPS** chứa dữ liệu theo đơn vị **GPS**.

Các thư mục ***Cell*** chứa dữ liệu khi đã mapping vào lưới 40 hoặc 60 mét, tức là chia TP. HCM thành các lưới ô vuông có cạnh là 40m hoặc 60m và điểm GPS được chuyển thành dòng, cột của lưới.

4 Nhiệm vụ

Gọi MD là **mã đề** riêng cho mỗi nhóm (phân biệt với mã nhóm N), nhóm sinh viên sẽ thực hiện các yêu cầu dưới đây với các giá trị xác định như sau:

1. Tập mẫu (dataset) mà nhóm có mã đề MD thực hiện được xác định như sau:

$$route\_id_{MD} = \{\, x \in route\_id \mid MD - 10 \leq x \leq MD + 50 \,\}$$
$$jurney\_id_{MD} = \{\, x \in jurney\_id \mid MD \times 3 \leq x \leq MD \times 10 \,\}$$

   Trong đó: **route_id** và **jurney_id** là tất cả các mã của tuyến và hành trình trong tập dữ liệu được cung cấp. Mã của tuyến được xác định là cột dữ liệu **Route_Id** (cột đầu tiên, A) trong mỗi file dữ liệu về tuyến, mã của hành trình chính là **số** trong tên file (không tính đuôi .xlsx)

2. Bài tập 4 mã tuyến cần làm được xác định theo công thức $h_4 = nearestid(MD)$ với hàm *nearestid* trả về mã tuyến gần với số MD nhất trong các tuyến hiện có. Nếu có nhiều hơn một số có cùng giá trị chênh lệch, hàm trả về id nhỏ nhất

3. Bài tập 6 mã hành trình thực tiễn nhóm MD cần thực hiện được xác định như sau: $h_6 = MD \times 5$

Tùy theo yêu cầu từng mã đề, (nhóm) sinh viên thực hiện các câu hỏi dưới đây:

1. Hãy trích xuất từ tập data chung của đề ra các tập dataset theo đúng mã đề của nhóm để sử dụng cho các bài ở sau. **Đây được coi là các tập mẫu để các bạn làm bài sau này (1)**
2. Xác định số lượng xe buýt trong tập mẫu
3. Xác định số lượng tuyến trong tập mẫu
4. Nhóm câu hỏi liên quan đến hành trình của một tuyến xe buýt
   a) số lượng cell mà một tuyến xe buýt đi qua
   b) tổng quãng đường di chuyển của một tuyến xe buýt
   c) danh sách cell mà một tuyến xe buýt chứa nhiều lần
   d) khoảng cách trung bình giữa các trạm liên tiếp trên hành trình của tuyến xe buýt là bao nhiêu
5. Nhóm câu hỏi liên quan đến một tập các tuyến xe buýt
   a) số lượng cell mà một tuyến đi qua
   b) số lượng tuyến xe đi qua một cell cho trước
   c) tuyến nào dài nhất
   d) tuyến nào dài nhì
   e) danh sách các tuyến thuộc một phần ba đầu theo thứ tự chiều dài tuyến giảm dần
   f) tuyến nào chứa nhiều cell nhất
   g) tuyến nào chứa nhiều cell nhì
   h) danh sách các tuyến thuộc một phần ba đầu theo thứ tự số lượng cell đi qua giảm dần
   i) xác định phổ phân bố theo số lần chứa cell của các tuyến xe buýt
   j) danh sách nhóm cell có lượng số lớn nhất mà trong đó các cell được gom nhóm sao cho số lượng tuyến xe buýt đi qua mỗi cell trong nhóm là như nhau.
   k) phân tích khoảng cách giữa 2 trạm liên tiếp trên mỗi tuyến: tính giá trị trung bình, phương sai, độ lệch chuẩn, giá trị trung vị, giá trị lớn nhất, nhỏ nhất. Hãy vẽ histogram cho biến ngẫu nhiên này. Đưa ra các nhận xét.
6. Nhóm câu hỏi liên quan đến hành trình thực tiễn của một xe buýt
   a) số lượng cell mà một xe buýt đi qua
   b) tổng quãng đường di chuyển của xe buýt
   c) vận tốc trung bình của xe buýt
   d) xác định phổ phân bố theo số lần xe buýt đi qua các cell
   e) danh sách nhóm cell có lượng số lớn nhất mà trong đó các cell được gom nhóm sao cho số lần xe buýt đi qua mỗi cell trong nhóm là như nhau.
7. Nhóm câu hỏi liên quan đến hành trình thực tiễn của một tập các xe buýt
   a) tổng số lần di chuyển qua một cell cho trước
   b) quãng đường di chuyển trung bình của các xe buýt
   c) vận tốc trung bình di chuyển của các xe buýt
   d) số lượng cell trung bình đi qua của xe buýt
   e) xe buýt nào di chuyển dài nhất
   f) xe buýt nào di chuyển dài nhì
   g) danh sách các xe buýt thuộc một phần ba đầu theo thứ tự chiều dài di chuyển giảm dần
   h) danh sách các cell có lượng xe buýt qua nhiều nhất
   i) danh sách các cell có lượng xe buýt qua nhiều nhì

j) danh sách các cell có lượng xe buýt qua nhiều nhất hoặc nhiều nhì

k) danh sách các cell thuộc một phần ba đầu theo thứ tự số lượng xe buýt đi qua giảm dần

l) danh sách nhóm cell có lượng số lớn nhất mà trong đó các cell được gom nhóm sao cho số lần xe buýt đi qua mỗi cell trong nhóm là như nhau.

m) khảo sát thời gian trung bình của một xe buýt để đi từ điểm đầu tới điểm cuối của một tuyết nào đó theo đơn vị phút. Hãy tính giá trị trung bình, trung vị, giá trị lớn nhất, nhỏ nhất. Hãy vẽ histogram cho biến ngẫu nhiên này. Đưa ra các nhận xét.

n) hãy tìm tuyến (mã tuyến) mà trong tập dataset này có nhiều hành trình nhất chạy trên tuyến đó

o) với kết quả ở câu 7n, chọn 3 trạm bất kỳ ở trên tuyến đó không phải là 2 điểm đầu cuối để khảo sát, mỗi trạm chỉ xét theo một chiều (đi hoặc về). Khảo sát thời gian để 2 chiếc xe buýt liên tiếp nhau qua trạm (tính theo phút). Tính trung bình, trung vị, lớn nhất, nhỏ nhất. Hãy vẽ histogram cho biến ngẫu nhiên này và đưa ra các nhận xét.

p) với kết quả khảo sát ở 7o, giả sử một người hoàn toàn không có thông tin gì về thời gian chạy của xe buýt, đi ra trạm để đón xe buýt vào một thời điểm **bất kỳ và ngẫu nhiên** (trong khoảng thời gian có xe buýt chạy, tức từ lúc có chuyến sớm nhất cho tới khi cho chuyến muộn nhất theo dataset chạy qua trạm đó). Hãy tính thời gian trung bình mà người đó phải chờ để bắt được xe buýt.

8. **Điểm ùn tắc**: có ba định nghĩa như dưới đây.

   a) có nhiều tuyến xe buýt giao nhau tại một cell;

   b) nhiều hơn ***hai*** xe buýt xuất hiện tại một cell trong cùng một ***khoảng thời gian*** Δt;

   c) có nhiều hơn ***bốn*** xe buýt di chuyển trong các cell liền kề trong cùng một ***khoảng thời gian*** Δt.

   Định nghĩa đầu tiên có ý nghĩa xét về mặt lý thuyết, hai định nghĩa sau cần thông qua dữ liệu di chuyển thực tế của các xe buýt. Theo tập dữ liệu nhận được, hãy xác định các điểm ùn tắc theo từng định nghĩa trên.

9. **Điểm kẹt xe**: là nơi mà xe buýt di chuyển chậm. Đặc điểm là xe buýt di chuyển ***chậm*** qua ***hai*** cell liền kề, cụ thể là hơn 60 giây để di chuyển qua một cell.

   Lưu ý rằng có nhiều trường hợp xe buýt dừng lâu tại một cell, ví dụ như là trạm, hoặc gặp đèn đỏ hoặc gặp sự cố khác. Do vậy, không chắc chắn rằng xe buýt đang tại vị trí kẹt xe nếu chỉ có thông tin tại một cell.

   Theo tập dữ liệu nhận được, hãy xác định các điểm kẹt xe.

10. **Điểm thông thoáng**: là điểm không hề bị ùn tắc hoặc kẹt xe tại bất kỳ thời điểm nào.

11. **Điểm bất thường**: là điểm ***thường xuyên*** bị ùn tắc hoặc kẹt xe.

12. **Điểm cần quan sát**: là điểm mà các xe di chuyển có độ ***chênh lệch lớn*** về vận tốc.

    Lưu ý rằng cần loại bỏ các điểm là đầu mút của các tuyến xe buýt.

5 Hướng dẫn và yêu cầu

5.1 Hướng dẫn

- Cài đặt đồng thời cả R và Rstudio, đây là môi trường hỗ trợ lập trình với R, (nhóm) sinh viên sẽ cần sử dụng để làm bài tập lớn.
- Tìm hiểu kĩ cách soạn thảo văn bản bằng LaTeX và cách sử dụng phần mềm R trong các file hướng dẫn và tìm hiểu thêm trong các tài liệu khác.
- Tạo một folder chung chứa mọi thứ cần thiết để share giữa các thành viên trong nhóm trên các cloud services như Google Drive hay Dropbox,...
- Dùng Doodle để lên kế hoạch họp nhóm.
- Dùng Trello để quản lý project.

  Một số nguồn tham khảo thêm:

  – Microsoft Azure Notebooks có hỗ trợ lập trình R, trực quan (vẽ biểu đồ) để dễ dàng chia sẽ và làm việc nhóm (tập làm quen với notebook)

  – Overleaf soạn thảo LaTeXonline

5.2 Yêu cầu

Mỗi nhóm, từ 3 đến 6 sinh viên, đề xuất giải pháp. Nhóm cần nộp báo cáo trình bày về lời giải cho các câu hỏi và kết quả thực nghiệm. Đồng thời, nhóm cũng cần nộp source code, và trình bày các kết quả của nhóm trong khoảng 5 phút (có slide). **Hôm báo cáo, (nhóm) sinh viên cần in và cầm theo báo cáo.**

Báo cáo và slide trình bày cần được viết dưới dạng LaTeX.

- Thời gian làm bài: Từ 25/11/2017 đến 25/12/2017 (sinh viên theo dõi deadline nộp cụ thể ở mục Assignments).

  Đối với mỗi bài toán, yêu cầu sinh viên trình bày lời giải theo lối truyền thống, sử dụng các công thức, kết quả lý thuyết trong phần kiến thức chuẩn bị. Đồng thời, sau đó trình bày kết quả tính toán và biểu đồ minh họa bằng R.
- Trình bày cả code R và kết quả tính toán trong R giống như file mẫu.
- Viết báo cáo theo đúng **bố cục như trong file mẫu** bằng LaTeX.
- Trong báo cáo cần phải có **log (nhật ký)** ghi rõ: tiến độ công việc, phân công nhiệm vụ, trao đổi của các thành viên,... (*theo template*)

5.3 Nộp bài

- Sinh viên chỉ nộp bài qua hệ thống Sakai: nén tất cả các file cần thiết (file .tex, .pdf, file .R, ...) thành một file tên là "*BTL-CO1007-MT171-Nhom-N.zip*" và nộp trong mục Assignment. Trong đó *N* là mã nhóm. *Chú ý: file nén chỉ được chứa các file quan trọng và không được chứa bất cứ thư mục nào khác*
- *Không chấp nhận bất kỳ một định dạng nào khác (ví dụ .rar, .7z). Bài nộp sai định dạng, thiếu các file quan trọng được xem là một bài nộp không hợp lệ. Bài nộp hợp lệ sẽ không được xem xét để đánh giá, cho điểm.*
- Mỗi nhóm **chỉ cần một thành viên là nhóm trưởng nộp bài.**
- Không nén kèm thư mục dữ liệu **data** và các file/thư mục tạm thời không cần dùng khác. File nộp có **kích thước không quá 10MB**.

6 Cách đánh giá và xử lý gian lận

6.1 Đánh giá

Mỗi bài làm sẽ được đánh giá như sau.

Nội dung	Tỉ lệ điểm (%)
Giải đúng các bài toán bằng công thức và lập luận	30%
Các lệnh (hàm) R được sử dụng đúng đắn và hợp lý	30%
Trình bày kiến thức chuẩn bị rõ ràng, phù hợp	20%
Trình bày báo cáo, slide	20%

6.2 Xử lý gian lận

Bài tập lớn phải được sinh viên (nhóm) TỰ LÀM. Sinh viên (nhóm) sẽ bị coi là gian lận nếu:

- Có sự giống nhau bất thường giữa các bài thu hoạch (nhất là phần kiến thức chuẩn bị). Trong trường hợp này, TẤT CẢ các bài nộp có sự giống nhau đều bị coi là gian lận. Do vậy sinh viên (nhóm) phải bảo vệ bài làm của mình.
- Sinh viên (nhóm) không hiểu bài làm do chính mình viết. Sinh viên (nhóm) có thể tham khảo từ bất kỳ nguồn tài liệu nào, tuy nhiên phải đảm bảo rằng mình hiểu rõ ý nghĩa của tất cả những gì mình viết.

Bài bị phát hiện gian lận thì sinh viên sẽ bị xử ý theo quy định của nhà trường.

Tài liệu

[Dal] Dalgaard, P. *Introductory Statistics with R.* Springer 2008.

[K-Z] Kenett, R. S. and Zacks, S. *Modern Industrial Statistics: with applications in R, MINITAB and JMP,* 2nd ed., John Wiley and Sons, 2014.

[Ker] Kerns, G. J. *Introduction to Probability and Statistics Using R,* 2nd ed., CRC 2015.