Phân Tích Ảnh Hưởng Của Yếu Tố Dân Số   
Và Dự Báo Số Lượng Ca Mắc, Tử Vong Do COVID-19  
Ở Các Tỉnh Thành Của Việt Nam

Huỳnh Ngô Trung Trực, Lê Phạm Lan Anh, Nguyễn Đại Nghĩa

{[19120040, 19120447, 19120735}@student.hcmus.edu.vn](mailto:19120040@student.hcmus.edu.vn)

Khoa Công nghệ Thông tin

Trường Đại học Khoa học Tự nhiên – Đại học Quốc Gia TP. Hồ Chí Minh

***Tóm tắt nội dung* – Sự bùng nổ số ca mắc và tử vong do COVID-19 ở các tỉnh thành Việt Nam cho thấy cần phải xem xét kỹ lưỡng nhiều yếu tố khác nhau có thể tác động đến dịch bệnh cũng như cần có các công cụ dự báo về mức độ lây lan và tử vong do dịch bệnh để các tỉnh thành kịp thời đưa ra các biện pháp can thiệp có hiệu quả, giúp kiểm soát tình hình dịch bệnh. Xem xét trên nhiều yếu tố khác nhau về kinh tế - xã hội, nhóm nghiên cứu tiến thành đánh giá sự liên hệ giữa các yếu tố dân số như mật độ dân số, tỉ lệ dân số thành thị với số lượng ca nhiễm cũng như số ca tử vong do Covid-19 của 63 tỉnh thành ở Việt Nam. Hơn nữa, nhóm sử dụng dữ liệu về số ca mắc và tử vong [1] theo thời gian để xây dựng mô hình dự báo ngắn hạn về tình hình dịch bệnh. Sử dụng công cụ hồi quy tuyến tính trong việc đánh giá sự liên quan giữa các biến và xây dựng mô hình dự báo dịch bệnh trong thời gian 7 ngày tiếp theo, qua quá trình thực nghiệm, nhóm đã có nhiều kết quả đáng chú ý.**

*Từ khóa:* Covid-19, Hồi quy tuyến tính, Linear Regression, Time Series, ARIMA

**I. Giới thiệu**

Đại dịch Covid-19 tính đến hiện tại đã tồn tại một khoảng thời gian dài và để lại những ảnh hưởng to lớn đến mọi mặt từ kinh tế, xã hội, chính trị của thế giới. Việt Nam cũng không ngoại lệ khi tính đến thời điểm hiện tại nước ta đã ghi nhận hơn 2 triệu ca mắc Covid-19. Chính phủ đang tích cực đề ra những chính sách và biện pháp để ngăn ngừa và đẩy lùi đại dịch như ban hành chỉ thị 15, 16, tuyên truyền thông điệp 5K,...

Chúng em nhận thấy có rất nhiều yếu tố ảnh hưởng đến việc lây nhiễm và tỉ lệ tử vong ở các địa phương. Trong các yếu tố đó, các yếu tố như mật độ, tỉ lệ dân thành thị có thể là những yếu tố có ảnh hưởng. Do đó việc tiến hành đánh giá mối liên quan giữa các yếu tố dân số đến tình hình dịch bệnh là cần thiết cho việc đưa ra các quyết định và chính sách phòng chống phù hợp cho từng địa phương. Bên cạnh đó, nhóm chúng em đề xuất sử dụng mô hình ARIMA để dự báo số ca mắc mới và tử vong và sử dụng kết quả trong việc so sánh, đánh giá và các vấn đề liên quan khác.

Để đánh giá sự liên quan giữa các biến trong các mô hình sử dụng trong nghiên cứu, cần các thông tin về nhiều yếu tố khác nhau do sự tác động qua lại của các yếu tố này. Ngoài ra, để việc đánh giá sự liên quan cũng như mô hình dự báo được tốt, số lượng mỗi loại thông tin cần dùng cần phải đủ lớn. Trong khi đó thông tin về dịch bệnh ở các tỉnh thành ở Việt Nam chủ yếu gồm 2 nhóm: về số ca (mắc, tử vong, hồi phục…) và về tình hình tiêm vaccine (1 mũi, 2 mũi, tổng số mũi), thiếu một số nội dung quan trọng như số ca theo độ tuổi, giới tính, hay loại vaccine tiêm, tổng số lượng theo từng loại… Sự thiếu đầy đủ các loại thông tin cần thiết là một khó khăn trong quá trình tiến hành thực nghiệm của nhóm.

Bài nghiên cứu này nhằm mục đích tạo cơ sở khoa học cho những dự đoán mang yếu tố chủ quan cụ thể là sự tương quan giữa yếu tố dân số và số lượng ca mắc, tử vong, hay sự ảnh hưởng của số ca mắc, tử vong trước đó với những ca mắc, tử vong trong tương lai. Việc xác định một cách khách quan này tạo cơ sở để thực hiện những chính sách phù hợp trong tương lai như các chính sách về dân số để kiểm saots dịch bệnh. Tầm nhìn được đề ra trong bài này là có thể phát triển bài nghiên cứu với nhiều hướng tiếp cận hơn và theo nhiều phương diện hơn như các yếu tố khác không chỉ về dân số mà còn có thể về địa lý, môi trường,... Đó là những đóng góp lớn cho công việc hàng đầu trong bối cảnh hiện nay đó là chung tay phòng chống dịch bệnh.

**II. Phương pháp**

***A. Hồi quy tuyến tính***

Hồi quy tuyến tính (Linear Regression) là phương pháp thống kê giúp hồi quy và dự báo dữ liệu theo thuật toán giữa một một giá trị liên tục với một hoặc nhiều các giá trị liên tục, định danh hay phân loại có liên quan. Hồi quy tuyến tính là một mô hình đơn giản nhưng mang lại hiệu quả cao.

Hồi quy dùng để xem xét mối quan hệ tuyến tính giữa hai biến và . Trong đó là biến độc lập (ảnh hưởng đến ), còn là biến phụ thuộc (chịu ảnh hưởng bởi biến ). Trong hồi quy tuyến tính, output của bài toán chính là sự kết hợp các input theo một tỷ lệ nào đó. Tỷ lệ này là các hệ số , và thường được gọi là trọng số của mô hình. Các giá trị có thể được viết thành vector , các trọng số có thể được viết thành vector . Việc tối ưu mô hình Linear Regression là tìm ra vector sao cho từ input ta có thể tính ra được output của bài toán.

Ta có đầu vào là là các thể hiện các thuộc tính dưới dạng số và đầu ra sẽ là , với là giá trị thật. có thể được biểu diễn như sau:

Giá trị dự đoán và giá trị thật thường là hai giá trị khác nhau do sai số của mô hình.[8]

Trong nghiên cứu này, nhóm tiến hành hồi quy đơn biến và hồi quy đa biến để đánh giá sự phụ thuộc của số lượng ca mắc và số lượng ca tử vong vào các biến giải thích bao gồm:

* popdens: mật độ dân số (người/km2)
* pop: dân số (nghìn người)
* pop\_city: dân số thành thị (nghìn người)

Giả định rằng các biến phụ thuộc có quan hệ tuyến tính với biến giải thích, hoặc có quan hệ phi tuyến có thể tuyến tính hóa qua phép đổi biến bằng logarit tự nhiên. Một điều lưu ý là dữ liệu về số ca mắc có mức độ chênh lệch rất lớn giữa các tỉnh thành (từ dưới 1000 cho đến hơn 500000) [1] nên việc đổi biến qua phép lấy logarit tự nhiên nhìn chung là phù hợp. Các biến giải thích cũng được làm tương tự.

Công thức chung cho mô hình hồi quy sử dụng trong thử nghiệm:

Đơn biến:

Đa biến:

Sử dụng bộ dữ liệu đầu vào, nhóm tiền hành hồi quy với phần mềm R để tính toán giá trị ước lượng các tham số , đó là các . Các giá trị được coi là có ý nghĩa thống kê nếu miền bác bỏ có giá trị đủ nhỏ hơn một ngưỡng cho trước, thường chọn là 0.01.

Bên cạnh xem xét các hệ số ước lượng, các thông tin được xem xét để đánh giá độ tốt của mô hình hồi quy là hệ số tương quan giữa biến phụ thuộc và biến giải thích Corr, và gía trị R-squared (), cho biết khả năng giải thích sự biến động của biến phụ thuộc bởi biến giải thích.

***B. ARIMA***

Mô hình ARIMA (Autoregressive Integrated Moving Average) là một trong số những mô hình phổ biến của Time Series (Chuỗi thời gian). Mô hình dựa trên giả thuyết chuỗi dừng và phương sai sai số không đổi. Đầu vào của ARIMA là những tín hiệu quá khứ của chuỗi được dự báo để dự báo nó.

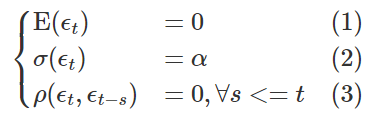
Chúng ta biết rằng hầu hết các chuỗi thời gian đều có sự tương quan giữa giá trị trong quá khứ đến giá trị hiện tại. Mức độ tương quan càng lớn khi chuỗi càng gần thời điểm hiện tại. Chính vì thể mô hình ARIMA sẽ tìm cách đưa vào các biến trễ nhằm tạo ra một mô hình dự báo fitting tốt hơn giá trị của chuỗi.[6]

Mô hình dựa trên phương trình hồi quy tuyến tính đa biến (Multiple Linear Regression) của các biến đầu vào gồm các thành phần chính sau:

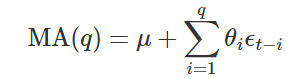
● **Tự hồi quy (Auto Regression - AR):**  bao gồm tập hợp các độ trễ của biến hiện tại. Độ trễ bậc p chính là giá trị lùi về quá khứ p bước thời gian của chuỗi. Độ trễ dài hoặc ngắn trong quá trình AR phụ thuộc vào tham số trễ p. Cụ thể, quá trình của chuỗi được biểu diễn như bên dưới[6]:



● **Trung bình trượt (Moving Average - MA):** là quá trình dịch chuyển hoặc thay đổi giá trị trung bình của chuỗi theo thời gian. Quá trình trung bình trượt sẽ tìm mối liên hệ về mặt tuyến tính giữa các phần tử ngẫu nhiên (Stochastic Term). Chuỗi này phải là một chuỗi nhiễu trắng thỏa mãn các tính chất[6]:



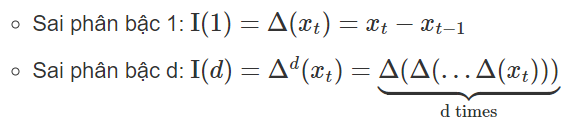
Vế (1) có nghĩa rằng kỳ vọng của chuỗi bằng 0 để đảm bảo chuỗi dừng không có sự thay đổi về trung bình theo thời gian. Vế (2) là phương sai của chuỗi không đổi. Do kỳ vọng và phương sai không đổi nên chúng ta gọi phân phối của nhiễu trắng là phân phối xác định (Identical Distribution) và được kí hiệu là . Nhiễu trắng là một thành phần ngẫu nhiên thể hiện cho yếu tố không thể dự báo của mô hình và không có tính quy luật. Quá trình trung bình trượt được biểu diễn theo nhiễu trắng như sau[6]:



Quá trình này có thể được biểu diễn theo [dịch chuyển trễ](https://en.wikipedia.org/wiki/Lag_operator) B như sau:



● **Đồng tích hợp / Lấy sai phân (Integrated - I):** Yêu cầu chung của các thuật toán trong chuỗi thời gian là chuỗi phải đảm bảo tính dừng. Hầu hết các chuỗi đều tăng hoặc giảm theo thời gian. Do đó yếu tố tương quan giữa chúng chưa chắc thực sự mà là do chúng cùng tương quan theo thời gian. Khi biến đổi sang chuỗi dừng, các nhân tố ảnh hưởng thời gian được loại bỏ và chuỗi sẽ dễ dự báo hơn. Để tạo thành chuỗi dừng, một phương pháp đơn giản nhất là chúng ta sẽ lấy sai phân. Một số còn quy đổi sang logarit hoặc lợi suất. Bậc của sai phân để tạo thành chuỗi dừng còn gọi là bậc của quá trình đồng tích hợp (Order of Integration). Quá trình sai phân bậc d của chuỗi được thực hiện như sau[6]:



Thông thường chuỗi sẽ dừng sau quá trình đồng tích hợp I(0) hoặc I(1). Rất ít chuỗi chúng ta phải lấy tới sai phân bậc 2. Một số trường hợp chúng ta sẽ cần biến đổi logarit hoặc căn bậc 2 để tạo thành chuỗi dừng. Phương trình hồi quy ARIMA(p, d, q) có thể được biểu diễn dưới dạng:



Trong đó là giá trị sai phân bậc d và là các chuỗi nhiễu trắng.

**III. Thực nghiệm**

***A. Tập dữ liệu***

Tập dữ liệu về Covid-19 được lấy từ trang <https://covid19.ncsc.gov.vn/dulieu/>[1] với các trường thông tin khác nhau theo thời gian như: số ca mắc mới, tổng số ca mắc, số ca tử vong trong ngày, tổng số ca tử vong,... của các tỉnh thành trên cả nước. Ngoài ra, nhóm còn lấy thêm dữ liệu về dân số từ trang <https://www.gso.gov.vn/dan-so/>[2]

● Ở thử nghiệm 1, nhóm em trích ra tổng số ca mắc, tổng số ca tử vong, dân số, mật độ dân số và số dân thành thị của ở tổng cộng 63 tỉnh thành trên cả nước.

● Ở thử nghiệm 2, nhóm em trích ra số ca mắc theo ngày và số ca tử vong theo ngày từ ngày 1/8/2021 đến ngày 7/1/2022 của thành phố Hồ Chí Minh, Bà Rịa - Vũng Tàu và Đắk Lắk.

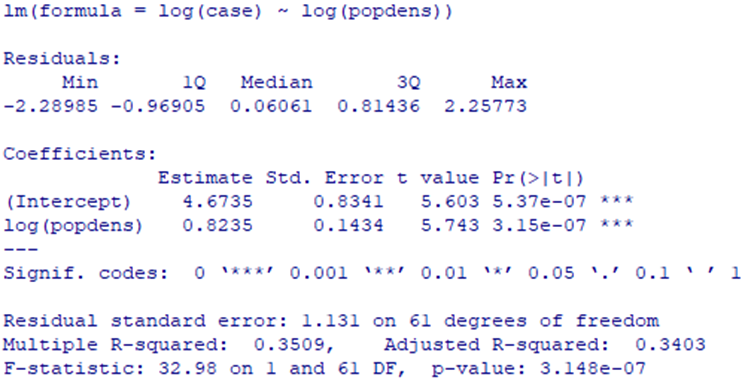
***B. Thử nghiệm 1: Sự tương quan của các yếu tố dân số đến số ca mắc và tử vong do COVID-19***

**1. Số ca mắc**

**a. Hồi quy đơn biến**

*Mật độ dân số*

Kết quả hồi quy

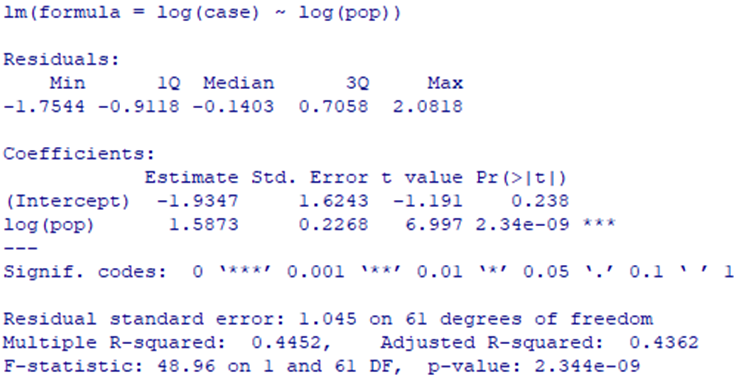


Kết quả cho thấy mức ý nghĩa trong ước lượng các hệ số đều nhỏ hơn 0.001, tức các ước lượng này có ý nghĩa thống kê.. Phương trình hồi quy số ca mắc theo mật độ dân số là:

Hệ số góc của ước lượng trên cho thấy mật độ dân số cao hơn mỗi 1% tương ứng với khoảng 0.8235% về số ca mắc. Mô hình này giải thích được khoảng 35% sự biến động của số ca mắc theo mật độ dân số

*Dân số*

Kết quả hồi quy:

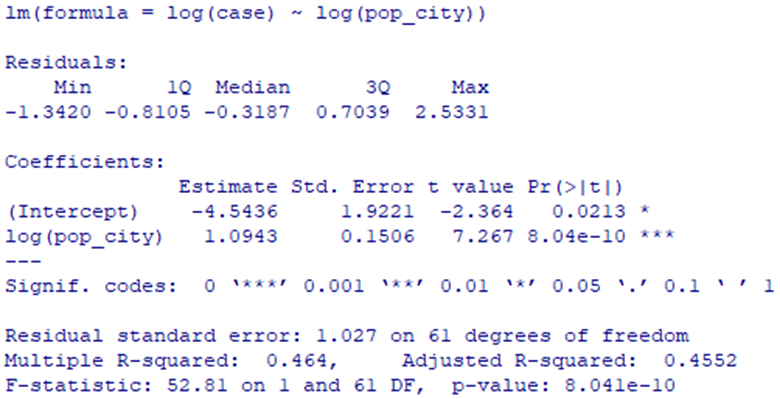


Trong các hệ số của đường thẳng hồi quy thì chỉ có hệ số góc có mức ý nghĩa nhỏ hơn 0.001, còn hệ số có giá trị này khá lớn, khoảng 0.238, nên chỉ có có ý nghĩa thống kê. Phương trình hồi quy số ca mắc theo dân số là:

Hệ số góc của ước lượng trên cho thấy dân số cao hơn mỗi 1% tương ứng với khoảng 1.5873% về số ca mắc. Mô hình này giải thích được khoảng 44.52% sự biến động của số ca mắc theo dân số. Tuy nhiên như trình bày ở trên, hệ số chặn -1.9347 chưa có ý nghĩa thống kê.

*Dân số thành thị*

Kết quả hồi quy:



Kết quả cho thấy mức ý nghĩa trong ước lượng các hệ số là nhỏ hơn 5% đối với và nhỏ hơn đối với , tức các ước lượng này có ý nghĩa thống kê. Phương trình hồi quy số ca mắc theo dân số thành thị là:

Hệ số góc của ước lượng trên cho thấy mật độ dân số thành thị cao hơn mỗi 1% tương ứng với khoảng 1.0943% về số ca mắc. Mô hình này giải thích được khoảng 46.4% sự biến động của số ca mắc theo dân số thành thị.

*Tóm tắt*

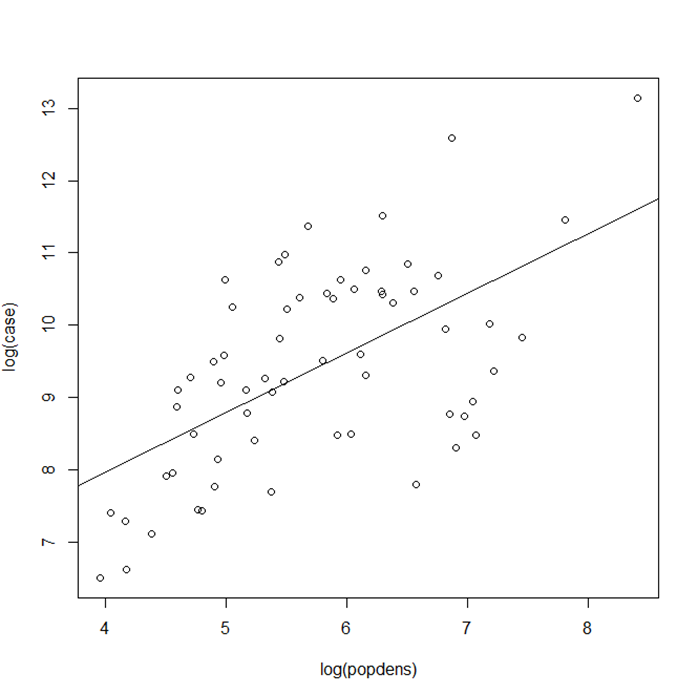
Tóm tắt các hệ số ước lượng, chỉ số R-squared và hệ số tương quan trong các mô hình hồi quy đơn biến đã thực hiện

|  | X = popdens | X = pop | X = pop\_city |
| --- | --- | --- | --- |
|  | 4.6735 | -1.9347\* | -4.5436 |
|  | 0.8235 | 1.5873 | 1.0943 |
|  | 0.3509 | 0.4452 | 0.464 |
| Corr | 0.5924 | 0.6673 | 0.6812 |

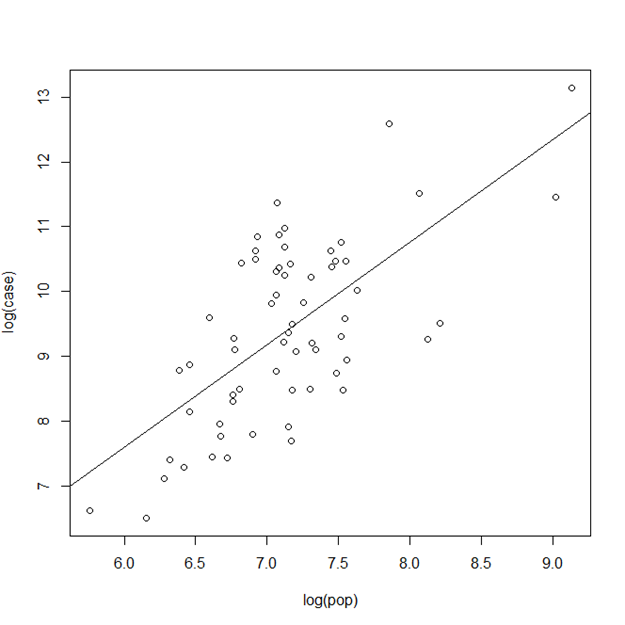
Bảng 1. Các hệ số trong hồi quy đơn biến log(case).

\* Màu đỏ - Hệ số ước lượng không có ý nghĩa thống kê

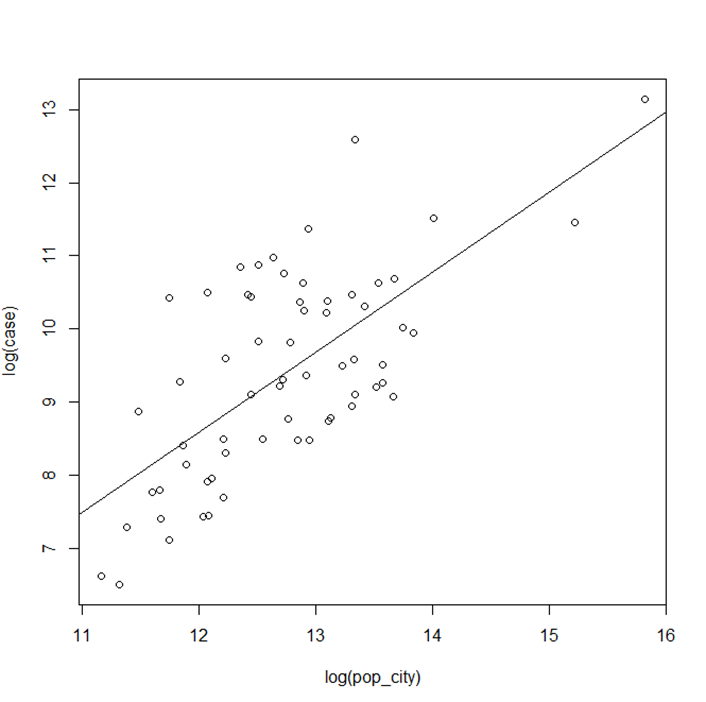
Đồ thị phân tán của các mô hình hồi quy đơn biến



Hình 1. Đồ thị phân tán log(case) - log(popdens).



Hình 2. Đồ thị phân tán log(case) - log(pop).



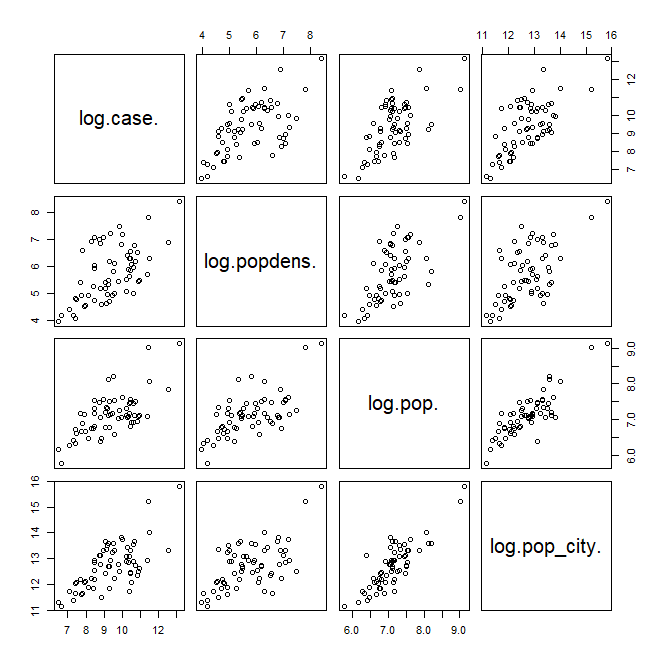
Hình 3. Đồ thị phân tán log(case) - log(pop\_city).

Có thể thấy hệ số tương quan giữa biến phụ thuộc là số ca mắc và biến giải thích lần lượt là mật độ dân số, dân số, dân số thành thị đều dương và khá cao, chứng tỏ tương quan thuận và có liên quan tốt giữa từng biến giải thích với biến phụ thuộc.

Từ kết quả các mô hình đã nêu, cho thấy các biến độc lập đều có tác động đến số lượng ca mắc, theo quan hệ log-log. Tiếp theo, nhóm đồ án sẽ tiến hành hồi quy đa biến để đánh giá sự tác động chung của nhiều biến trong việc giải thích số lượng ca mắc ở từng địa phương

**b. Hồi quy đa biến**

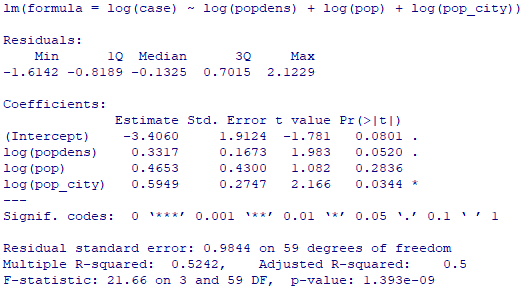
Xem xét tất cả các biến vàtrong việc giải thích sự thay đổi về . Biểu đồ phân tán theo từng cặp biến trong số 4 biến trên như sau.



Hình 4. Đồ thị phân tán theo cặp trong hồi quy đa biến cho log(case).

Như ở phần hồi quy đơn biến, có hệ số tương quan dương khá cao với cả 3 biến còn lại, tuy nhiên không thực sự tương quan quá lớn (gần 1). Để xem xét chọn biến cho mô hình, ta có thể thử hồi quy với cả 3 biến giải thích hoặc sử dụng công cụ *step* của R. Ở đây nhóm đồ án thử nghiệm hồi quy theo cả 3 biến, sau đó sử dụng công cụ step để xem xét tất cả 3 biến đang quan tâm và chọn ra mô hình tối ưu [3].

Hồi quy theo cả 3 biến

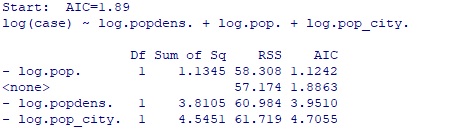


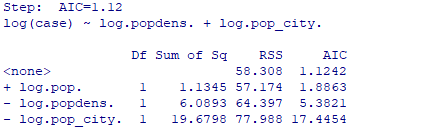
Kết quả quan sát được cho thấy chỉ có hệ số tự do, hệ số của log(popdens) và của log(pop\_city) là có ý nghĩa thống kê. Hệ số của log(pop) có trị số p khá lớn nên không có ý nghĩa thống kê.

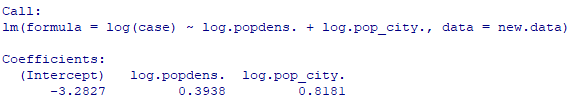
Tiếp tục tiến hành sử dụng step của R để lựa chọn mô hình



Quá trình tìm mô hình tối ưu dừng ở mô hình với 2 biến giải thích là vì mô hình này có giá trị AIC thấp nhất [3].

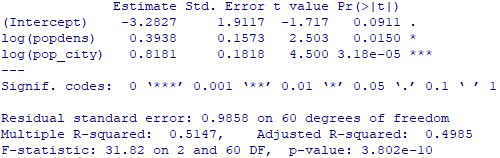






Phương trình tuyến tính tiên đoán cho log(case) là:

Thực hiện đánh giá lại ý nghĩa thống kê của các hệ số ước lượng



Các hệ số ước lượng đều có ý nghĩa thống kê, và mô hình này giải thích được khoảng 51.47% sự thay đổi của . Điều đáng chú ý trong việc lựa chọn mô hình hồi quy đa biến là vai trò của biến log(pop) không có ý nghĩa thống kê. Mặc dù pop và popdens trên thực tế phụ thuộc nhau thông qua diện tích, tuy nhiên việc không đưa diện tích của tỉnh thành đang khảo sát vào mô hình giúp ta có thể tạm coi là hai biến có ý nghĩa khác nhau.

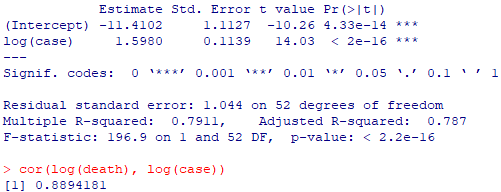
Hệ số của log(popdens) và của log(pop\_city) lần lượt là 0.3938 và 0.8181 có ý nghĩa thể hiện sự tương quan thuận giữa mật độ dân số và số dân thành thị đối với số ca mắc của một tỉnh thành, trong đó tác động của dân số thành thị là mạnh hơn.

**2. Số ca tử vong.**

**a. Hồi quy đơn biến.**

Trước hết, ngoài các yếu tố dân số, cần đánh giá sự liên quan của số ca tử vong và số ca mắc của.

Sử dụng R thực hiện hồi quy trên số liệu của 55 tỉnh thành có ít nhất 1 ca tử vong (dữ liệu tính đến 19/01/2022).



Các hệ số ước lượng đều có ý nghĩa thống kê, phương trình hồi quy:

.

Mỗi 1% sự gia tăng của số ca mắc COVID-19 tương ứng với khoảng 1.598% sự gia tăng của số ca tử vong. Sử dụng mô hình này, số ca mắc giải thích được 79.11% sự thay đổi của số ca tử vong. Hệ số tương quan của log(death) và log(case) cũng rất cao, xấp xỉ 0.89.

Tiếp theo, tương tự khi xem xét số lượng ca mắc, ta có kết quả của các mô hình hồi quy đơn biến của log(death) lần lượt theo các biến giải thích là log(popdens), log(pop), log(pop\_city).

Kết quả của các mô hình hồi quy đơn biến của log(death) với lần lượt các biến giải thích được tóm tắt trong bảng.

|  | X = popdens | X = pop | X = pop\_city |
| --- | --- | --- | --- |
|  | -0.7295\* | -8.8653 | -12.5469 |
|  | 0.8210 | 1.7950 | 1.2912 |
|  | 0.1135 | 0.2118 | 0.2315 |
| Corr | 0.3369 | 0.4602 | 0.4811 |

Bảng 2. Các hệ số trong hồi quy đơn biến log(death)

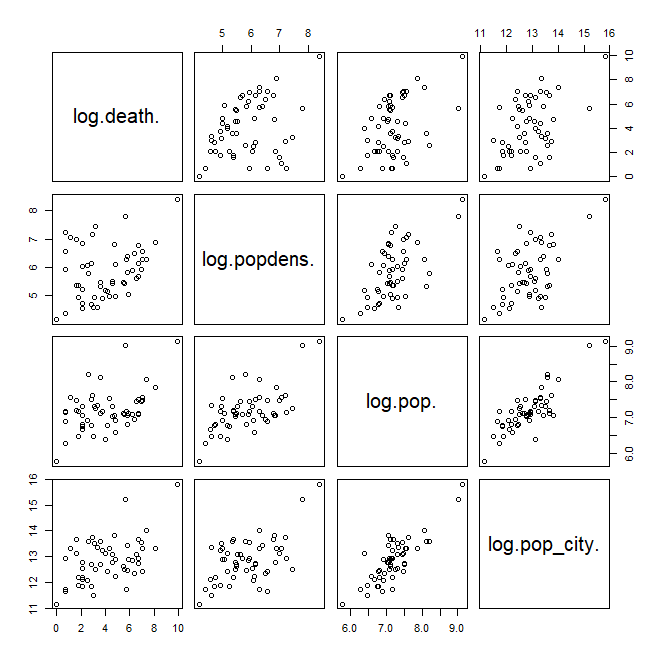
\* Màu đỏ - Hệ số ước lượng không có ý nghĩa thống kê

Có thể thấy có tương quan thuận giữa từng biến giải thích đối với biến log(death) nhưng các hệ số tương quan không thực sự cao, ở mức nhỏ hơn 0.5.

Ngoài ra chỉ số R-squared của các mô hình này ở mức thấp, chỉ giải thích được 11%-23% sự thay đổi của số lượng ca tử vong. Kết quả này cho thấy số lượng ca tử vong còn bị phụ thuộc vào nhiều yếu tố khác chưa được đưa vào mô hình.

**b. Hồi quy đa biến.**

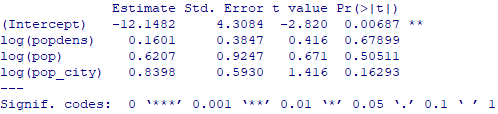
Xem xét tất cả các biến vàtrong việc giải thích sự thay đổi về . Biểu đồ phân tán theo từng cặp biến trong số 4 biến trên như sau.



Hình 5. Đồ thị phân tán theo cặp trong hồi quy đa biến cho log(death).

Như ở mục trước, biến log(death) có tương quan dương nhưng không chặt chẽ với 3 biến giải thích.

Kết quả hồi quy log(death) theo 3 biến giải thích cho thấy các hệ số ước lượng không có ý nghĩa thống kê (p-value lớn)



Từ các mô hình hồi quy đơn biến và đa biến, cho thấy số ca tử vong do COVID-19 ở một tỉnh thành có tương quan thuận nhưng chưa liên quan chặt chẽ đến các yếu tố dân số như: mật độ dân số, dân số, dân số thành thị.

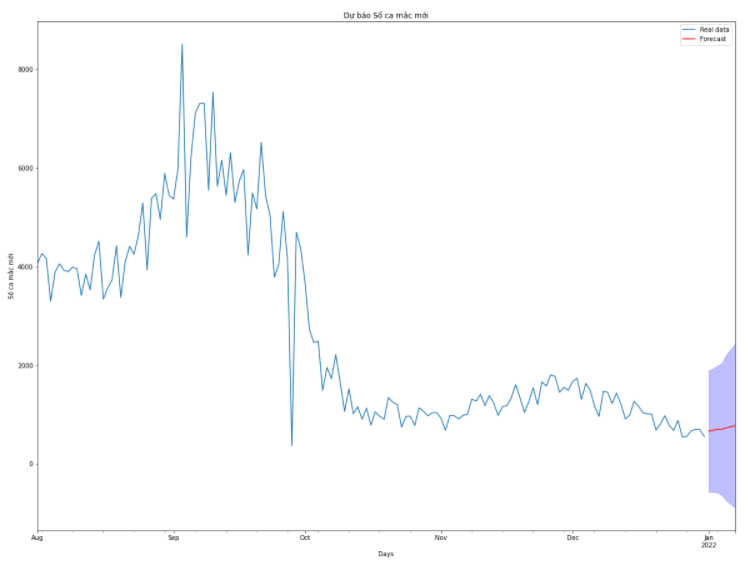
**3. Thảo luận.**

Việc phân tích hồi quy số lượng ca bệnh và số lượng ca tử vong của các tỉnh theo một số yếu tố dân số của địa phương cho một số kết quả trái ngược nhau. Số lượng ca mắc có mối liên hệ khá rõ với từng yếu tố: mật độ dân số, dân số, dân số thành thị, đặc biệt là liên quan đến cặp yếu tố (mật độ dân số, dân số thành thị) khi hai biến này có thể giải thích được khoảng 51.47% sự thay đổi về số ca mắc. Ở chiều ngược lại, số lượng ca tử vong, mặc dù có sự tương quan mạnh với số lượng ca mắc, nhưng lại không liên quan đủ mạnh đến các biến về dân số nêu trên. Điều này chứng tỏ còn nhiều yếu tố khác ngoài các yếu tố dân số đã nêu, tác động đến số ca tử vong của mỗi tỉnh thành.

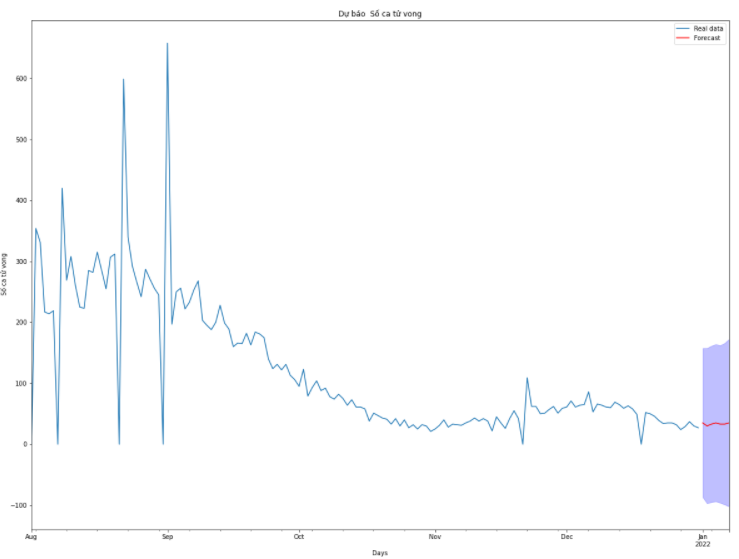
***C. Thử nghiệm 2: Dự đoán số ca mắc, số ca chết do Covid-19 ở các tỉnh thành***

Trong thử nghiệm này, nhóm em chọn mô hình ARIMA của nhóm Time Series để dự đoán số ca mắc và số ca tử vong theo ngày của các tỉnh thành ở Việt Nam. Nhóm em tiến hành thử nghiệm với tập dữ liệu training là số người mắc và tử vong do Covid-19 từ ngày 1/8/2021 đến ngày 31/12/2021 và tập dữ liệu test là 7 ngày tiếp theo. Nhóm em quyết định khảo sát thành phố Hồ Chí Minh, Bà Rịa Vũng Tàu và Đắk Lắk tương ứng với các mức mật độ dân số cao, trung bình và thấp. Ngoài ra, đây là các tỉnh có số liệu thu thập thực tế phù hợp với việc dự báo mô hình.

***1. Thành phố Hồ Chí Minh***

******

Hình 6. Biểu đồ thể hiện những ca mắc cũ và dự đoán tương ứng của Thành phố Hồ Chí Minh.

******

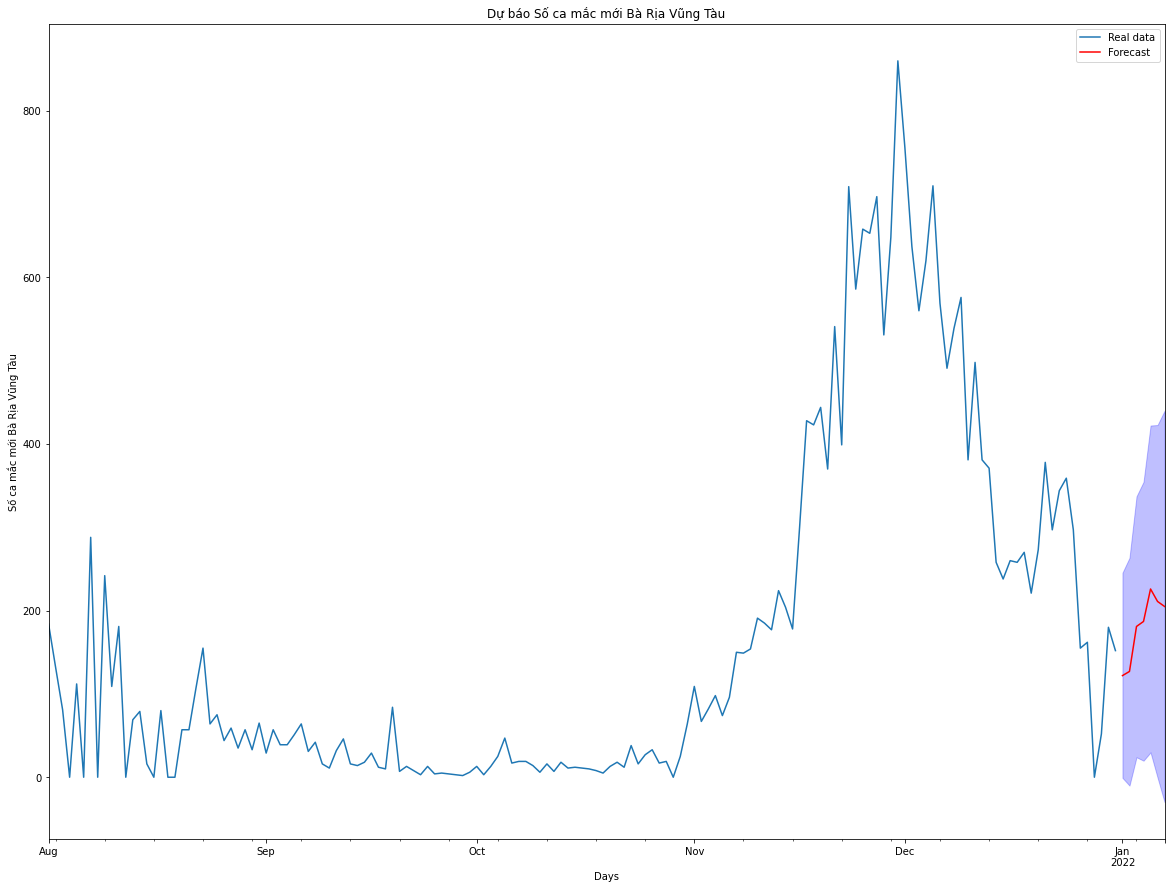
Hình 7. Biểu đồ thể hiện những ca tử vong cũ và dự đoán tương ứng của Thành phố Hồ Chí Minh.

| NGÀY | CA MẮC MỚI | | CA CHẾT MỚI | |
| --- | --- | --- | --- | --- |
| DỰ ĐOÁN | THỰC TẾ | DỰ ĐOÁN | THỰC TẾ |
| 1/1/2022 | 666 | 569 | 35 | 27 |
| 2/1/2022 | 683 | 384 | 30 | 52 |
| 3/1/2022 | 707 | 662 | 33 | 25 |
| 4/1/2022 | 705 | 664 | 35 | 21 |
| 5/1/2022 | 735 | 448 | 33 | 20 |
| 6/1/2022 | 755 | 442 | 35 | 18 |
| 7/1/2022 | 777 | 489 | 33 | 18 |

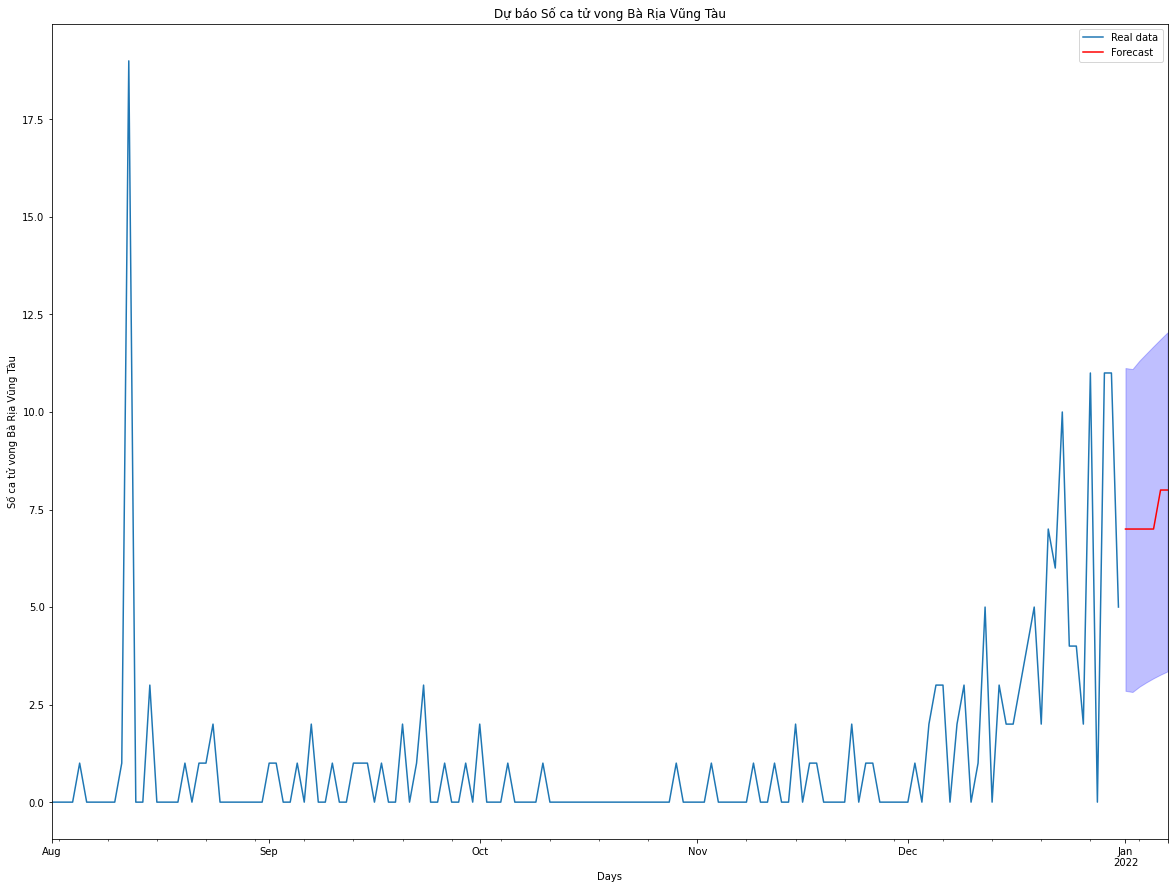
Bảng 3. So sánh kết quả dự đoán và số liệu thực tế của thành phố Hồ Chí Minh.

Ở thành phố Hồ Chí Minh, ta nhận thấy số người mắc và tử vong từ trước tháng 10/2021 biến động khá lớn. Số ca mắc trong ngày có thể lên tới trên 8000, còn số ca tử vong có thể lên tới trên 600 người/ngày. Từ thời gian đó trở đi, tình hình đã ngày càng ổn định hơn. Theo kết quả mô hình, ta thấy rằng số ca mắc mỗi ngày được dự đoán cao hơn thực tế. Số ca chết cũng tương tự. Tuy nhiên, chỉ có ngày 2/1/2022, số ca chết do Covid-19 tăng nhanh và gấp khoảng hai lần những ngày còn lại (52 người). Mô hình dự báo không quá sát sao, tuy nhiên ổn định và không chênh lệch quá nhiều so với thực tế.

***2. Bà Rịa Vũng Tàu***



Hình 10. Biểu đồ thể hiện những ca mắc cũ và dự đoán tương ứng của Bà Rịa Vũng Tàu.



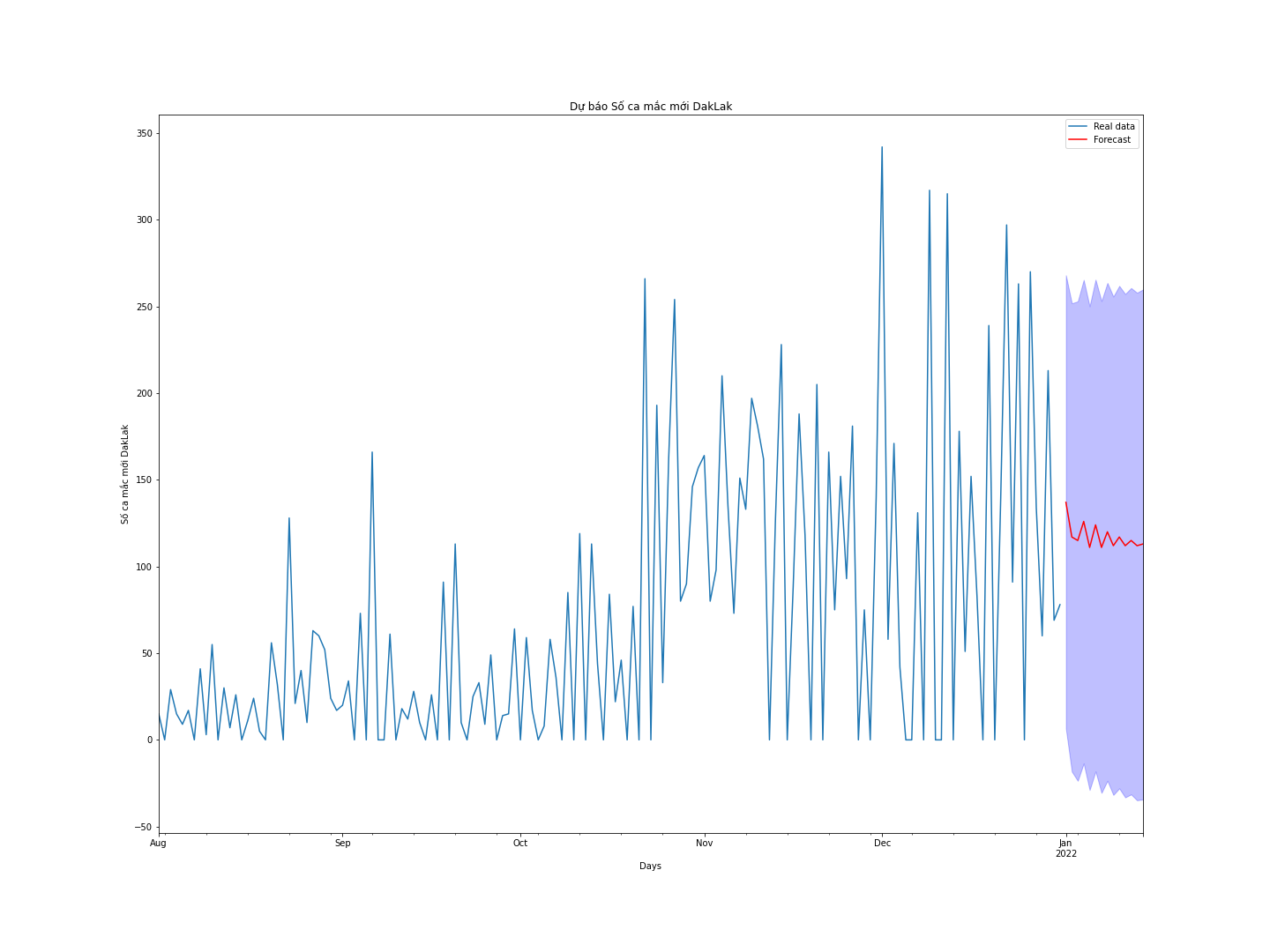
Hình 11. Biểu đồ thể hiện những ca tử vong cũ và dự đoán tương ứng của Bà Rịa Vũng Tàu.

| NGÀY | CA MẮC MỚI | | CA CHẾT MỚI | |
| --- | --- | --- | --- | --- |
| DỰ ĐOÁN | THỰC TẾ | DỰ ĐOÁN | THỰC TẾ |
| 1/1/2022 | 122 | 104 | 7 | 16 |
| 2/1/2022 | 127 | 72 | 7 | 6 |
| 3/1/2022 | 181 | 73 | 7 | 9 |
| 4/1/2022 | 187 | 42 | 7 | 4 |
| 5/1/2022 | 226 | 81 | 7 | 10 |
| 6/1/2022 | 211 | 152 | 8 | 13 |
| 7/1/2022 | 205 | 263 | 8 | 7 |

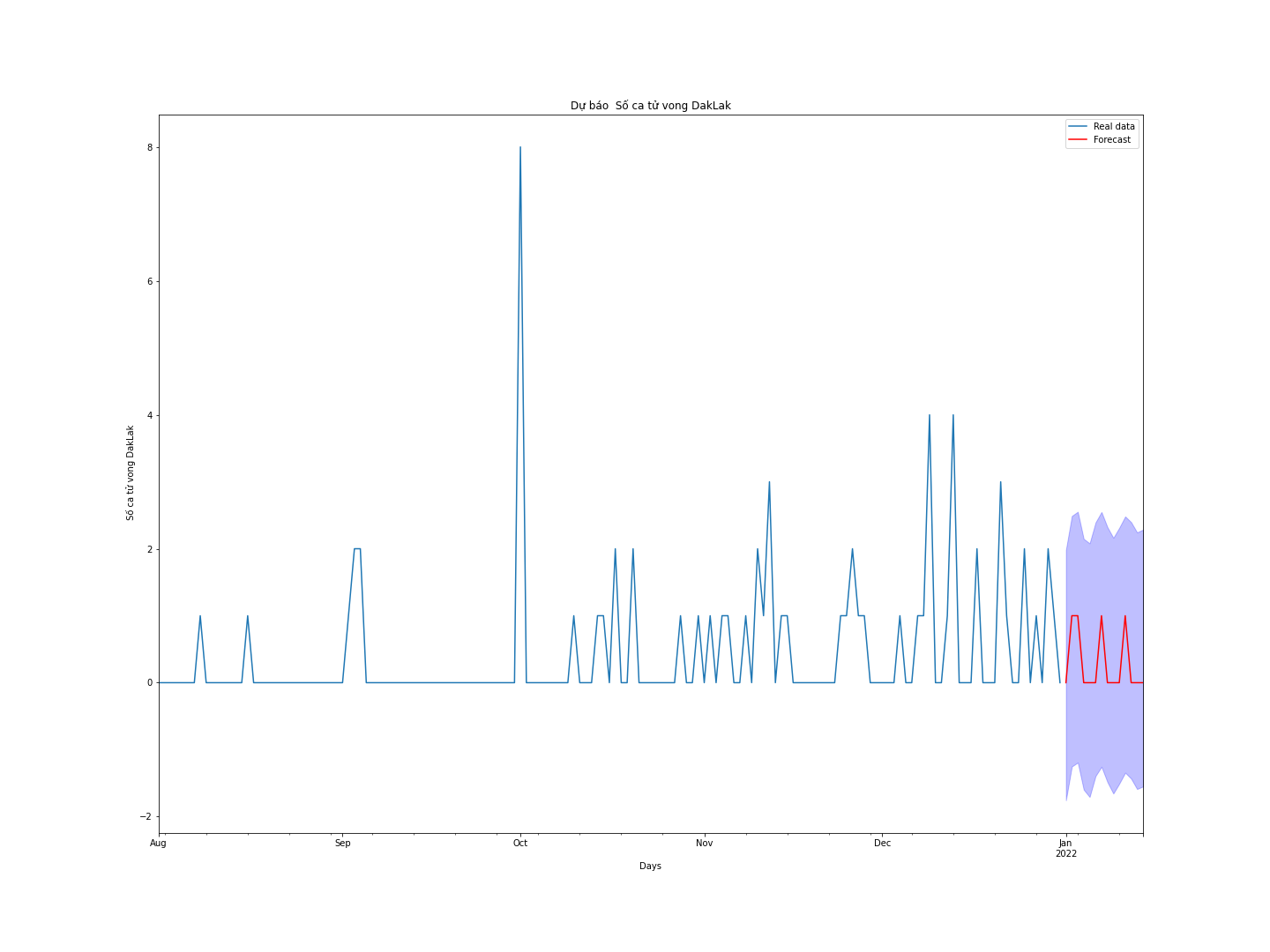
Bảng 5. So sánh kết quả dự đoán và số liệu thực tế của Bà Rịa Vũng Tàu.

Số liệu của tỉnh Bà Rịa Vũng Tàu có xu hướng tăng và mô hình dự báo tốt về xu hướng của dữ liệu này. Hơn nữa, có thể nhận thấy, số liệu cụ thể không quá sai lệch với dữ liệu thực tế, điều này cho thầy các tiếp cận này khá hiệu quả trong một số trường hợp.

***3. Đắk Lắk***



Hình 8. Biểu đồ thể hiện những ca mắc cũ và dự đoán tương ứng của Đắk Lắk.



Hình 9. Biểu đồ thể hiện những ca tử vong cũ và dự đoán tương ứng của Đắk Lắk.

| NGÀY | CA MẮC MỚI | | CA CHẾT MỚI | |
| --- | --- | --- | --- | --- |
| DỰ ĐOÁN | THỰC TẾ | DỰ ĐOÁN | THỰC TẾ |
| 1/1/2022 | 137 | 40 | 0 | 0 |
| 2/1/2022 | 117 | 185 | 1 | 1 |
| 3/1/2022 | 115 | 0 | 1 | 0 |
| 4/1/2022 | 126 | 367 | 0 | 0 |
| 5/1/2022 | 111 | 273 | 0 | 1 |
| 6/1/2022 | 124 | 0 | 0 | 0 |
| 7/1/2022 | 111 | 0 | 1 | 0 |

Bảng 4. So sánh kết quả dự đoán và số liệu thực tế của Đắk Lắk.

Ở Đắk Lắk, số liệu thực tế thấp hơn thành phố Hồ Chí Minh và không có tính ổn định, biểu hiện bằng việc số liệu cao thấp liên tục. Tuy nhiên có thể thấy mô hình vẫn dự báo tốt và có các trường hợp cao thấp tương tự như dữ liệu, nhưng vẫn chưa đáp ứng được hoàn toàn số liệu thực tế.

***4. Một số tỉnh khác***

Ngoài các tính trên, nhóm em đã thực nghiệm trên một số tỉnh khác như Hà Nội, Bắc Giang, Bình Dương, Cần Thơ. Tuy nhiên các tỉnh trên có số liệu không phù hợp để thực thi mô hình ARIMA vì vậy trong quá trình thử nghiệm đã phát sinh lỗi cũng như không có kết quả.

**VI. Kết luận**

Số ca mắc COVID-19 ở mỗi tỉnh thành có sự liên quan đến sự đông đúc của dân số, như mật độ dân số, hay tỉ lệ dân thành thị. Điều này dẫn đến việc phòng dịch ở các địa phương cần xem xét đến yếu tố mật độ dân số, thay đổi cường độ các biện pháp phòng dịch cho phù hợp với tình hình dân số cụ thể ở địa phương. Trong khi đó, việc giảm số ca tử vong cần phải xem xét thêm nhiều yếu tố khác, chẳng hạn như cơ sở vật chất, tỉ lệ nhân viên y tế trên dân số,...

Việc phân tích dự báo số ca dương tính, số ca tử vong mới dựa trên mô hình hồi quy những ca tử vong cũ có thể dự báo trước một kết quả ở mức trung bình và không sai lệch quá nhiều so với thực tế. Xu hướng của dự báo của mô hình đúng trong nhiều trường hợp. Tuy nhiên, mô hình có yêu cầu về dữ liệu nên có nhiều trường hợp thực tế không thể áp dụng mô hình trên.

**Tham khảo**

1. Bộ Y tế, Việt Nam. Số liệu ca mắc và tử vong do COVID-19 ở Việt Nam.
2. Tổng cục Thống kê. Số liệu dân số và mật độ dân số 63 tỉnh thành, 2021.
3. Nguyễn Văn Tuấn. Phân tích dữ liệu với R. Nhà xuất bản Tổng hợp TP.HCM, 2018.
4. R Documentation.<https://cran.r-project.org/manuals.html>
5. [AlessandroMinervini/Covid-19-Forecasting-Using-ARIMA-models: Covid-19 short-term forecasting using ARIMA models. (github.com)](https://github.com/AlessandroMinervini/Covid-19-Forecasting-Using-ARIMA-models)
6. [Khoa học dữ liệu (phamdinhkhanh.github.io)](https://phamdinhkhanh.github.io/2019/12/12/ARIMAmodel.html#32-kh%E1%BB%9Fi-t%E1%BA%A1o-chu%E1%BB%97i-l%E1%BB%A3i-su%E1%BA%A5t-v%C3%A0-kh%E1%BA%A3o-s%C3%A1t-d%E1%BB%AF-li%E1%BB%87u)
7. [Linear Regression - Hồi quy tuyến tính trong Machine Learning (viblo.asia)](https://viblo.asia/p/linear-regression-hoi-quy-tuyen-tinh-trong-machine-learning-4P856akRlY3)
8. [Machine Learning cơ bản (machinelearningcoban.com)](https://machinelearningcoban.com/2016/12/28/linearregression/)