LỜI CẢM ƠN

Qua quá trình học tập và nghiên cứu, được sự giúp đỡ nhiệt tình của các thầy cô giáo trường Đại học Công nghệ thông tin và truyền thông Thái Nguyên, Khoa Công nghệ thông tin, Phòng Đào tạo, tôi đã hoàn thành chương trình học tập và nghiên cứu luận văn với đề tài "***Tìm hiểu mô hình ARIMA phân tích dữ liệu y tế chuỗi thời gian***".

Tôi xin chân thành cảm ơn các thầy cô giáo trường Đại học Công nghệ thông tin và Truyền thông Đại học Thái Nguyên đã tạo điều kiện thuận lợi cho tôi trong quá trình học tập, nghiên cứu và hoàn thành luận văn.

Xin cảm ơn sự quan tâm, giúp đỡ chu đáo của Hội đồng khoa học, Ban Chủ nhiệm Khoa Công nghệ thông tin và các thầy cô giáo Phòng Đào tạo trường Đại học Công nghệ thông tin và Truyền thông - Đại học Thái Nguyên đã tạo mọi điều kiện thuận lợi và góp nhiều ý kiến quý báu cho luận văn.

Tôi xin trân trọng bày tỏ lòng biết ơn sâu sắc tới: TS. Trần Quang Quý - người Thầy đã tận tình hướng dẫn, chỉ bảo, động viên tôi trong suốt quá trình thực hiện luận văn, bổ sung cho tôi nhiều kiến thức chuyên môn và những kinh nghiệm quý báu trong nghiên cứu.

Cuối cùng, tôi xin bày tỏ lòng biết ơn và chia sẻ thành quả nhỏ bé này với tất cả những người thân trong gia đình tôi, bè bạn đã luôn động viên, giúp đỡ, tạo những điều kiện tốt nhất để tôi hoàn thành tốt chương trình học tập và thực hiện thành công luận văn này.

Thái Nguyên, ngày 24 tháng 6 năm 2023

Nguyễn Văn Cường

LỜI CAM ĐOAN

Tôi tên là: **Nguyễn Văn Cường**

Lớp: Cao học Khoa học máy tính K20

Tôi xin cam đoan đề tài luận văn thạc sỹ: "***Tìm hiểu mô hình ARIMA phân tích dữ liệu y tế chuỗi thời gian***" là do tôi thực hiện với sự hướng dẫn của TS. Trần Quang Quý. Đây không phải là bản sao chép của bất kỳ một cá nhân, tổ chức nào. Các số liệu, nguồn thông tin trong Luận văn là do tôi điều tra, trích dẫn và tham khảo.

Tôi xin hoàn toàn chịu trách nhiệm về những nội dung mà tôi đã trình bày trong Luận văn này.

Thái Nguyên, ngày 24 tháng 6 năm 2023

Người viết cam đoan

Nguyễn Văn Cường

MỤC LỤC

LỜI CẢM ƠN i
LỜI CAM ĐOAN ii
LỜI MỞ ĐẦU 7
CHƯƠNG 1 : PHÂN TÍCH DỮ LIỆU CHUỖI THỜI GIAN 9
1.1. Khái niệm về chuỗi thời gian 9
1.2. Các thành phần của chuỗi thời gian 9
1.3. Tính chất của dữ liệu chuỗi thời gian 11
1.4. Tính dừng của dữ liệu chuỗi thời gian 14
1.4.1. Tính dừng 14
1.4.2. Kiểm tra tính dừng chuỗi thời gian 15
1.4.3. Biến đổi chuỗi không dừng thành chuỗi dừng 17
1.5. Các chỉ số liên quan: Tự tương quan và tương quan chéo 18
1.6. Hồi quy cổ điển trong chuỗi thời gian 21
1.7. Các chỉ số để lựa chọn mô hình 24
1.7.1. AIC - Akaike information criterion 24
1.7.2. BIC 25
1.8. Phân tích dữ liệu khám phá 26
CHƯƠNG 2: CÁC MÔ HÌNH ARIMA 28
2.1. Sai Phân 28
2.2. Các mô hình tự hồi quy AR 29
2.3. Mô hình trung bình trượt MA 31
2.4. Mô hình trung bình trượt và tự hồi quy ARMA 32
2.5. Mô hình trung bình trượt tự hồi quy ARIMA 33
2.6. Các bước phân tích dữ liệu chuỗi thời gian với mô hình ARIMA 35
CHƯƠNG 3: MÔ HÌNH ARIMA DỰ ĐOÁN DỮ LIỆU COVID-19 46
3.1. Giới thiệu dữ liệu Covid-19 46

3.2. Thu thập và tiền xử lý dữ liệu 47
3.3. Dữ liệu Covid-19 Việt Nam 52
3.4. Xây dựng mô hình ARIMA dự đoán 54
3.5. Dự đoán 57
TÀI LIỆU THAM KHẢO 63
PHỤ LỤC 65

DANH MỤC HÌNH

Hình 1 Số liệu diễn biến cúm tại phía Nam châu Phi từ 2006-2015.....9
Hình 2 Biểu diễn xu hướng giảm của dữ liệu 9
Hình 3 Biểu diễn thay đổi chuỗi theo từng khoảng 10
Hình 4 Biểu diễn chu kỳ chuỗi thời gian 10
Hình 5 Sơ đồ chuỗi với định lượng Y(t) diễn tiến theo thời gian t.....11
Hình 6 Lợi nhuận hàng quý của Johnson & Johnson.....12
Hình 7 Biểu đồ nhiệt độ toàn cầu qua các năm.....12
Hình 8 Dữ liệu về tần số âm thanh.....13
Hình 9 Dữ liệu chuỗi thời gian tài chính.....14
Hình 10 Đồ thị ACF 17
Hình 11 Đồ thị PACF.....17
Hình 12 Chuỗi có nhiễu trắng 19
Hình 13 Chuỗi được làm mịn.....19
Hình 14 Trực quan dữ liệu toàn cầu.....23
Hình 15 Giá cổ phiếu công ty Amazon sử dụng AR 29
Hình 16 Giá cổ phiếu Amazon sử dụng trung bình động 31
Hình 17 Các bước chính trong phương pháp Box-Jenkins.....42
Hình 18 Miêu tả dữ liệu Covid-19 47
Hình 19 Tóm tắt các nước có số ca nhiễm nhiều nhất 47
Hình 20 Thống kê ca nhiễm và tử vong các nước có tỷ lệ cao 48
Hình 21 Biểu đồ tích lũy từ tháng 06/2020 đến tháng 01/2023 49
Hình 22 Biểu đồ Treemap theo các quốc gia 49
Hình 23 Tóm tắt số liệu vắc xin theo quốc gia 50
Hình 24 Biểu đồ tương quan giữa tỷ lệ tiêm vắc xin và quy mô dân số các quốc gia..... 50
Hình 25. Dữ liệu Covid thu được.....51
Hình 26. Dữ liệu Covid sau khi tiền xử lý 51
Hình 27. Kiểm định dữ liệu.....52
Hình 28 Trực quan dữ liệu Covid-19 tại Việt Nam từ 02/2022 đến 04/2023.....53
Hình 29 Tách dữ liệu từ 02/2022-05/2022.....53

Hình 30 Các biểu đồ ACF và PACF 54
Hình 31 Kết quả dự đoán 57
Hình 32 Biểu đồ so sánh giá trị thực tế và dự đoán 58
Hình 33. Dự đoán với dữ liệu Ấn Độ.....59
Hình 34 Dự đoán với dữ liệu Brazil.....59
Hình 35 Kết quả dự đoán 73

LỜI MỞ ĐẦU

Chuỗi thời gian là một lĩnh vực quan trọng trong phân tích dữ liệu, đặc biệt là trong lĩnh vực dự báo và dự đoán. Việc nghiên cứu và xây dựng mô hình chuỗi thời gian có vai trò quan trọng trong việc hiểu và dự đoán sự biến động của các hiện tượng theo thời gian.

Trong toán học, dữ liệu chuỗi thời gian được định nghĩa là những điểm dữ liệu đã được đánh chỉ số theo thời gian và có khoảng cách đều nhau giữa những quan sát liên tiếp. Đó có thể là dữ liệu về giá chứng khoán hàng ngày, tổng thu nhập quốc dân của một quốc gia hàng năm, tổng doanh số công ty hàng quí,…

Ưu điểm của chuỗi thời gian là nó có thể lưu trữ được trạng thái của một trường dữ liệu theo thời gian. Trong khi đó thế giới luôn vận động, các sự vật, hiện tượng hiếm khi dừng lại ở trạng thái tĩnh mà thường thay đổi. Do đó dữ liệu chuỗi thời gian có tính ứng dụng rất cao và được áp dụng trong rất nhiều lĩnh vực khác nhau như: thống kê, kinh tế lượng, toán tài chính, dự báo thời tiết, dự đoán động đất, điện não đồ, kỹ thuật điều khiển, thiên văn, kỹ thuật truyền thông, xử lý tín hiệu.

Mô hình ARIMA có tên tiếng Anh là Autoregressive Integrated Moving Average, đây là mô hình quan trọng trong việc phân tích và sử dụng để dự đoán dữ liệu chuỗi thời gian. Mô hình này lần đầu tiên được đưa ra bởi Box & Jenkins (1970). ARIMA được kết hợp bởi 3 thành thành phần chính: AR (thành phần tự hồi quy), I (tính dừng của chuỗi thời gian) và MA (thành phần trung bình trượt). Theo Gujarati (2004), để ước lượng mô hình ARIMA ta cần đi qua 4 bước chính sau:

Bước 1: Nhận dạng mô hình

Bước 2: Ước lượng các tham số và lựa chọn mô hình

Bước 3: Kiểm định mô hình

Bước 4: Dự báo

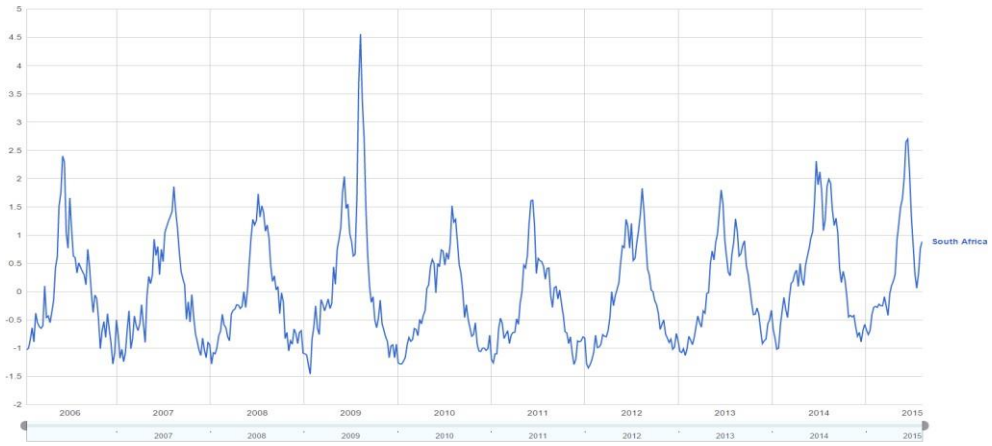
Trong bối cảnh đại dịch Covid-19 đang lan rộng trên toàn thế giới, việc dự đoán số ca nhiễm là một vấn đề cấp bách và có tính thiết yếu. Nội dung quyển luận văn này nhằm mục đích trình bày quá trình nghiên cứu và xây dựng mô hình ARIMA để dự đoán số ca nhiễm COVID-19 tại Việt Nam. Trong đó sẽ sử dụng các dữ liệu về

số ca nhiễm đã ghi nhận trong quá khứ để xây dựng mô hình và tiến hành dự đoán số ca nhiễm trong tương lai. Từ việc phân tích mô hình, thu thập dữ liệu và đưa ra các nhận xét sẽ rút ra được cái nhìn tổng quan về xu hướng của dữ liệu, từ đó đưa ra được các khuyến nghị. Dữ liệu đang đề cập ở đây là dữ liệu Covid-19, một dạng dữ liệu điển hình trong lĩnh vực y tế dự phòng.

CHƯƠNG 1 : PHÂN TÍCH DỮ LIỆU CHUỖI THỜI GIAN

1.1. Khái niệm về chuỗi thời gian

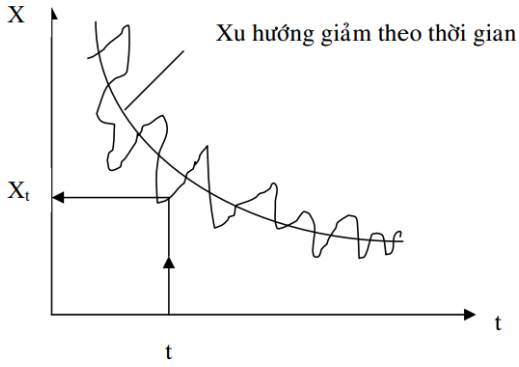
Chuỗi thời gian là một tập hợp các giá trị được ghi nhận tại các thời điểm khác nhau, có thể được sử dụng để mô tả các biến đổi theo thời gian. Các ví dụ về chuỗi thời gian bao gồm số lượng sản phẩm bán ra hàng tháng, giá cổ phiếu theo ngày, nhiệt độ theo giờ, và số lượng ca nhiễm Covid-19 hàng ngày.



Hình 1 Số liệu diễn biến cúm tại phía Nam châu Phi từ 2006-2015

1.2. Các thành phần của chuỗi thời gian

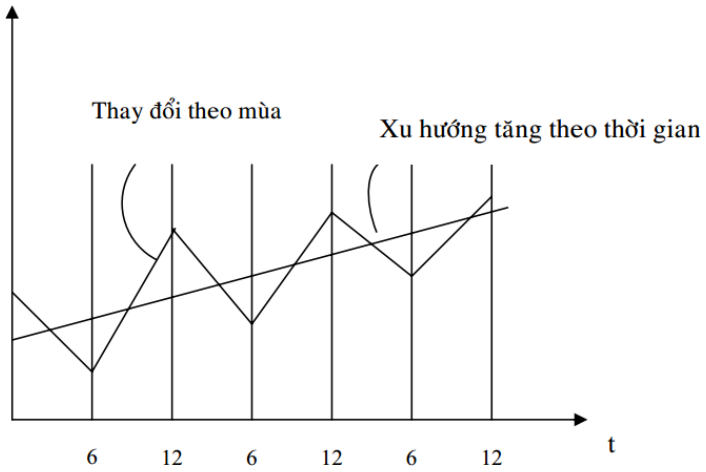
Dữ liệu chuỗi thời gian có các thành phần cơ bản như: thành phần xu hướng; thành phần mùa (thời vụ); thành phần chu kỳ (dài hạn); các điểm bất thường và ngẫu nhiên.



Hình 2 Biểu diễn xu hướng giảm của dữ liệu

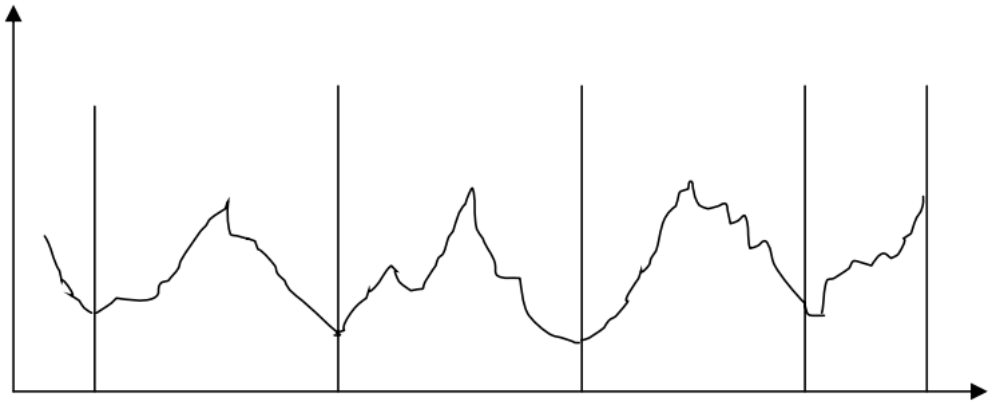
- ***Thành phần xu hướng***: để chỉ xu hướng tăng hay giảm của dữ liệu y trong chuỗi thời gian. Thành phần xu hướng này thường được biểu diễn trên đồ thị bởi

một đường thẳng hay đường cong trơn. Chuỗi dữ liệu không tồn tại thành phần xu hướng (tức là dữ liệu không tăng hoặc không giảm) thì chuỗi đó dừng theo giá trị trung bình.



Hình 3 Biểu diễn thay đổi chuỗi theo từng khoảng

- ***Thành phần mùa (thời vụ):*** để chỉ chiều hướng tăng hay giảm của giá trị y được tính theo giai đoạn thời gian (khoảng thời gian ngắn). Ví dụ: số lượng trẻ em mắc các bệnh về hô hấp tăng lên vào dịp cao điểm rét đậm, rét hại ở nước ta.



Hình 4 Biểu diễn chu kỳ chuỗi thời gian

- ***Thành phần chu kì (dài hạn):*** biểu thị bằng sự tăng, giảm của dữ liệu chuỗi thời gian xoay quanh xu hướng. Thường trong chuỗi dữ liệu dài hạn thì khó đoán chu kì.