**Lesson 12 Time Series Analysis**

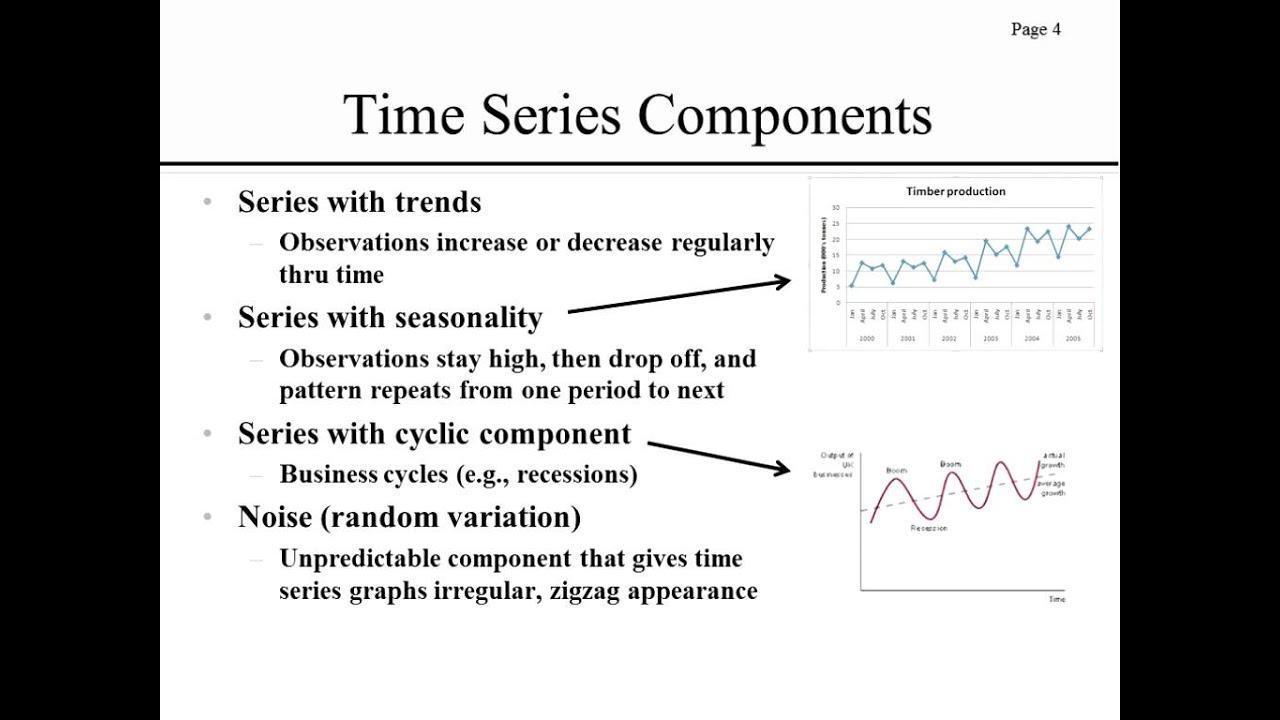
1. **Lý thuyết**

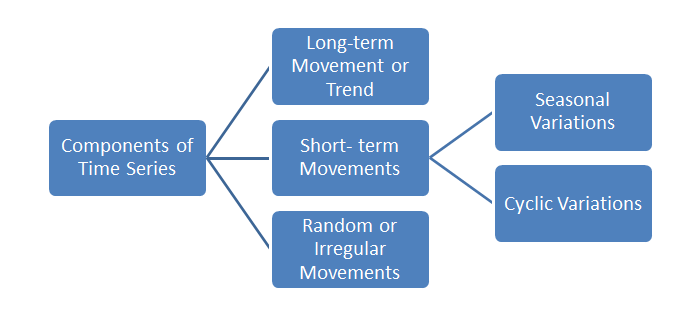
Rất dễ nhầm lẫn giữa time series analysis và thuật toán regression.

Có rất nhiều người mới học Machine Learning muốn sử dụng Machine Learning để phân tích giá cổ phiếu , giá vàng ,... Và các bạn sử dụng Regression để phân tích.

Tuy nhiên, việc phân tích một giá trị dựa vào thời gian không phải sử dụng thuật toán Regression , mà nó là riêng 1 mảng Time Series Analysis.

Trong đó, Time Series chính là khoảng thời gian mà chúng ta cần phân tích + dự đoán. Nói chung quy là phân tích sự tăng tiến , biến động của đối tượng cần phân tích trong 1 khoảng thời gian để rút ra những quy luật và áp dụng quy luật đó cho việc phân tích trong tương lai



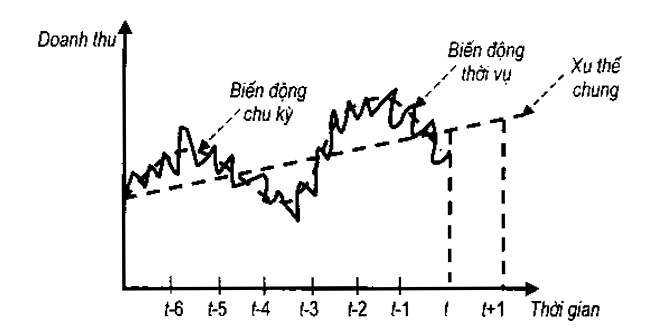


Những thuật toán dùng để phân tích Time Series tiêu biểu : ARIMA , SARIMA , ARIMAX , Holt-Winters exponential smoothing... ngoài ra , còn có các phương pháp forecasting của tài chính như Moving Average , Exponential Moving Average, ...hoặc thậm chí deep learning như LSTM

Nhưng nhìn chung, việc phân tích dữ liệu theo thời gian bị ảnh hưởng bởi những yếu tố sau :

* Xu hướng ( tăng hay giảm , hay side way )
* Yếu tố mùa vụ ( seasonal )
* Nhiễu ngẫu nhiên ( thi thoảng sẽ có yếu tố tăng giảm bất ngờ , xác suất ntn ,... )

**Tổng quan**

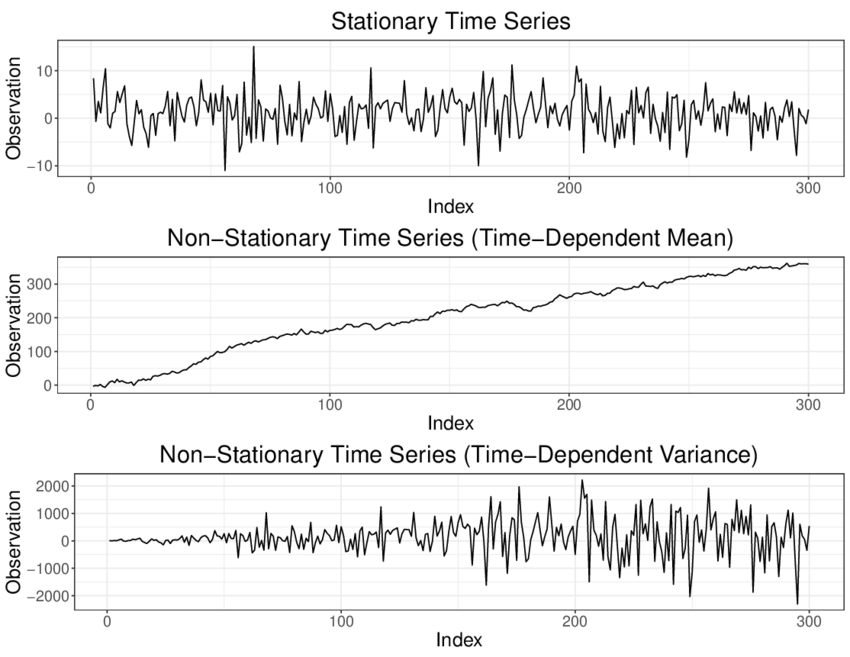


**Seasonal**

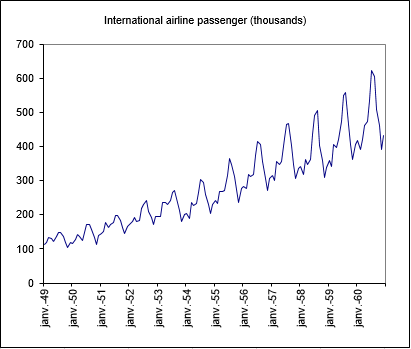
Chart, line chart

Description automatically generated

**Stationary ( không đổi )**



Trend



1. **Cách làm**

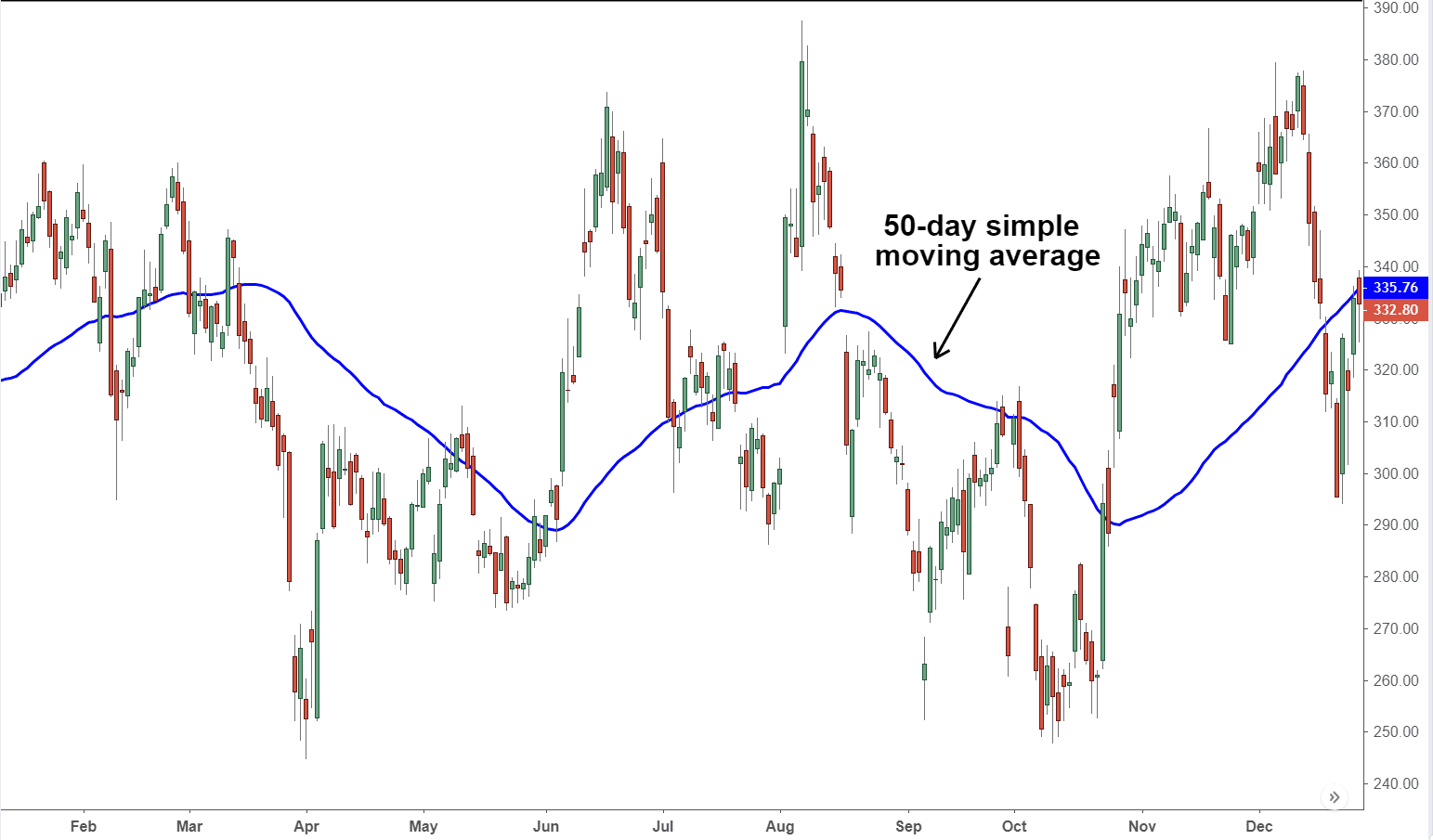
* Công việc đầu tiên đương nhiên vẫn là visualize dữ liệu , để thấy được bức tranh tổng quan, sau khi thấy được bức tranh tổng quan , chúng ta sẽ dễ chọn ra model phù hợp.
* Stationary test , để xem dữ liệu có stationary không , có nhiều loại test , Dickey Fuller, Autocorrelation Function , hoặc thậm chí việc nhìn vào data sau khi visualize cũng nói lên phần nào
* Diagram

  Description automatically generated with medium confidence
* Hình d , h , g thì ta thấy rõ sự seasonality
* A, c, e, f, i ta thấy rõ trendline
* B và g dao động với mean, standard deviation , variance ,... hầu như không đổi => có thể có sự stationary .
* Hình i dao động càng ngày càng lớn => variance tăng dần đều.
* Chọn thuật toán phù hợp : moving average , exponential smoothing hoặc ARIMA.

1. **Thuật toán**

**3.1 Moving Average**

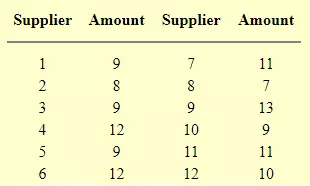
Với moving average, chỉ đơn giản là dự đoán giá trị mới bằng cách lấy giá trị trung bình của các dữ liệu quá khứ



Đây là phương pháp forecast cơ bản nhất. Nó cũng được dùng để dự đoán trendline.

Chúng ta sẽ cùng xem xét 1 số phương pháp trung bình, chẳng hạn như phương pháp trung bình đơn giản với các dữ liệu trong quá khứ.

1 quản lý nhà băng muốn biết khách hàng điển hình sẽ gửi tiết kiệm bao nhiêu tiền(đơn vị 1000$). Anh ta chọn 12 khách hàng bất kì và cho kết quả sau:

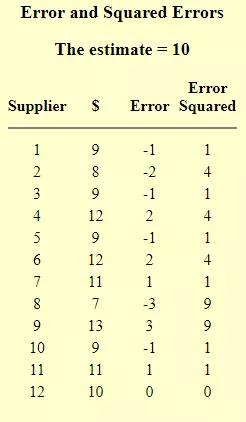


Giá trị trung bình mà máy tính tính ra là 10. Người quản lý quyết định sử dụng số liệu này làm dự toán chi tiêu của một nhà cung cấp thông thường.

Nhưng đây là cách tính toán tốt hay cách tính toán tồi?

Chúng ta sử dụng phương pháp “Bình phương bình quân lỗi”:

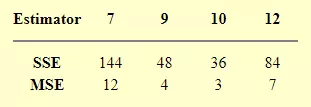
* "error" = số tiền thực đã trừ trừ số tiền ước tính.
* "error squared" là lỗi ở trên, bình phương.
* Các "SSE" là tổng của các bình phương lỗi.
* Các "MSE" là trung bình của các bình phương lỗi.  
  Ta có kết quả:



SSE=36 và MSE=36/12=3

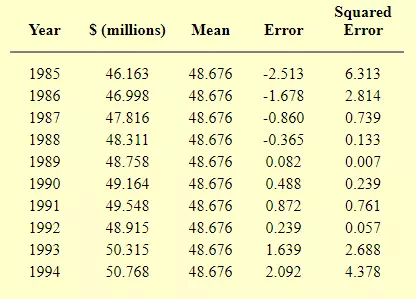
Vậy ước tính của số tiền chi cho mỗi nhà cung cấp là bao nhiêu? Hãy so sánh ước tính (10) với các ước tính sau: 7, 9 và 12. Tức là, chúng ta ước tính rằng mỗi nhà cung cấp sẽ chi tiêu $ 7, hoặc $ 9 hoặc $ 12.

Tiếp tục tính toán chúng ta sẽ có bảng sau:



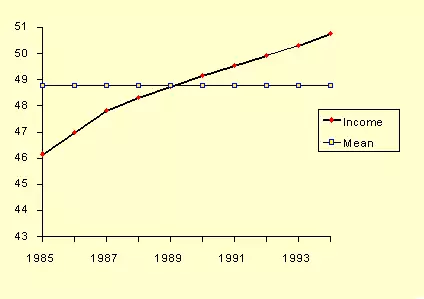
Ước tính tốt nhất ở đây là ước tính với MSE thấp nhất nên trong trường hợp này thì 10 là OK

Bảng tiếp theo đây sẽ là thống kê doanh thu chưa qua thuế của các cty máy tính từ 1985-1994:

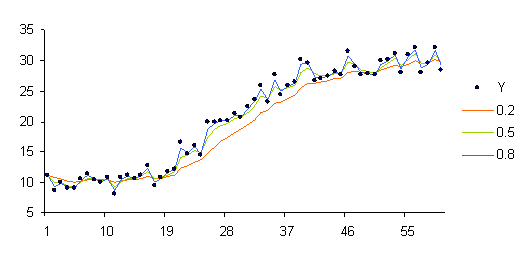


MSE = 1.8129

Chúng ta có thể dựa vào số liệu trên để có 1 biểu đồ dự đoán xu hướng



**3.2 Exponential Smoothing**

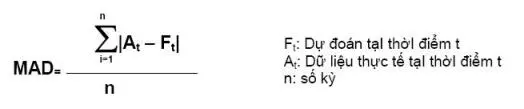


Phương Pháp exponential Smoothing cũng gần giống như moving average, tuy nhiên , nó sử dụng một giá trị gọi là hằng số liên tiến để diễn đạt cái sự “smooth” ( tức sự chênh lệch giữa thực tế và dự đoán, vì các điểm dự đoán sẽ có sự sai lệch với nhau rất nhiều, nếu dự đoán chính xác quá thì alpha sẽ cao, và cái đường dự đoán sẽ không smooth ). Alpha là một hằng số và có giá trị từ 0 – 1.



Hằng số liên tiến (alpha) thể hiện độ đáp ứng đối với sự chênh lệch giữa dự đoán và thực tế. Hằng số này được quyết định bởi bản thân của sản phẩm cũng như cảm nhận của nhà quản lý. Chẳng hạn, một nhà sản xuất một sản phẩm mà nhu cầu tương đối ổn định thì tỷ lệ đáp ứng với sự chênh lệch giữa dự đoán và thực tế sẽ nhỏ. Nếu nhu cầu tăng trưởng nhanh, tỷ lệ này sẽ tăng cao. Hằng số này có giá trị từ 0 đến 1. Ðể tìm số alpha chính xác nhất thì phảI dùng phương pháp thử sai, alpha sẽ là giá trị mà làm cho độ lệch trị tuyệt đốI trung bình (Mean Absolute Deviation-MAD) là nhỏ nhất.

MAD được tính bằng trung bình cộng của độ lệch giữa dự đoán và thực tế, sau khi đã lấy trị tuyệt đối.

Công thức tính MAD

**3.3 ARIMA (Auto Regressive Integrated Moving Average )**

Auto Regressive là sử dụng giá trị quá khứ để dự đoán giá trị tương lai, còn Moving Average thì các bạn đã biết.

Auto Regressive là một model có dạng như linear regression mà các bạn đã biết y = ax + b

Tuy nhiên với auto regressive thì y tại 1 điểm thời gian t sẽ được tính dựa vào các lag thời gian trước đó của nó

Text

Description automatically generated

Còn Moving average thì có dạng

Diagram, schematic

Description automatically generated

Vậy ARIMA là 1 model có dạng combine giữ MA và AR

Text

Description automatically generated

( các bạn chỉ cần biết và hiểu phần này là được, không cần phải quá nặng về toán )

**ARIMA model in words:**

Predicted Yt = Constant + Linear combination Lags of Y (upto p lags) + Linear Combination of Lagged forecast errors (upto q lags)

The objective, therefore, is to identify the values of p, d and q

Với d là differencing , p là đại lượng của hàm AR , q là đại lượng của hàm MA

Với python thì đã có thư viện auto arima hỗ trợ các bạn tìm ra p , d và q fit nhất với dữ liệu của các bạn. Nên với buổi học này thì chúng ta chỉ cần hiểu thuật toán và biết cách sử dụng là được.

Tóm lại,

Có rất nhiều cách để dự đoán , nhưng điều đầu tiên là chúng ta phải nhận biết được đặc trưng của dữ liệu để chọn cách tốt nhất. Thậm chí đôi khi phải sử dụng nhiều cách để so sánh và lấy ra kết quả tốt nhất.