

DATA ANALYTICS

Unit 2:Confusion matrices and Metrics

Mamatha H R

Department of Computer Science and Engineering

The confusion matrix is a metric that is often used to measure the performance of a classification algorithm.

In binary classifiers as with the spam filtering example, in which each email can be either spam or not spam.

The confusion matrix will be of the following form:

	Predicted: Real Email	Predicted: Spam Email
Actual: Real Email	True Negatives (TN)	False Positives (FP)
Actual: Spam Email	False Negatives (FN)	True Positives (TP)

Accuracy: Out of all the classes, how much we predicted correctly

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

The ability of the model to correctly classify positives and negatives are called sensitivity and specificity

Sensitivity =

$$\frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Negative (FN)}}$$

Specificity =

$$\frac{\text{True Negative (TN)}}{\text{True Negative (TN)} + \text{False Positive (FP)}}$$

Precision =

$$\frac{\text{True Positive (TP)}}{\text{True Positive (TP)} + \text{False Positive (FP)}}$$

$$\text{F - Score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Confusion matrix to compare two classifiers

- Which is the better classification model (wrt Class A)?

		Predicted	
		A	A'
Actual	A	60	34
	A'	1	12

Model 1

$$\begin{aligned}\text{Recall}(A) &= 60/94 \\ \text{Precision}(A) &= 60/61 \\ \text{F1_score}(A) &= 2RP/(R+P) \\ &= 0.774 \\ \text{Accuracy}(\text{Model1}) & \\ &= 0.673\end{aligned}$$

$$\begin{aligned}\text{Recall}(A') &= 12/13 \\ \text{Precision}(A') &= 12/34 \\ \text{F1_score}(A') &= 0.51\end{aligned}$$

		Predicted	
		A	A'
Actual	A	90	4
	A'	8	5

Model 2

$$\begin{aligned}\text{Recall}(A) &= 90/94 \\ \text{Precision}(A) &= 90/98 \\ \text{F1_score}(A) &= 2RP/(R+P) \\ &= 0.937 \\ \text{Accuracy}(\text{Model2}) & \\ &= 0.888\end{aligned}$$

$$\begin{aligned}\text{Recall}(A') &= 5/13 \\ \text{Precision}(A') &= 5/9 \\ \text{F1_score}(A') &= 0.45\end{aligned}$$

Confusion matrices for multiple classes

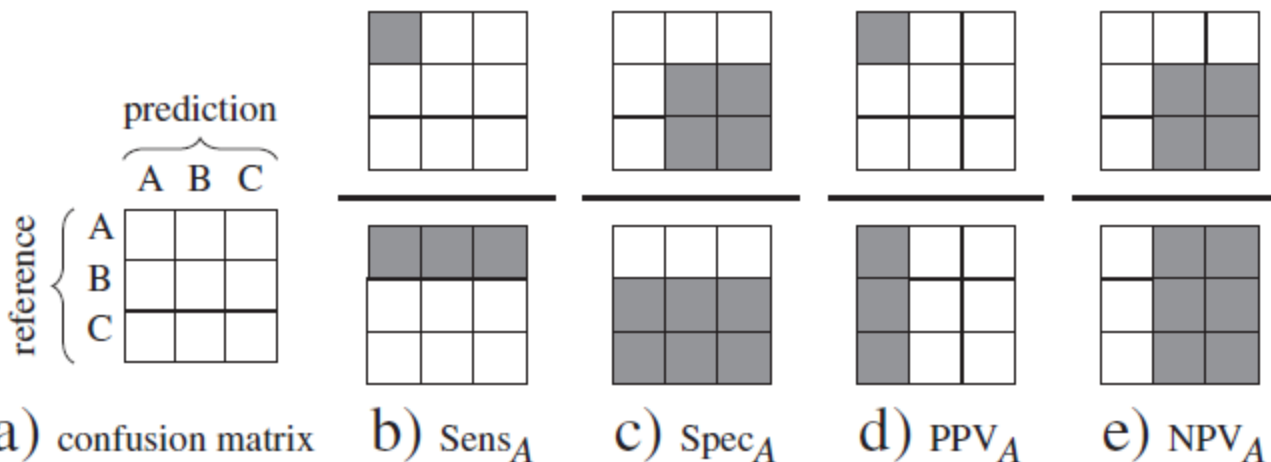
- What is Recall(A)?
- What is Specificity(B)?
- What is Precision(C)?
- What is the average accuracy of this model?

	A	B	C
A			
B			
C			

		Predicted class		
		Cat	Dog	Rabbit
Actual class:	Cat	5	3	0
	Dog	2	3	1
	Rabbit	0	2	11

		Predicted class	
		Cat	NotCat
Actual class:	Cat	5	3
	NotCat	2	17

		Predicted class	
		Dog	NotDog
Actual class:	Dog	3	3
	NotDog	5	16



Revisiting the example

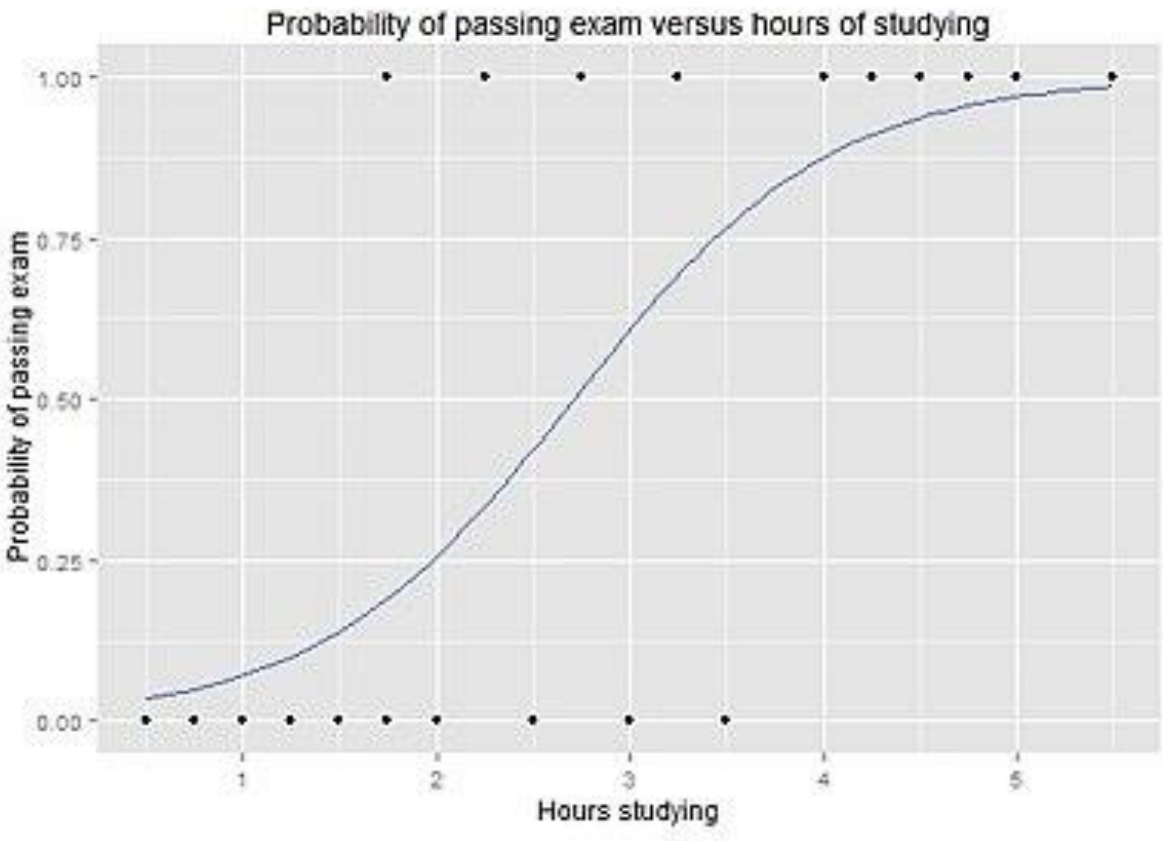


Hours of study	Passing exam		
	Log-odds	Odds	Probability
1	-2.57	0.076 \approx 1:13.1	0.07
2	-1.07	0.34 \approx 1:2.91	0.26
3	0.44	1.55	0.61
4	1.94	6.96	0.87
5	3.45	31.4	0.97

One hr of study increases log odds of passing by 1.5046

	Coefficient	Std.Error	P-value (Wald)
Intercept	-4.0777	1.7610	0.0206
Hours	1.5046	0.6287	0.0167

Hours	0.50	0.75	1.00	1.25	1.50	1.75	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50	4.00	4.25	4.50	4.75	5.00	5.50
Pass	0	0	0	0	0	0	1	0	1	0	1	0	1	0	1	1	1	1	1	1



https://en.wikipedia.org/wiki/Logistic_regression

Concordant and Discordant Pairs

- **Discordant Pairs.** A pair of positive and negative observations for which the model has no cut-off probability to classify both of them correctly are called discordant pairs.
- **Concordant Pairs.** A pair of positive and negative observations for which the model has a cut-off probability to classify both of them correctly are called concordant pairs.
- Divide the dataset into positives ($y=1$) and negatives ($y=0$).
- For a randomly chosen positive and negative, if the probability of positive (obtained using logistic regression model) is greater than probability of negative then such pairs are called concordant pairs.
- For a randomly chosen positive and negative, if the probability of positive is less than probability of negative then such pairs are called discordant pairs.
- Area under the ROC curve is the proportion of concordant pairs in the dataset.

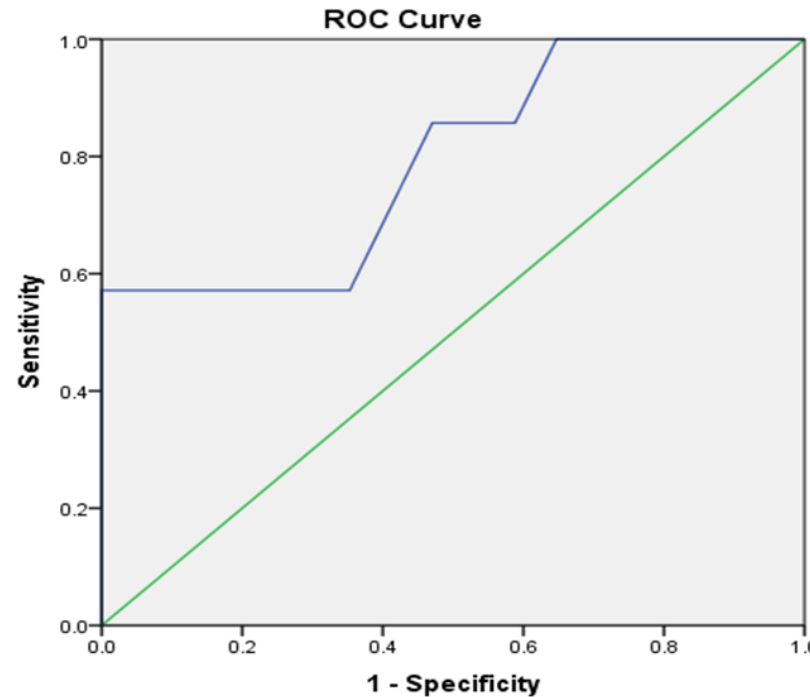
Hours of study	Passing exam	
	Probability	Label
1	0.070	0
2	0.260	0
3	0.610	0
4	0.870	1
5	0.950	1
6	0.970	1
7	0.980	0

(1,5): concordant pair

(4,7): discordant pair

Receiver Operating Characteristics (ROC) Curve

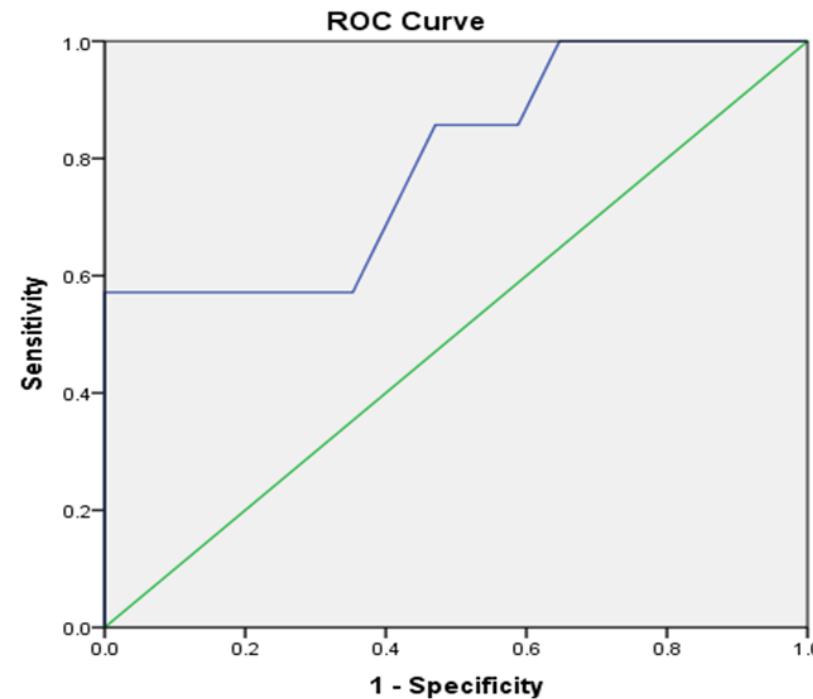
- **ROC curve** is a plot between sensitivity (true positive rate) in the vertical axis and $1 - \text{specificity}$ (false positive rate) in the horizontal axis.
- The higher the area under the ROC curve, the better the prediction ability.



Diagonal segments are produced by ties.

Receiver Operating Characteristics (ROC) Curve

- **ROC curve** is a plot between sensitivity (true positive rate) in the vertical axis and $1 - \text{specificity}$ (false positive rate) in the horizontal axis.
- The higher the area under the ROC curve, the better the prediction ability.

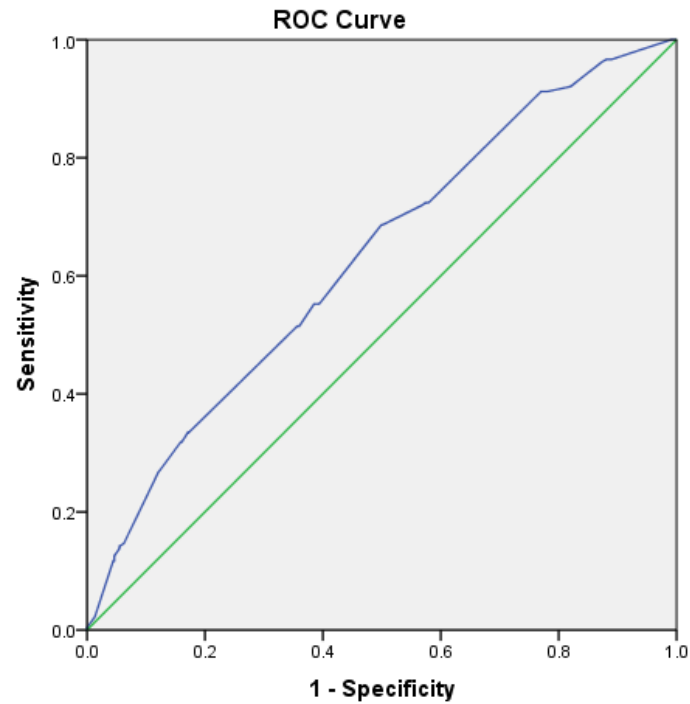


Diagonal segments are produced by ties.

- Area under the ROC (AUC) curve is interpreted as the probability that the model will rank a randomly chosen positive higher than randomly chosen negative.
- If n_1 is the number of positives (1s) and n_2 is the number of negatives (0s), then the area under the ROC curve is the proportion of cases in all possible combinations of (n_1, n_2) such that n_1 will have higher probability than n_2 .

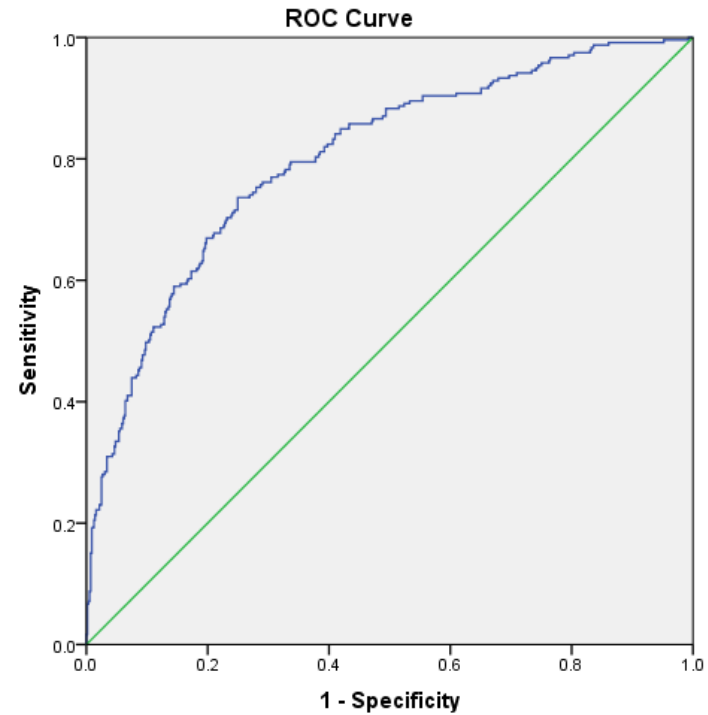
AUC = P (Random Positive Observation) > P(Random Negative Observation)

Area Under the ROC Curve (AUC) is a measure of the ability of the logistic regression model to discriminate positives and negatives correctly.



Diagonal segments are produced by ties.

AUC = 0.629



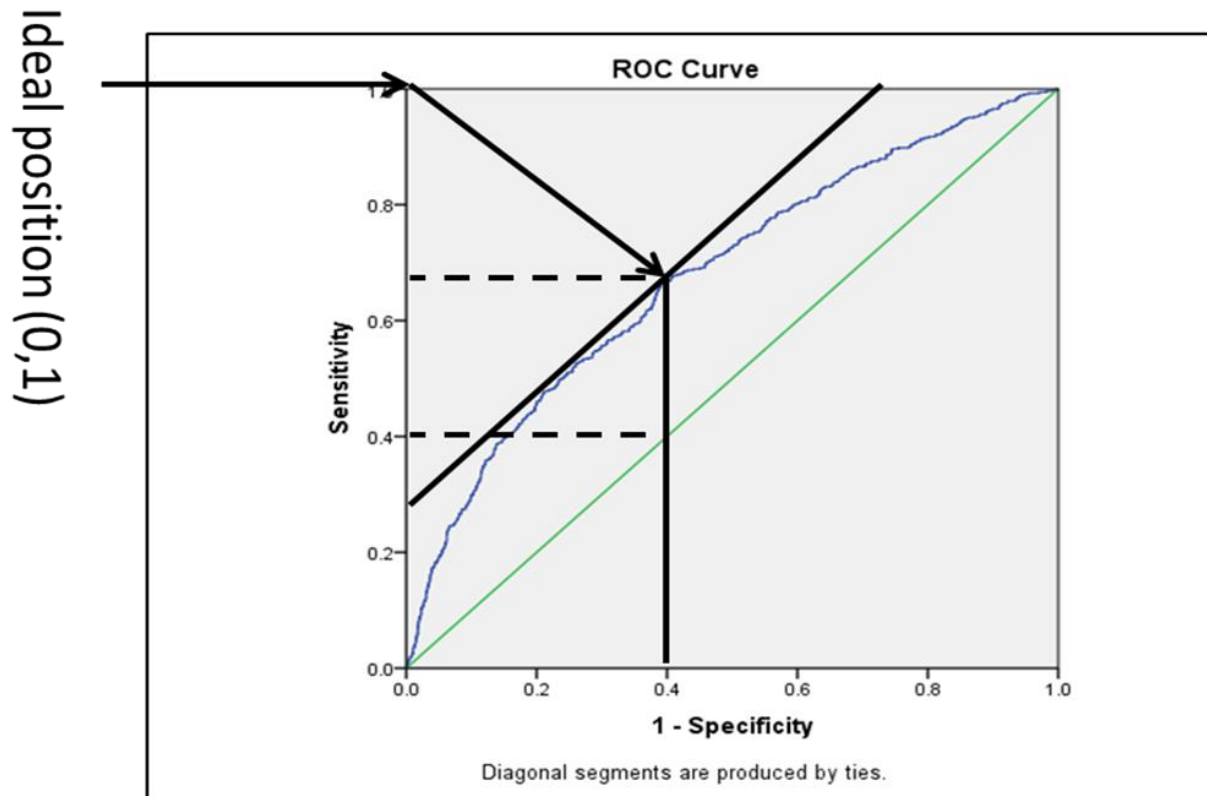
AUC = 0.801

- General rule for acceptance of the model:
- If the area under ROC is:
 - $0.5 \Rightarrow$ No discrimination
 - $0.7 \leq \text{ROC area} < 0.8 \Rightarrow$ Acceptable discrimination
 - $0.8 \leq \text{ROC area} < 0.9 \Rightarrow$ Excellent discrimination
 - $\text{ROC area} \geq 0.9 \Rightarrow$ Outstanding discrimination

Youden's Index for Optimal Cut-Off Probability

Youden's Index (1950) is a classification cut-off probability, for which the following function is maximized (also known as J statistic):

$$\text{Youden's Index} = \text{J Statistic} = \max_P [\text{Sensitivity}(p) + \text{Specificity}(p) - 1]$$



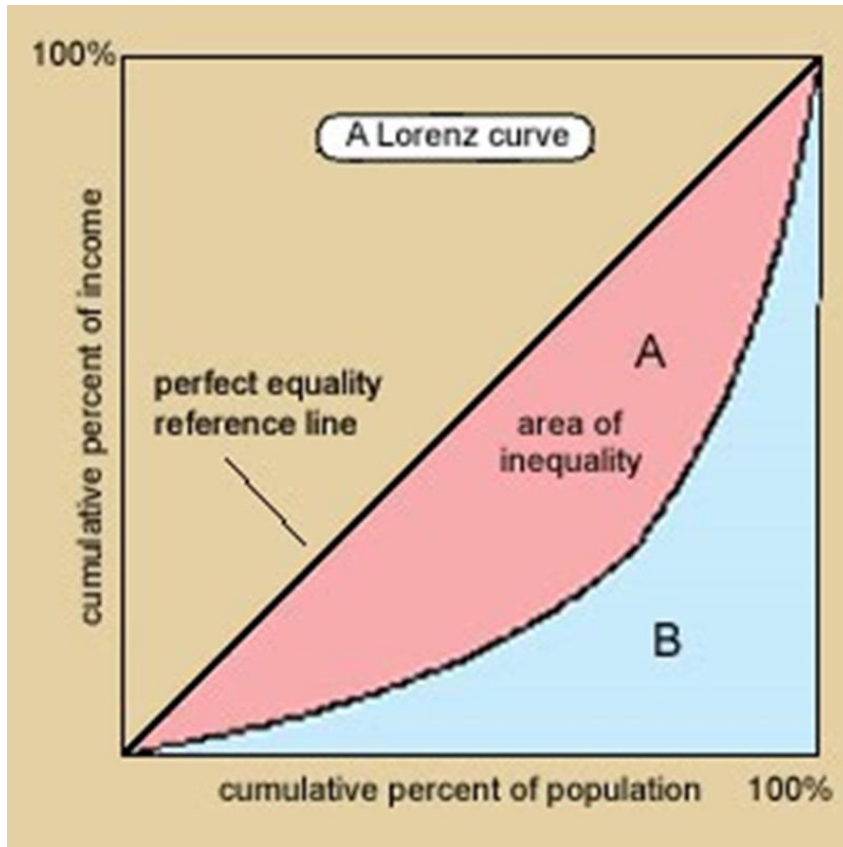
Cost-Based Cut-Off Probability

In cost-based approach, we assign penalty cost for misclassification of positives and negatives. Assume that cost of misclassifying negative (0) as positive (1) is C_{01} and cost of misclassifying positive (1) as negative (0) is C_{10} as shown in Table

Observed	Classified	
	0	1
0	---	C_{01}
1	C_{10}	---

The optimal cut-off probability is the one which minimizes the total penalty cost and is given by

$$\underset{p}{\text{Min}} [C_{01}P_{01} + C_{10}P_{10}]$$



Gini Index is a statistical measure of dispersion

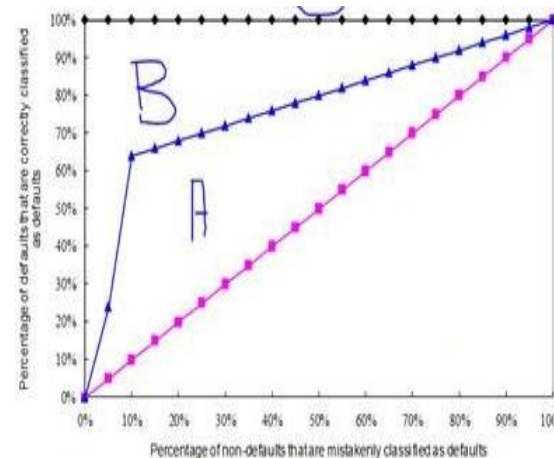
$$\text{Gini Coefficient} = A / (A+B)$$

$$\text{Gini Coefficient} = 2 \text{ AUC} - 1$$

AUC = Area Under the ROC Curve

- Gini coefficient measures individual impact of the an explanatory variable.
- $\text{Gini coefficient} = 2 \text{ AUC} - 1$
- AUC = Area under the ROC Curve

- Can discordant pairs be thought of as outliers?
 - Indeed, discordancy tests are used to detect outliers and one or both the points in the discordant pair could be outliers.
 - However, we must also be aware there are other possibilities:
 - (a) The parameters (coefficients) could be better estimated
 - (b) The current model (logistic regression) is not suitable for modeling the data on hand (some preprocessing may be required before we can model the data using Logistic Regression or we could explore alternatives)
- Why is Gini coefficient = $(2 \text{ AUC} - 1)$?
 - (AUC = Area under the ROC Curve)
 - $\text{Gini} = A/(A+B)$
 - A = area under the curve and the diagonal
 - B = area under the perfect model and diagonal
 - $\text{Gini (in RoC)} = A/(A+B) = A/0.5 = 2A$
 - $\text{AUC for this case} = A + \frac{1}{2}$
 - $\text{AUC} = \text{Gini}/2 + \frac{1}{2}$



Text Book:

“Business Analytics, The Science of Data-Driven Decision Making”, U. Dinesh Kumar, Wiley 2017 (Ch. 11.1-11.4, 11.6.5, 11.7.2-11.7.3)

<https://towardsdatascience.com/train-validation-and-test-sets-72cb40cba9e7>

<https://medium.com/hugo-ferreiras-blog/confusion-matrix-and-other-metrics-in-machine-learning>

<https://online.stat.psu.edu/stat504/node/216/>

<https://towardsdatascience.com/logit-of-logistic-regression-understanding-the-fundamentals-f384152a33d1>



DATA ANALYTICS

Unit 3: Time Series Analysis

Jyothi R.

Department of Computer Science
and
Engineering

DATA ANALYTICS

Unit 3: Introduction to Time Series Data

Jyothi R., Gowri Srinivasa

Department of Computer Science and Engineering

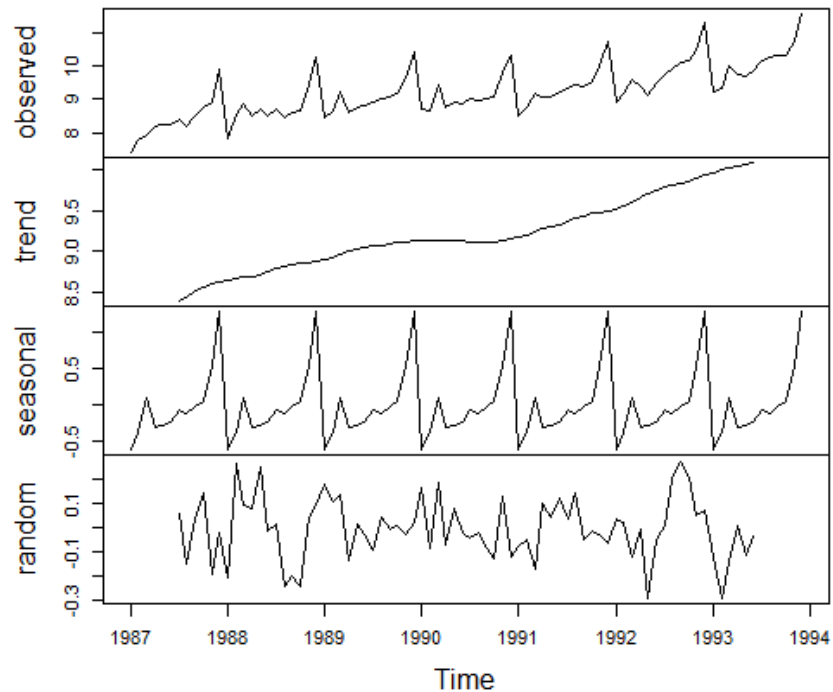
INTRODUCTION TO FORECASTING

- Forecasting - important and frequently addressed problems in analytics
- Inaccurate forecasting has a significant impact
- For example
 - non-availability of product → customer dissatisfaction
 - too much inventory → erodes the organization's profit
- Necessary to forecast the demand for a product and service as accurately as possible.
- Every organization prepares **long-range** and **short-range planning**
 - forecasting demand for product and service is an important input
for both long-range and short-range planning
- Budget allocation, manpower, warehouse capacity, machine resource planning, etc., based on forecast of demand for a product

- Forecasting can be very challenging with [stock keeping units \(SKUs\)](#) running into several millions.
 1. [Boeing 747-400](#) has more than 6 million parts and several thousand unique parts (Hill, 2011). Forecasting [demand for spare parts](#) is important since non-availability of mission critical parts can result in aircraft on ground (AOG) which can be very [expensive for airlines](#).
 2. [Amazon.com](#) sells more than 350 million products through its E-commerce portal. Amazon itself sells about [13 million SKUs](#) and has more (about 2 million) retailers selling their products through Amazon (Ali, 2017).
 3. [Walmart](#) sells more than 142,000 products through their supercenters. Being a brick-and-mortar retail store, Walmart [has to maintain stock](#) for each and every product sold and [predict demand](#) for the products [as accurately as possible](#).

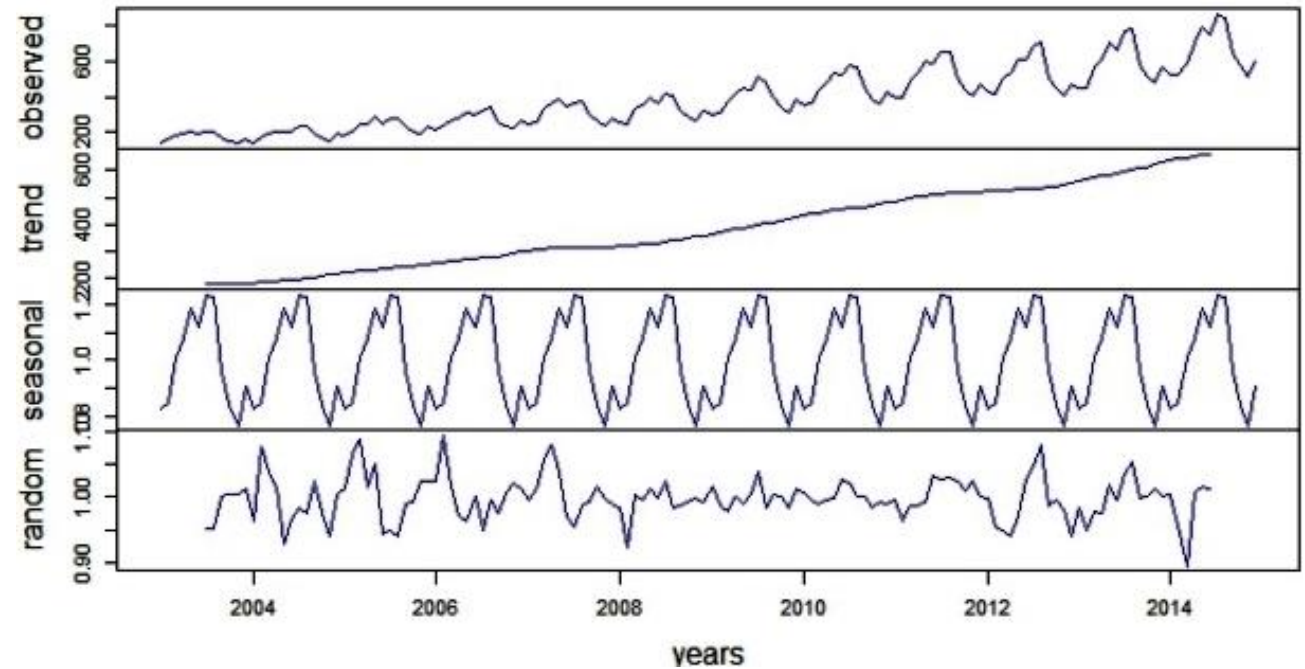
Additive and Multiplicative Time Series Data

Decomposition of additive time series



$$Y_t = T_t + S_t + C_t + I_t$$

Decomposition of multiplicative time series

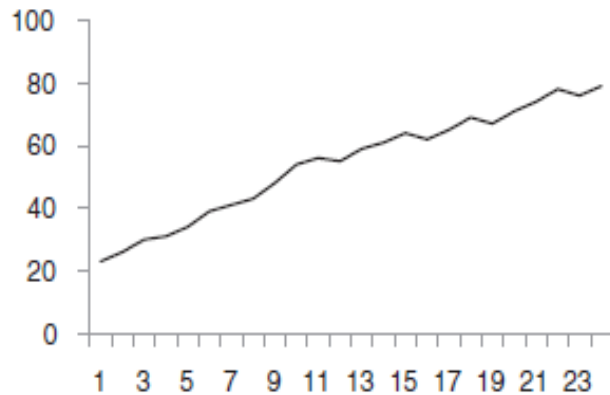


$$Y_t = T_t \times S_t \times C_t \times I_t$$

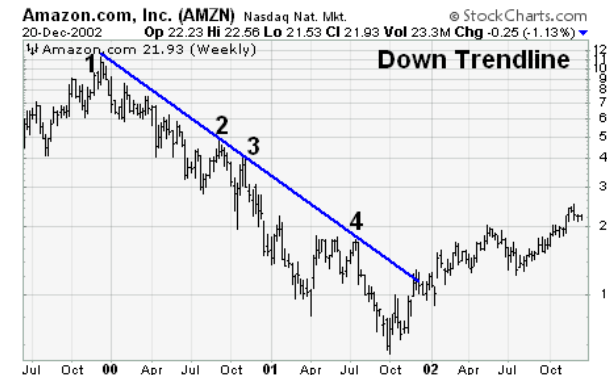
COMPONENTS OF TIME-SERIES DATA

From a forecasting perspective, a time-series data can be broken into the following components

1. **Trend Component (T_t):** Trend is the consistent long-term upward or downward movement of the data over a period of time.



(a) Trend



COMPONENTS OF TIME-SERIES DATA Contd.

2. Seasonal Component (S_t): Seasonal component is the **repetitive** upward or downward movement (or **fluctuations**) **from the trend** that occurs **within a calendar year** such as seasons, quarters, months, days of the week, etc.

- The upward or downward fluctuation may be caused due to festivals, customs within a society, school holidays, business practices within the market such as 'end of season sale', and so on.
- For example, in India demand for many products surge during the festival months of October - December.
- Seasonal fluctuation occurs at fixed intervals (such as months, quarters) known as periodicity of seasonal variation and repeats over time.

Seasonal Component (S_t): Contd.

The seasonal component consists of effects that are reasonably stable with respect to timing, direction and magnitude. It arises from systematic, calendar related influences such as:

- **Natural Conditions:** Weather fluctuations that are representative of the season (uncharacteristic weather patterns such as snow in summer would be considered irregular influences)
- **Business and Administrative procedures:**
 - Start and end of the school term
- **Social and Cultural behavior:**
 - Christmas

Seasonal Component (S_t):

It also includes calendar related systematic effects that are not stable in their annual timing or are caused by variations in the calendar from year to year, such as:

- **Trading Day Effects**

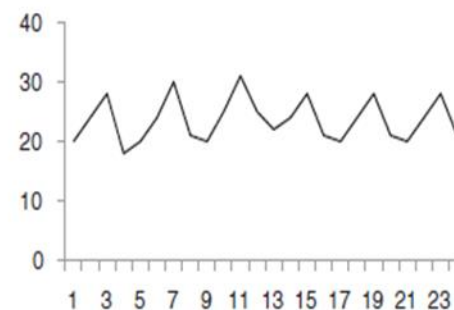
the number of occurrences of each of the day of the week in a given month will differ from year to year

- There were 4 weekends in March in 2000, but 5 weekends in March of 2002

- **Moving Holiday Effects**

holidays which occur each year, but whose exact timing shifts

- Diwali, Easter, Ramadan



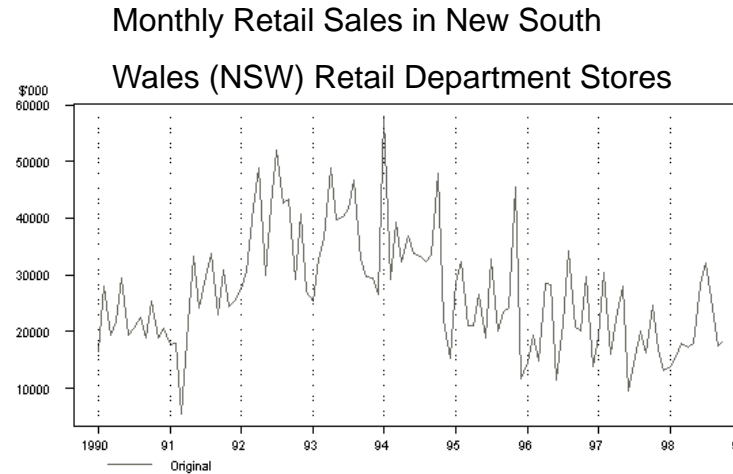
(b) Seasonality (fixed periodicity)

Identifying seasonal components:

Regularly spaced peaks and troughs which have a consistent direction and approximately the same magnitude every year, relative to the trend

Seasonal Component (S_t):

- Seasonality in a time series can be identified by regularly spaced peaks and troughs which have a consistent direction and approximately the same magnitude every year, relative to the trend.
- In this example, the magnitude of the seasonal component increases over time, as does the trend.

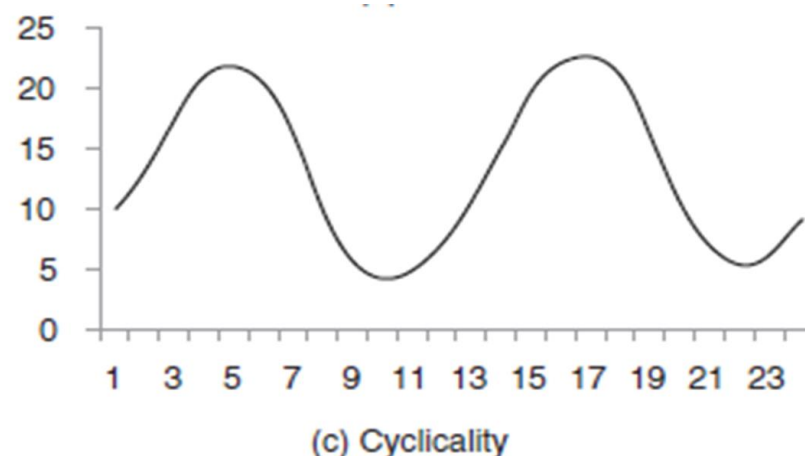


Obvious large seasonal increase in December retail sales in New South Wales due to Christmas shopping

COMPONENTS OF TIME-SERIES DATA

3. Cyclical Component (C_t): Cyclical component is fluctuation around the trend line that happens due to macro-economic changes such as recession, unemployment, etc.

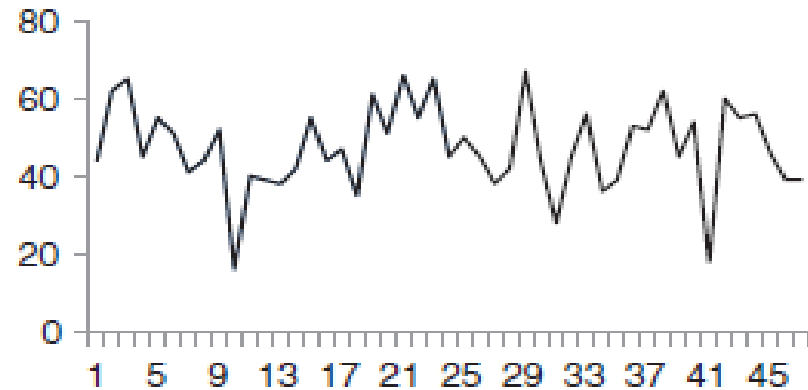
- Cyclical fluctuations have repetitive patterns with a time between repetitions of more than a year



- A major difference between seasonal fluctuation and cyclical fluctuation is that seasonal fluctuation occurs at fixed period within a calendar year, whereas cyclical fluctuations have random time between fluctuations.
- That is, periodicity of seasonal fluctuations is constant, whereas the periodicity of cyclical fluctuations is not constant.

COMPONENTS OF TIME-SERIES DATA contd.

4. Irregular Component (I_t): Irregular component is the white noise or random uncorrelated changes that follow a normal distribution with mean value of 0 and constant variance.



(d) Irregular

- What remains after the seasonal and trend components of a time series have been estimated and removed.
- It results from short term fluctuations in the series which are neither systematic nor predictable

Additive and Multiplicative Time Series Revisited

- The additive time-series model is given by

$$Y_t = T_t + S_t + C_t + I_t$$

- The additive models assume that the **seasonal and cyclical components are independent of the trend** component.
- Additive models are **not very common** since in many cases the seasonal component may not be independent of trend.
- The **additive model** is appropriate if the **seasonal component remains constant about the level** (or mean) and does not vary with the level of the series.

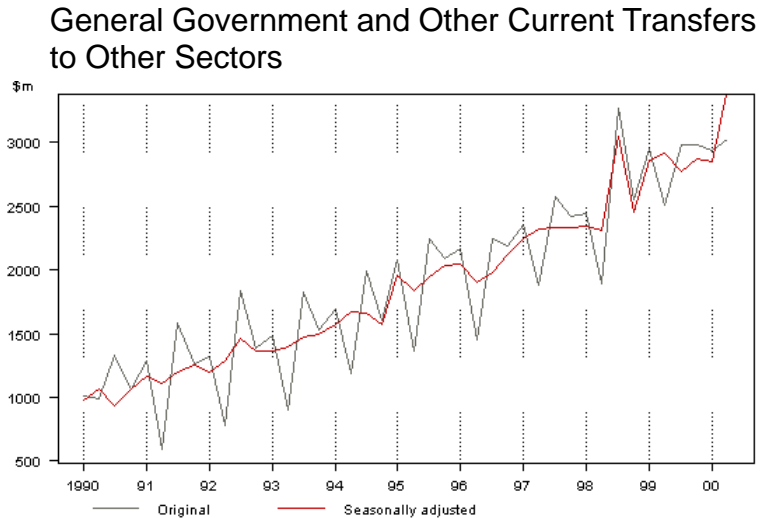
- The multiplicative time-series model is given by

$$Y_t = T_t \times S_t \times C_t \times I_t$$

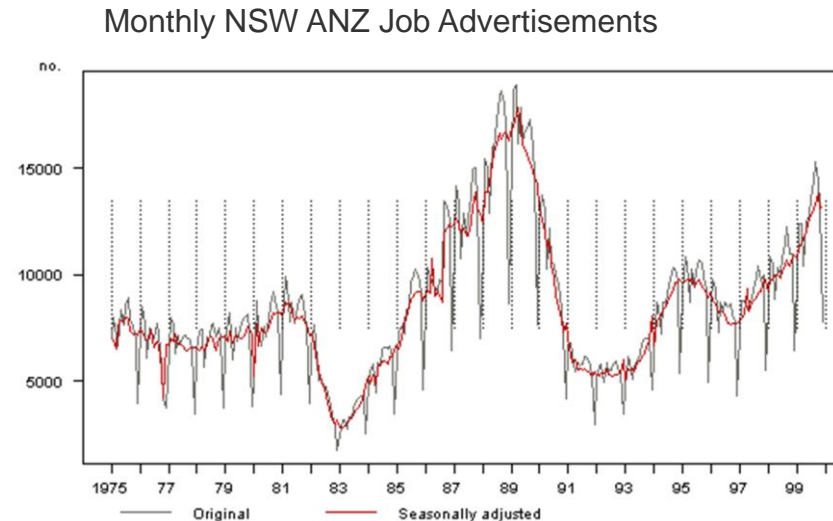
- Multiplicative models are **more common** and are a **better fit for many data sets**.
- In many cases, we will use the form
$$Y_t = T_t \times S_t$$
- To estimate the cyclical component we will need a large data set.
- The **multiplicative model** is more appropriate, if **seasonal variation is correlated with level (local mean)**.

Additive and Multiplicative Time Series Revisited

- The additive time-series model is given by $Y_t = T_t + S_t + C_t + I_t$
- The multiplicative time-series model is given by $Y_t = T_t \times S_t \times C_t \times I_t$



The underlying level of the series fluctuates but the magnitude of the seasonal spikes remain approximately stable



The trend has the same units as the original series, but the seasonal and irregular components are unitless factors, distributed around 1

Decomposition of Time Series Data - Additive

- Decomposition models are typically additive or multiplicative, but can also take other forms such as pseudo-additive.

Additive Decomposition

In some time series, the amplitude of both the seasonal and irregular variations do not change as the level of the trend rises or falls. In such cases, an additive model is appropriate.

In the additive model, the observed time series (O_t) is considered to be the sum of three independent components: the seasonal S_t , the trend T_t and the irregular I_t .

Observed series = Trend + Seasonal + Irregular

$$O_t = T_t + S_t + I_t$$

Seasonally adjusted series = Observed-Seasonal

$$\begin{aligned} SA_t &= O_t - \hat{S}_t \\ &= T_t + I_t \end{aligned}$$

COMPONENTS OF TIME-SERIES DATA contd.

- **Multiplicative Decomposition**

In many time series, the amplitude of both the seasonal and irregular variations increase as the level of the trend rises. In this situation, a multiplicative model is usually appropriate.

In the multiplicative model, the original time series is expressed as the product of trend, seasonal and irregular components.

- Observed series = Trend x Seasonal x Irregular

$$O_t = T_t \times S_t \times I_t$$

Seasonally Adjusted series = Observed \div Seasonal
= Trend \times Irregular

$$\begin{aligned} SA_t &= \frac{O_t}{\hat{S}_t} \\ &= T_t \times I_t \end{aligned}$$

COMPONENTS OF TIME-SERIES DATA contd.

Pseudo-Additive Decomposition:

- The multiplicative model cannot be used when the original time series contains very small or zero values
- This is because it is not possible to divide a number by zero
- In these cases, a pseudo additive model combining the elements of both the additive and multiplicative models is used
- This model assumes that seasonal and irregular variations are both dependent on the level of the trend but independent of each other.

The original data can be expressed in the following form:

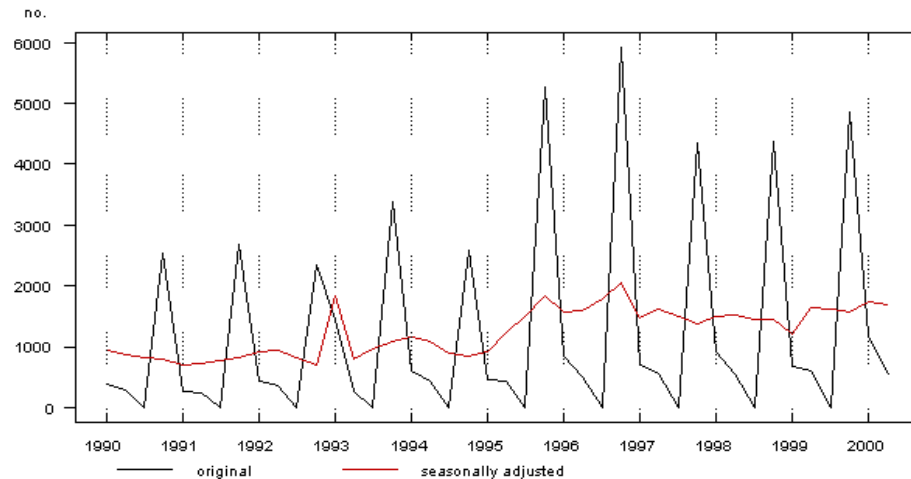
$$\begin{aligned}O_t &= T_t + T_t \times (S_t - 1) + T_t \times (I_t - 1) \\ &= T_t \times (S_t + I_t - 1)\end{aligned}$$

- Both the seasonal factor S_t and the irregular factor I_t centered around one
- We need to subtract one from S_t and I_t to ensure that the terms $T_t \times (S_t - 1)$ and $T_t \times (I_t - 1)$ are centered around zero.
- These terms can be interpreted as the additive seasonal and additive irregular components respectively; the original data O_t will be centered around the trend values T_t .

COMPONENTS OF TIME-SERIES DATA contd.

- An example of series that requires a pseudo-additive decomposition model is shown below.
- This model is used as cereal crops are only produced during certain months, with crop production being virtually zero for one quarter each year.

Quarterly Gross Value for the Production of Cereal Crops



This model is used as cereal crops are only produced during certain months, with crop production being virtually zero for one quarter each year.

Text Book:

“Business Analytics, The Science of Data-Driven Making”, U. Dinesh Kumar, Wiley 2017 (Chapter 13.1-13.2)

Additional reference and image courtesy:

<https://www.abs.gov.au/websitedbs/D3310114.nsf/home/Time+Series+Analysis:+The+Basics>



THANK YOU

Dr. Mamatha H R

Professor, Department of Computer Science

mamathahr@pes.edu

+91 80 2672 1983 Extn 834

Ms. Jyothi R.

Assistant Professor, Department of Computer Science

jyothir@pes.edu