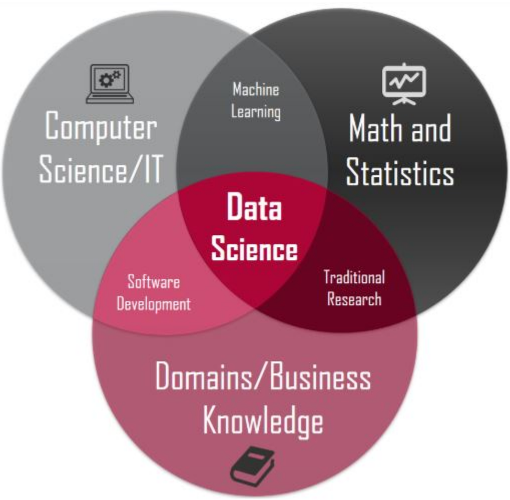
Statistics and Data Science, Enhanced Data Insight: Application in Real Life Problems

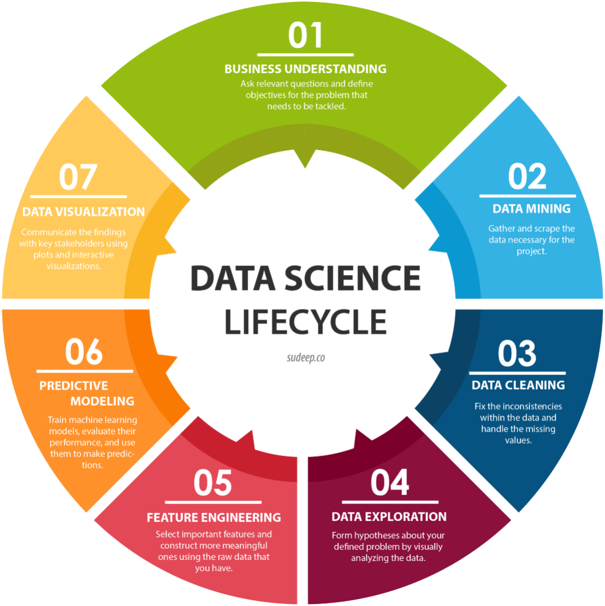
Associate Professor Dr Roslinazairimah Binti Zakaria
Centre for Mathematical Sciences
Universiti Malaysia Pahang

PROGRAM KOLOKIUM STATISTIK DAN
SCIENTIFIC POSTER DOSM 2019
25 -26 September 2019,
Institut Latihan Statistik Malaysia, Sungkai Perak

Data Science and Statistics



Data Science Lifecycle



DATA SCIENCE
LIFECYCLE
sudeep.co
01
BUSINESS UNDERSTANDING
Ask relevant questions and define objectives for the problem that needs to be tackled.
02
DATA MINING
Gather and scrape the data necessary for the project.
03
DATA CLEANING
Fix the inconsistencies within the data and handle the missing values.
04
DATA EXPLORATION
Form hypotheses about your defined problem by visually analyzing the data.
05
FEATURE ENGINEERING
Select important features and construct more meaningful ones using the raw data that you have.
06
PREDICTIVE MODELING
Train machine learning models, evaluate their performance, and use them to make predictions.
07
DATA VISUALIZATION
Communicate the findings with key stakeholders using plots and interactive visualizations.

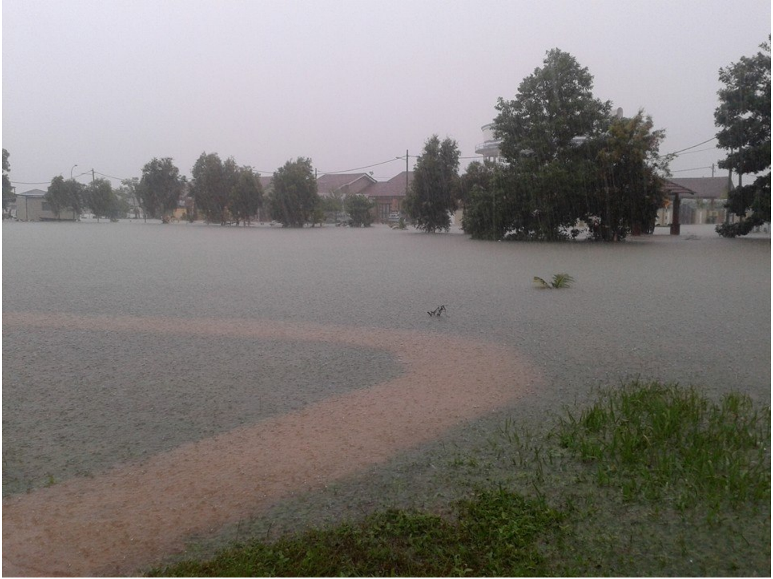
My Research

- Rainfall modelling
- Gold price modelling
- Students data: iCGPA and xCGPA
- Academic Staff Data: ADCAP (Academic Differentiated Career Pathways)

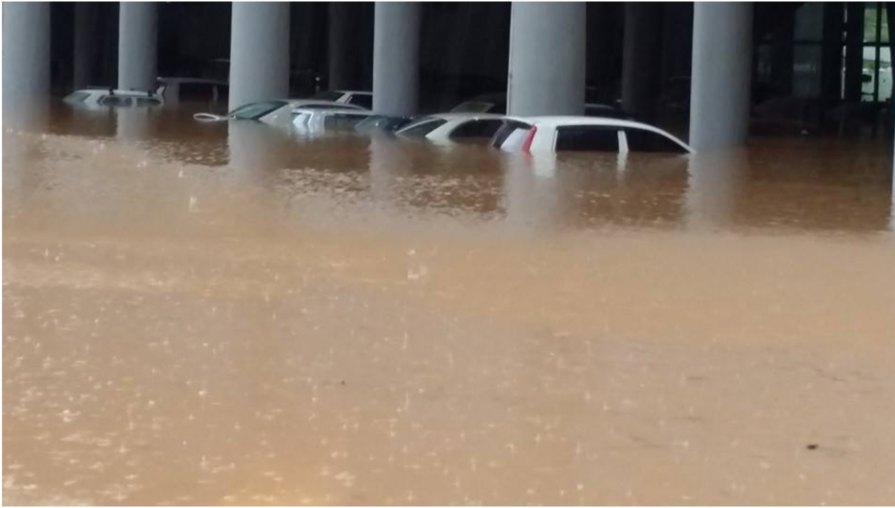
Rainfall Modelling

2 & 3 December 2013, Sg. Isap, Kuantan (407 mm)

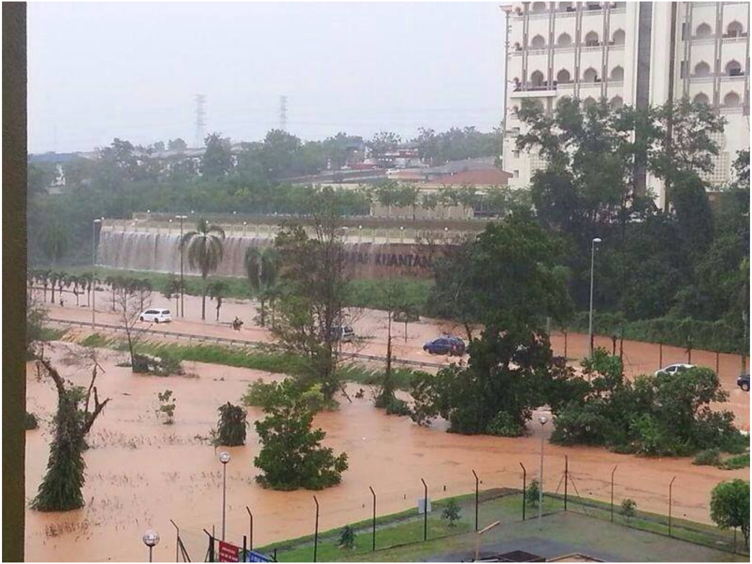
3 Dec 2013, Indera Sempurna, Kuantan



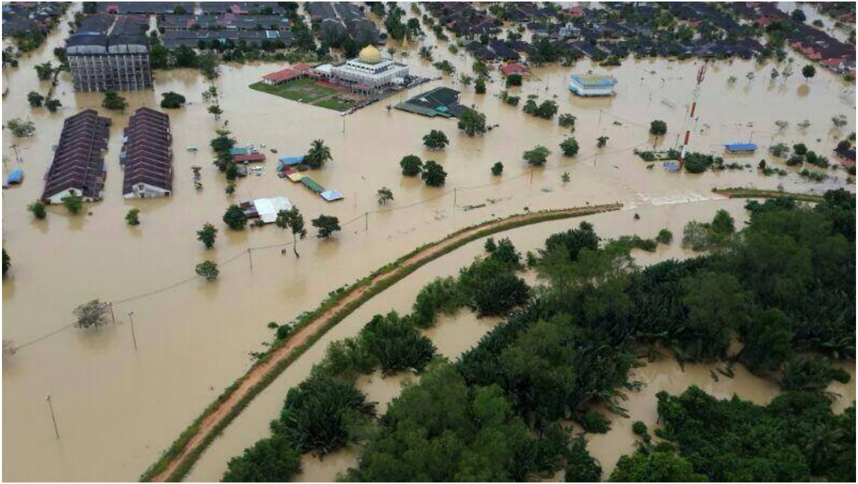
3 Dec 2013, Kuantan Central Terminal



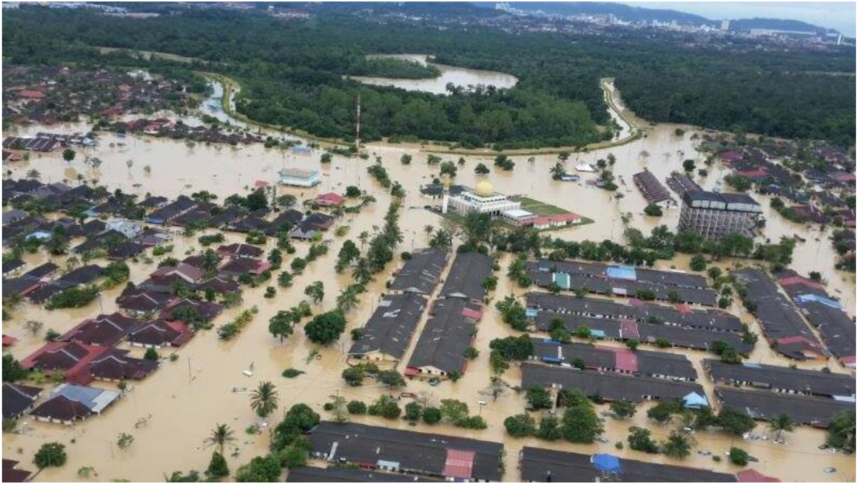
3 Dec 2013, Kuantan High Court



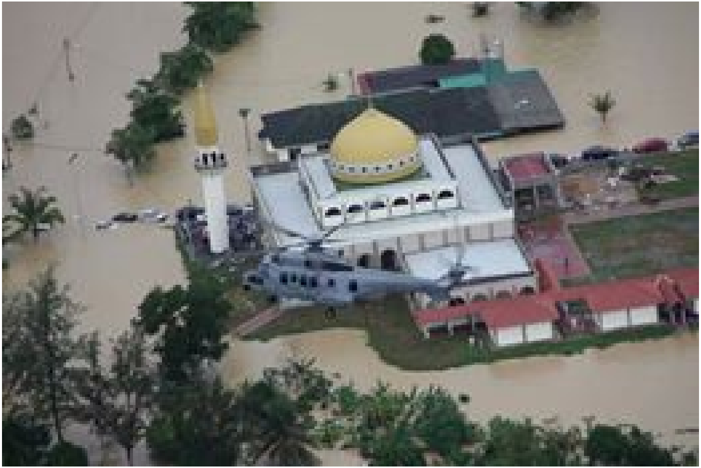
4 Dec 2013, Sungai Isap Kuantan



4 Dec 2013, Sungai Isap Kuantan



4 Dec 2013, Sungai Isap Kuantan



Motivation

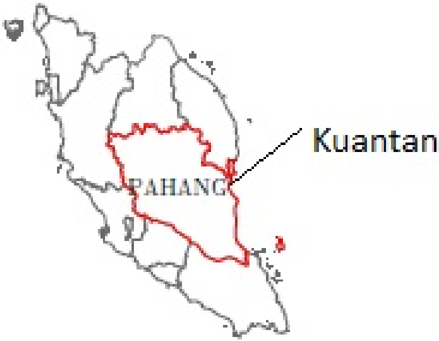
One of the major difficulties in simulating rainfall is the need to accurately represent rainfall accumulations. An accurate simulation of monthly rainfall should also provide an accurate simulation of yearly rainfall by summing the monthly totals.

To model monthly rainfall amounts during the monsoon season for Kuantan station located in the east coast of peninsular Malaysia.

Outline

- Introduction – Kuantan
- Marginal distribution – Gamma distribution
- Model for the sum of gamma variables
- Generating synthetic rainfall data
- Results and discussion

Kuantan Meteorological Station



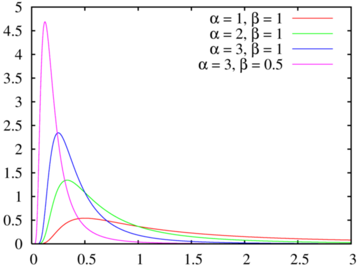
- Dry season (Southwest monsoon, May-August)
- Wet season (Northeast monsoon, October-March)
- State of Pahang (East-coast of Peninsular Malaysia)
- Monthly rainfall data (1971–2000)
- Secondary data obtained from Department of Meteorology Malaysia

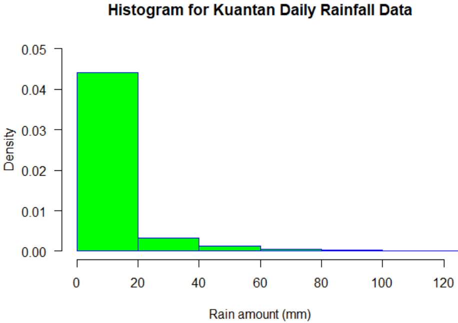
Why use gamma distribution?

The gamma distribution is chosen because it is

- suitable to model continuous variables that are always positive ($x_i > 0$) and have skewed distribution like rainfall totals.
- flexible as it involves two parameters:
  scale (β - spread of data) and shape (α - skewness of data distribution).

Gamma distribution and Kuantan Rainfall





Probability Density Function for Gamma Variables

The gamma distribution is used to model the individual rainfall totals and use maximum likelihood to find the best gamma distribution for each marginal data set.

Consider a set of n independent gamma variables $\{X_i\}_{i=1}^{n}$ with parameters α_i and β_i written as $X_i \sim G(\alpha_i, \beta_i)$ where the PDF of X_i is given by

$$f_i(x_i; \alpha_i, \beta_i) = \frac{x_i^{\alpha_i - 1} e^{-\frac{x_i}{\beta_i}}}{\Gamma(\alpha_i)\beta_i^{\alpha_i}} \;; \qquad x_i > 0 \text{ and } \alpha_i, \beta_i > 0. \qquad (1)$$

Parameters estimation: Maximum Likelihood Estimation

Table: Estimated parameters, means and variances of positive pairs for observed and formulae from gamma distribution

Data	Parameter		Mean (mm)		Variance	
	α	β	observed	$\alpha\beta$	observed	$\alpha\beta^2$
May	4.7890	35.9996	172.40	172.40	5434.57	6206.34
June	4.3822	37.4476	164.10	164.10	6580.34	6145.27

PDF for the sum of independent gamma variables

Type I McKay distribution for the sum of independent gamma variables where $\Sigma = \sum_{i=1}^{n} X_i$ is given by

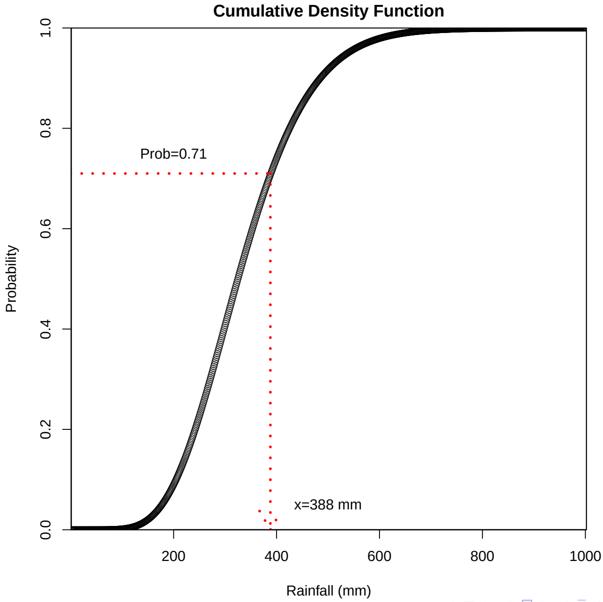
$$f_\Sigma(\sigma) = \frac{\sqrt{\pi}}{\Gamma(a+\frac{1}{2})} \frac{(c^2-1)^{a+\frac{1}{2}}}{2^a b^{a+1}} \, \sigma^a \, \exp\left(-\frac{\sigma c}{b}\right) \, I_a\left(\frac{\sigma}{b}\right) \qquad (2)$$

where a, b, c are real parameters with $a > -1/2$, $b > 0$ and $c > 1$
$\Gamma(\cdot)$ – gamma function
$I_a(\cdot)$ – modified Bessel function of the first kind of order a

Parameters a, b and c are expressed in terms of the parameters α and β for the sum of two independent gamma variables given in type I McKay distribution as

$$a = \alpha - \frac{1}{2}, \quad b = \frac{2\beta_1\beta_2}{|\beta_1 - \beta_2|} \quad \text{and} \quad c = \frac{\beta_1 + \beta_2}{|\beta_1 - \beta_2|}.$$

Generating Synthetic Rainfall Data



Matching the synthetic probability and rainfall amounts

Probability	Rainfall (mm)
0.99	661
0.71	388
0.60	352
0.34	280
0.11	208
⋮	⋮
0.39	295
0.64	365
0.54	334
0.85	450
0.07	191
0.93	511

Goodness of fit test: Kolmogorov–Smirnov

Use to compare the cumulative distributions between the observed data, $F_1(x)$ and generated data, $F_2(x)$.
The cumulative distribution function of a random variable X is defined as $F(x) = P(X \leq x)$. The hypothesis is

$$\begin{aligned} H_0 &: \quad F_1(x) = F_2(x) \\ H_1 &: \quad F_1(x) \neq F_2(x) \end{aligned}$$

and the test statistic is

$$D_{1,2} = \overset{\text{max}}{x} |F_1(x) - F_2(x)| \tag{3}$$

D – the absolute maximum difference between $F_1(x)$ and $F_2(x)$
If P-value $> \alpha = 0.05$, the two distributions are not significantly different from each other.

Results and discussion

Table: Means and variances for the sum of two months of observed and generated rainfall amounts

Data	Mean	Variance
Observed	336.5	12615.8
Generated	333.4	11735.4

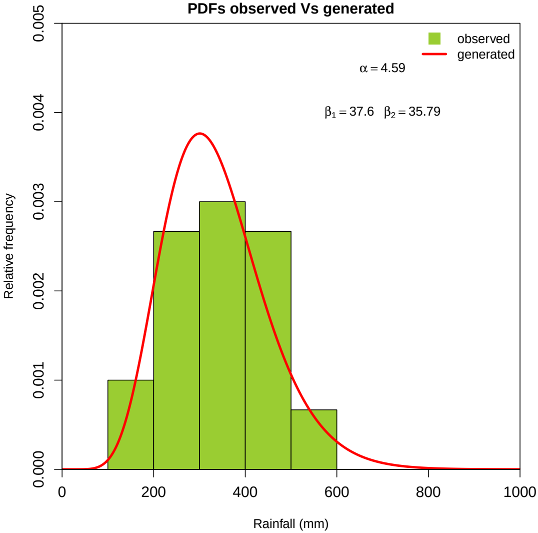
Results and discussion

Table: P-value of goodness of fit test for the sum of McKay distribution of independent gamma variables from two months

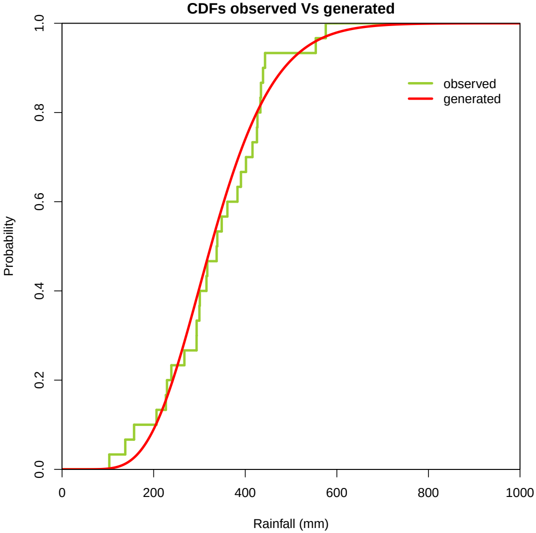
Goodness of fit test	P-value
Kolmogorov–Smirnov	0.8105

Based on the goodness of fit test, the P-value > 0.05 which shows that the observed and the generated data are not significantly different at 5% significance level.

PDF plots of observed and generated for the sum of independent gamma variables



CDF plots of observed and generated for the sum of independent gamma variables



Conclusion

- The McKay model is easy to implement for two independent variables only.

- The model of sum of two independent gamma variables based on McKay distribution is suitable in modelling rainfall amounts for independent random variables.

Extreme Rainfall Modelling

Research by Dr Noor Fadhilah Ahmad Radi
Title: SPATIAL MODELLING OF EXTREME RAINFALL

- Able to develop a spatial rainfall profile for Kelantan to support flood risk management and hydrological design.
- Develop user-friendly and practical software?

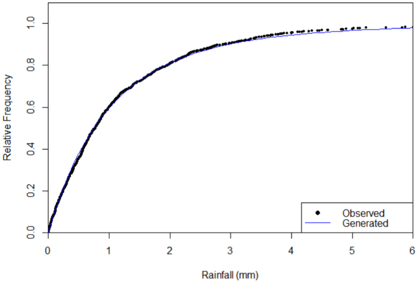
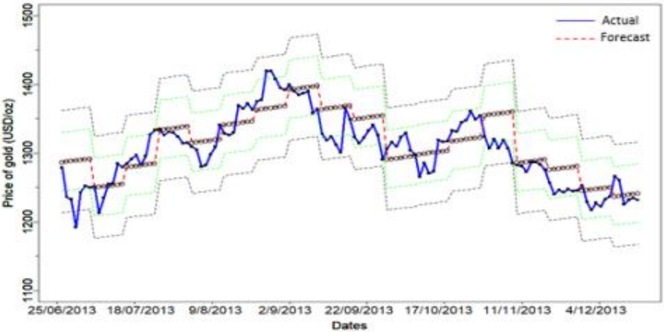


Figure 10: Cdf of GP distribution for observed and generated data for station V30

Time Series Modelling of Gold Price

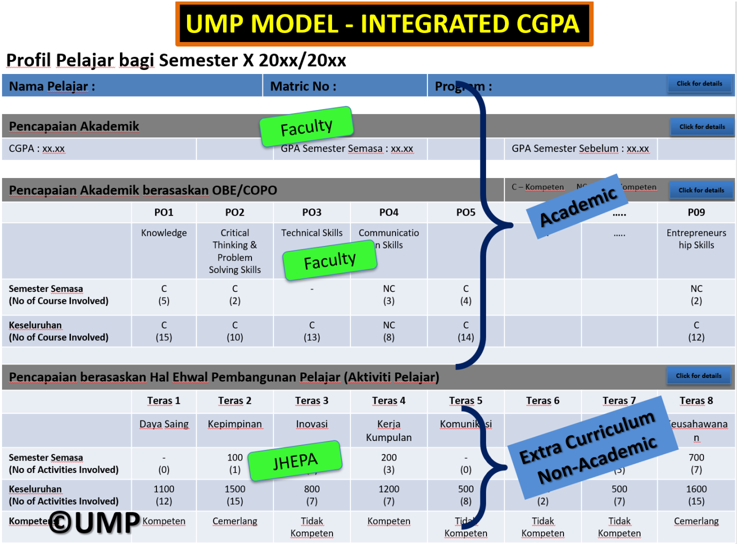
Research by Dr Siti Roslindar Yaziz
Title: NEW ALGORITHMS OF BOX-JENKINS AND GARCH FOR FORECASTING HIGHLY VOLATILE TIME SERIES DATA

- Able to forecast gold price up to 7 days ahead accurately.
- Develop user-friendly Mobile Apps and practical software



iCGPA and xCGPA

iCGPA - integrated Cummulative Grade Point Average
xCGPA - extra-curricular Cummulative Grade Point Average



UMP MODEL - INTEGRATED CGPA

Profil Pelajar bagi Semester X 20xx/20xx

Nama Pelajar :	**Matric No :**	**Program :**	Click for details

Pencapaian Akademik Click for details

CGPA : xx.xx		GPA Semester Semasa : xx.xx		GPA Semester Sebelum : xx.xx	

Pencapaian Akademik berasaskan OBE/COPO C – Kompeten NC [illegible] Kompeten Click for details

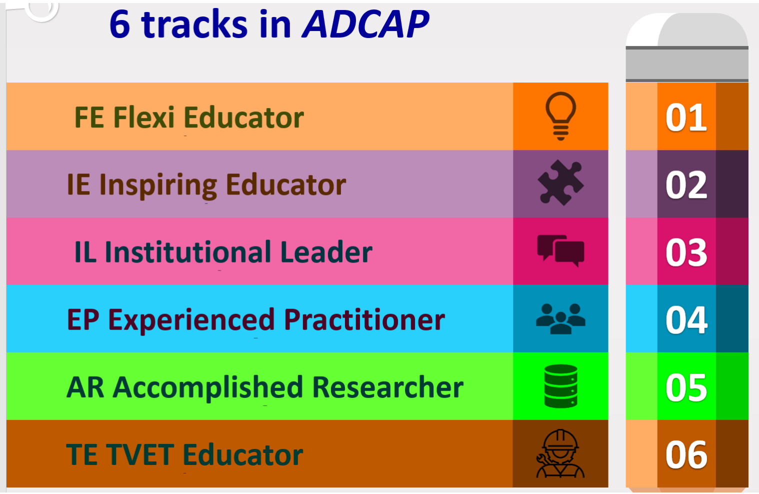
	PO1	PO2	PO3	PO4	PO5	[illegible]	.....	P09
	Knowledge	Critical Thinking & Problem Solving Skills	Technical Skills	Communication Skills		[illegible]	.....	Entrepreneurship Skills
Semester Semasa (No of Course Involved)	C (5)	C (2)	-	NC (3)	C (4)			NC (2)
Keseluruhan (No of Course Involved)	C (15)	C (10)	C (13)	NC (8)	C (14)			C (12)

Pencapaian berasaskan Hal Ehwal Pembangunan Pelajar (Aktiviti Pelajar) Click for details

	Teras 1	Teras 2	Teras 3	Teras 4	Teras 5	Teras 6	Teras 7	Teras 8
	Daya Saing	Kepimpinan	Inovasi	Kerja Kumpulan	Komunikasi	[illegible]	[illegible]	Keusahawanan
Semester Semasa (No of Activities Involved)	- (0)	100 (1)	[illegible]	200 (3)	- (0)	[illegible]	[illegible]	700 (7)
Keseluruhan (No of Activities Involved)	1100 (12)	1500 (15)	800 (7)	1200 (7)	500 (8)	[illegible] (2)	500 (7)	1600 (15)
Kompete	Kompeten	Cemerlang	Tidak Kompeten	Kompeten	Tidak Kompeten	Tidak Kompeten	Tidak Kompeten	Cemerlang

ADCAP

ADCAP - Academic Differentiated Career Pathways



6 tracks in *ADCAP*

- FE Flexi Educator 01
- IE Inspiring Educator 02
- IL Institutional Leader 03
- EP Experienced Practitioner 04
- AR Accomplished Researcher 05
- TE TVET Educator 06

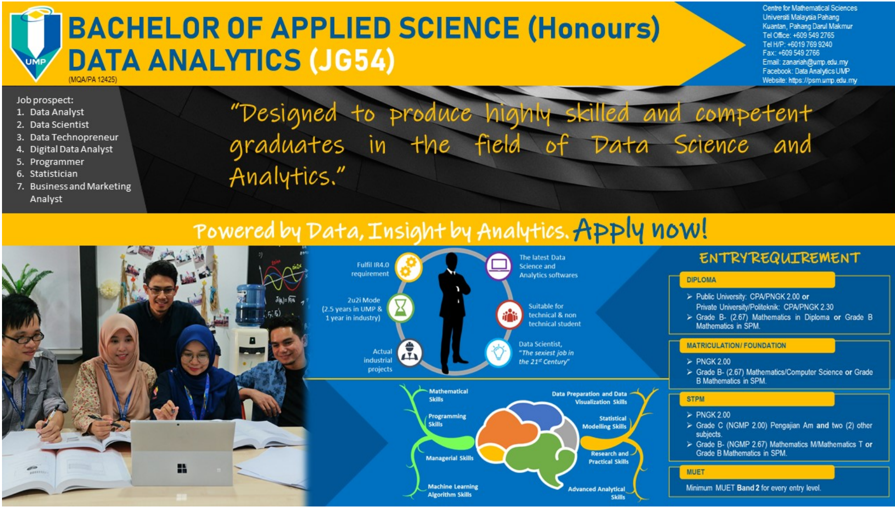
Challenges for Statistics and Data Science

- Data - quality (accuracy, missing values), aquisition (expensive)
- Develop user-friendly and practical softwares, dashboard (interactive), Mobile Apps, Infograhic
- Match customer/industry needs - commercial value

Way Forward Statistics and Data Science

- DOSM and universities can work hand in hand to enhance data analytics.
- Form a research team which compose of team members from different backgrounds including mathematics/statistics, computer science and expert domain.
- Regular meeting and discourse among the team members.
- Joint publication/seminar/conference.

Data Science Bachelor Programme



BACHELOR OF APPLIED SCIENCE (Honours)
DATA ANALYTICS (JG54)

(MQA/PA 12425)

Centre for Mathematical Sciences
Universiti Malaysia Pahang
Kuantan, Pahang Darul Makmur
Tel Office: +609 549 2765
Tel H/P: +6019 769 9240
Fax: +609 549 2766
Email: zanariah@ump.edu.my
Facebook: Data Analytics UMP
Website: https://psm.ump.edu.my

Job prospect:
1. Data Analyst
2. Data Scientist
3. Data Technopreneur
4. Digital Data Analyst
5. Programmer
6. Statistician
7. Business and Marketing Analyst

"Designed to produce highly skilled and competent graduates in the field of Data Science and Analytics."

Powered by Data, Insight by Analytics. Apply now!

UMP
Fulfil IR4.0 requirement
2u2i Mode (2.5 years in UMP & 1 year in industry)
Actual industrial projects
The latest Data Science and Analytics softwares
Suitable for technical & non technical student
Data Scientist, "The sexiest job in the 21st Century"
Mathematical Skills
Programming Skills
Managerial Skills
Machine Learning Algorithm Skills
Data Preparation and Data Visualization Skills
Statistical Modelling Skills
Research and Practical Skills
Advanced Analytical Skills

ENTRYREQUIREMENT

DIPLOMA

- Public University: CPA/PNGK 2.00 **or**
  Private University/Politeknik: CPA/PNGK 2.30
- Grade B- (2.67) Mathematics in Diploma **or** Grade B Mathematics in SPM.

MATRICULATION/ FOUNDATION

- PNGK 2.00
- Grade B- (2.67) Mathematics/Computer Science **or** Grade B Mathematics in SPM.

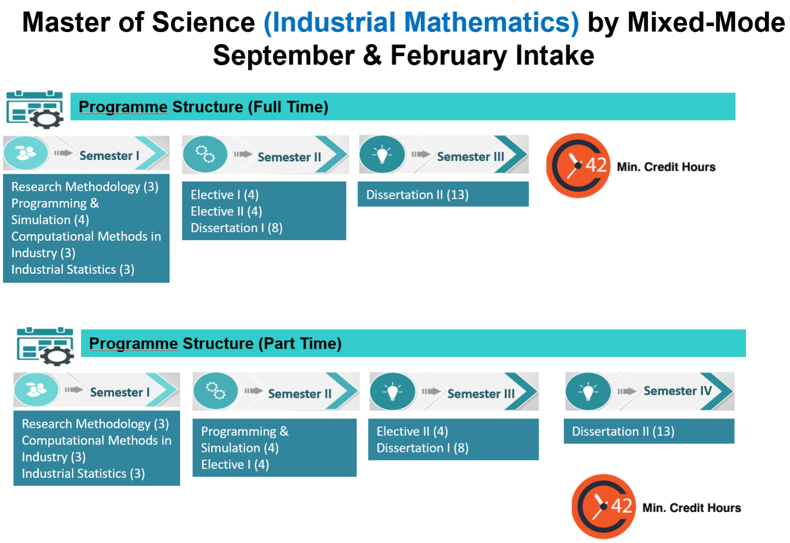
STPM

- PNGK 2.00
- Grade C (NGMP 2.00) Pengajian Am **and** two (2) other subjects.
- Grade B- (NGMP 2.67) Mathematics M/Mathematics T **or** Grade B Mathematics in SPM.

MUET

Minimum MUET **Band 2** for every entry level.

Data Science Master Programme



Master of Science (Industrial Mathematics) by Mixed-Mode September & February Intake

Programme Structure (Full Time)

Semester I

Research Methodology (3)
Programming & Simulation (4)
Computational Methods in Industry (3)
Industrial Statistics (3)

Semester II

Elective I (4)
Elective II (4)
Dissertation I (8)

Semester III

Dissertation II (13)

42 **Min. Credit Hours**

Programme Structure (Part Time)

Semester I

Research Methodology (3)
Computational Methods in Industry (3)
Industrial Statistics (3)

Semester II

Programming & Simulation (4)
Elective I (4)

Semester III

Elective II (4)
Dissertation I (8)

Semester IV

Dissertation II (13)

42 **Min. Credit Hours**

Thank you

roslinazairimah@ump.edu.my