Optimal Combination of Centrality Measures for Node Importance Determination

Ho Khai Shen (U2103646)

UNIVERSITI MALAYA

1. Abstract and Objectives

Introducing centrality measures, a way to rank nodes, plus how and why to construct a convex combination of them. We mainly put our focus on convex combinations of two centrality measures. We introduce some well-known centrality measures and discuss what aspects they individually emphasise to rank nodes of a graph. We will see that different centrality measures provide a variety of valuable insights into node importance, and forming a combination of multiple centrality measures is necessary to capture a graph's whole structure. Objectives To determine t_μ such that

- it maximise the correlation value between c_0 and c_{t_μ}, and between c_1 and c_{t_μ}
- it induces the most number of ranks (regular points)
- c_{t_μ} mimics both c_0 and c_1 better than c_0 and c_1 do to each other ($min\{\tau_b(c_0, c_{t_\mu}), \tau_b(c_1, c_{t_\mu})\} \geq \tau_b(c_0, c_1)$)

Definition 1: Centrality measures

A centrality measure of a graph G is a function $c : V \mapsto [0, \infty)$ with the following properties:

1. *If G is a star with centre u, then $c(u) > c(v)$ for all nodes $v \in V - \{u\}$.*
2. *If $G \cong H$ with bijection $\theta : V(G) \to V(H)$, then $c(v) = c(\theta(v))$ for all nodes $v \in V$. [1]*

2. Well-known Centralities

- Degree centrality
- Closeness centrality
- Harmonic centrality
- Betweenness centrality
- Eigenvector centrality
- Katz centrality
- PageRank centrality

Example: Different centrality measures give different rankings

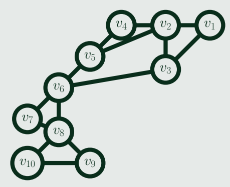


Figure 1. A custom drawn graph

Rank	Degree	Closeness	Harmonic	Betweenness	Eigenvector	Katz	PageRank
1	v_2, v_6, v_8	v_6	v_6	v_6	v_6	v_6	v_8
2	v_3, v_5	v_3, v_5, v_8	v_8	v_8	v_2	v_2	v_2
3	$v_1, v_4, v_7, v_9, v_{10}$	v_7	v_2, v_3, v_5	v_3, v_5	v_3, v_5	v_8	v_6
4	-	v_2	v_7	v_2	v_8	v_3, v_5	v_3, v_5
5	-	v_1, v_4	v_1, v_4	$v_1, v_4, v_7, v_9, v_{10}$	v_7	v_7	v_9, v_{10}
6	-	v_9, v_{10}	v_9, v_{10}	-	v_1, v_4	v_1, v_4	v_7
7	-	-	-	-	v_9, v_{10}	v_9, v_{10}	v_1, v_4

Table 1. Ranks of nodes of Figure 1 by degree, closeness, harmonic, betweenness, eigenvector, Katz and PageRank centrality.

Different centrality measures yield significantly different rankings of nodes, for this reason, a single centrality measure may not be enough to capture the whole network structure and the importance of nodes. Therefore, it is crucial to take into account the view of multiple centrality measures simultaneously, by taking the combination of them when ranking nodes. [2]

3. Convex Combinations of Centrality Measures

Given centrality measures c_0 and c_1 and a real number t, we define the functions tc_0 and $c_0 + c_1$ by setting

$$(tc_0)(v) = tc_0(v) \text{ and } (c_0 + c_1)(v) = c_0(v) + c_1(v)$$

for all $v \in V$.

Definition 2: Convex combinations of centrality measures

Given centrality measures $c_0, c_1, \ldots, c_n$ and non-negative real numbers $t_0, t_1, \ldots, t_n$ such that $\sum_{i=0}^{n} t_i = 1$, we define the convex combination c_t as follows:

$$c_t = t_0c_0 + t_1c_1 + \ldots + t_nc_n = \sum_{i=0}^{n} t_ic_i. \tag{1}$$

Theorem 3:

Let $c_0, c_1, \ldots, c_n$ be centrality measures. For any $(t_0, t_1, \ldots, t_n) \in \Delta^n$, the convex combination c_t is a centrality measure. [1]

Each score of node assigned by individual c_i contributes some weightage to c_t, which is determined by the value t_i. One way to determine the value of each t_i is by making sure c_t truly represents all individual c_i's, that is, the correlation between c_t and each individual c_i's are as large as possible.

4. Correlation Methods

Definition 4: Kendall's tau-b

Given normalized centrality measures c_0 and c_1, the Kendall's tau-b of the two ranks induced by c_0 and c_1 on a graph G on n vertices is

$$\tau_b(c_0, c_1) = \frac{m_C - m_D}{\sqrt{(m - T_0)(m - T_1)}}, \tag{2}$$

where m = #pairs of distinct nodes in $G = \frac{n(n-1)}{2}$, m_C = #concordant pairs in G with respect to c_0 and c_1, m_D = #discordant pairs in G with respect to c_0 and c_1, T_0 = #pairs of distinct nodes in G tied on c_0, T_1 = #pairs of distinct nodes in G tied on c_1.

Definition 5: Pearson's r

Given normalized centrality measures c_0 and c_1, the Pearson's r between the two ranks induced by c_0 and c_1 on a graph G on n vertices is

$$r(c_0, c_1) = \frac{n(\sum_{i=1}^{n} c_0(v_i)c_1(v_i)) - (\sum_{i=1}^{n} c_0(v_i))(\sum_{i=1}^{n} c_1(v_i))}{\sqrt{[n\sum_{i=1}^{n}(c_0(v_i))^2 - (\sum_{i=1}^{n} c_0(v_i))^2][n\sum_{i=1}^{n}(c_1(v_i))^2 - (\sum_{i=1}^{n} c_1(v_i))^2]}}. \tag{3}$$

7. Conclusion

In conclusion, we see different centralities emphasise different aspects. To construct a convex combination, we need it to represent all c_i's, where we used correlation methods. Finally, we demonstrated how to use Kendall's tau-b and Pearson's r to find the optimal t for c_t.

5. Convex Combinations of Two Centrality Measures

A convex combination of two centrality measures, $c_t = (1-t)c_0 + tc_1$, c_0, c_1 spanned by $\Delta^1 = \{(1-t, t) \in | t \in [0, 1]\}$

Lemma 6:

If any pair of distinct nodes of G are concordant with respect to or tied on both c_0 and c_1, then c_t will preserve this property for all $t \in [0, 1]$. [1]

Lemma 7:

Any pair of distinct nodes of G are either tied on c_{t^*} for at most one $t^* \in [0, 1]$ or tied on c_t for all $t \in [0, 1]$. [1]

Definition 8: Regular points

A point $t^* \in [0, 1]$ is a critical point if there exist distinct nodes u and v of G that are ranked equally by c_{t^*} but u and v are either not tied on c_0 nor on c_1. [1]

Definition 9: Critical points

A point $t^* \in [0, 1]$ is a critical point if there exist distinct nodes u and v of G that are ranked equally by c_{t^*} but u and v are either not tied on c_0 nor on c_1. [1]

Lemma 10:

For any pair of distinct nodes of G that is neither concordant wrt nor tied on both c_0 and c_1, there exists a unique critical point $t^* \in [0, 1]$. Furthermore, for any $r, s \in [0, 1]$, the pair of nodes are concordant with respect to c_r and c_s if and only if either $r, s \in [0, t^*)$ or $r, s \in (t^*, 1]$. [1]

Lemma 11:

The ranking of nodes of a graph G induced by c_t, as t varies in $[0, 1]$, changes only finitely many times and the number of times the ranking changes is

$$2|T^*| - |T^* \cap \{0, 1\}|, \tag{4}$$

where T^* is the set of critical points $t^* \in [0, 1]$ and $|T^*|$ is the cardinality of the set T^*. [1]

Lemma 12:

For a regular point t, the ranking of nodes of a graph G induced by c_t is insensitive to sufficiently small variation in t. Particularly, for any two consecutive critical points t_i^* and t_j^*, where $t_i^* < t_j^*$, the ranking induced by c_t stays the same as t varies in $t \in (t_i^*, t_j^*) \cap [0, 1]$. [1]

6. Optimizing Convex Combinations of Two Centrality Measures

Theorem 13:

Using Kendall's tau-b, there is always a $t_b \in [0, 1]$ such that t_b is a regular point and c_{t_b} mimics both c_0 and c_1 better than c_0 and c_1 do to each other. [1]

Via minimizing the objective function $g(t; c_0, c_1) = (1 - \tau_b(c_0, c_t))^2 + (1 - \tau_b(c_1, c_t))^2$

- By Lemma 11, each $\tau_b(c_0, c_t)$ and $\tau_b(c_1, c_t)$ changes its values finitely many times as t varies in $[0, 1]$. By Lemma 6 and 10, changes in values occur only when it passes through a critical point.
- Therefore together with the Theorem 13, it is feasible to determine t_b and hence compute the optimum c_{t_b} by minimizing the objective function $g(t; c_0, c_1)$ with regular points.

Theorem 14:

Using Pearson's r, there is always a $t_r \in [0, 1]$ such that c_{t_r} mimics both c_0 and c_1 better than c_0 and c_1 do to each other.

Via minimizing the objective function $h(t; c_0, c_1) = (1 - r(c_0, c_t))^2 + (1 - r(c_1, c_t))^2$

- With similar sayings as before, by Lemma 6, 10, 11 and Theorem 14, it is feasible to determine t_r and hence compute the optimum c_{t_r} by minimizing the objective function $h(t; c_0, c_1)$.

Example: c_{t_b} and c_{t_r} splits ranks of nodes better

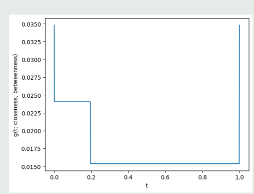


Figure 2. $g(t; c_0, c_1)$ for Figure 1

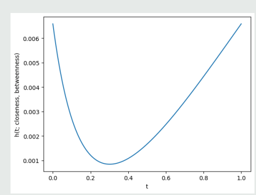


Figure 3. $h(t; c_0, c_1)$ for Figure 1

Rank	c_0 (Closeness)	c_1 (Betweenness)	c_{t_r}	c_{t_b}
1	v_6	v_6	v_6	v_6
2	v_3, v_5, v_8	v_8	v_8	v_8
3	v_7	v_3, v_5	v_3, v_5	v_3, v_5
4	v_2	v_2	v_2	v_2
5	v_1, v_4	$v_1, v_4, v_7, v_9, v_{10}$	v_7	v_7
6	v_9, v_{10}	-	v_1, v_4	v_1, v_4
7	-	-	v_9, v_{10}	v_9, v_{10}

Table 2. Ranking of nodes by $c_0, c_1, c_{t_b}, c_{t_r}$ on Figure 1

When c_0, c_1 = closeness, betweenness centrality,

- $g(t; c_0, c_1)$ minimized at $t_b \approx 0.1959$, $h(t; c_0, c_1)$ minimized at $t_r \approx 0.3013$
- $c_{t_b} = 0.8041c_0 + 0.1959c_1$, $c_{t_r} = 0.6987c_0 + 0.3013c_1$
- $\min\{\tau_b(c_0, c_{t_b}) \approx 0.9271, \tau_b(c_1, c_{t_b}) \approx 0.8997\} \geq \tau_b(c_0, c_1) \approx 0.8315$, $\min\{r(c_0, c_{t_b}) \approx 0.9795, r(c_1, c_{t_b}) \approx 0.9795\} \geq r(c_0, c_1) \approx 0.9188$

8. References

[1] Y. Y. Keng, K. H. Kwa, and C. McClain. "Convex combinations of centrality measures". In: *The Journal of Mathematical Sociology* 45.4 (2021), pp. 195–222. DOI: 10.1080/0022250X.2020.1765776. eprint: https://doi.org/10.1080/0022250X.2020.1765776. URL: https://doi.org/10.1080/0022250X.2020.1765776.

[2] A. Abbasi and L. Hossain. "Hybrid Centrality Measures for Binary and Weighted Networks". In: *Complex Networks*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 1–7. ISBN: 978-3-642-30287-9. DOI: 10.1007/978-3-642-30287-9_1. URL: https://doi.org/10.1007/978-3-642-30287-9_1.

u2103646@siswa.um.edu.my

UNSTEADY MAGNETOHYDRODYNAMIC (MHD) FLUID FLOW OVER A STRETCHING SHEET WITH VISCOUS DISSIPATION AND HEAT SOURCE EFFECTS

UNIVERSITI MALAYA
SIN3015 MATHEMATICAL SCIENCE PROJECT
MOHAMAD ADAM DANIAL BIN SHAHRIMAN (U2004713)
SUPERVISOR: DR KOHILAVANI NAGANTHRAN

Abstract

This study analyzes the unsteady magnetohydrodynamic (MHD) boundary layer flow and heat transfer of a fluid over a stretching sheet in the presence of viscous dissipation and heat source. By using the similarity transformation, the governing non-linear partial differential equations are transformed into a solvable form, which is ordinary differential equations, before being solved numerically by a collocation method using the bvp4c function in MATLAB software. The influence of physical parameters such as the local Nusselt number and skin friction coefficient on the developed model is presented and discussed. The impacts of various physical parameters on the dimensionless temperature profiles are graphically illustrated and analyzed. The graph should be asymptotic to the x-axis as the similarity variable increases. From the result, we could see that as the magnetic parameter *M* increases, the skin friction coefficient, increases while the Nusselt number decreases.

Research Objectives

- To construct a mathematical model which describes the unsteady two-dimensional laminar boundary layer flow past a continuously stretching sheet immersed in an incompressible electrically conducting field.
- To solve the mathematical model by using the collocation method (bvp4c function) available in MATLAB software.
- To analyze the transport phenomena in the fluid regime when the governing parameter varies in terms of the physical quantities.

Literature Review

1. Boundary layer flow over a stretching sheet
   - Sakiadis (1961) analyzed boundary layer flow on a continuously stretching surface at a constant speed.
   - Chen and Char (1988) analyzed laminar boundary layer flow and heat transfer on a stretching sheet with various thermal conditions.

2. Magnetohydrodynamics (MHD)
   - In 1940s, geophysicists theorized that Earth's magnetic field was generated by the dynamo action in its liquid-metal core, an idea first proposed in 1919 by Larmor for the Sun.
   - In the 1950s, plasma physicists became interested in MHD for controlled thermonuclear fusion, focusing on the stability of magnetically confined plasmas.

3. Unsteady Flow
   - Stokes has laid the foundation for the Navier-Stokes equations, which describe the motion of viscous fluids and are fundamental to the analysis of unsteady flows.
   - Ludwig Prandtl developed the concept of the boundary layer and made significant advances in the understanding of unsteady flow phenomena, particularly in the context of aerodynamics.
   - Unsteady flow is characterized by flow parameters at any point that vary with time.

4. Viscous Dissipation
   - Morini (2013) stated that the irreversible process by means of which the work done by a fluid on adjacent layers due to the action of shear forces is transformed into heat is defined as viscous dissipation.
   - Stokes (1851) studied the motion of spheres through a viscous fluid, leading to the formulation of Stokes' law, which describes the drag force experienced by a sphere moving through a viscous fluid.

5. Heat Source/Sink
   - In the context of heat transfer, a heat source adds thermal energy to a system, increasing its temperature, while a heat sink removes thermal energy, decreasing the temperature.
   - According to West (2014), Joseph Black was a pioneer in careful measurements of heat transfer in 1761 who noticed that the change of state can occur over a prolonged period when a substance is heated or cooled without a change in temperature.

Methodology

This section explains how viscous dissipation and heat sources affect unsteady MHD flow over a stretching sheet. It derives the governing equations which are continuity equation, momentum equation and energy equation and uses similarity transformation to turn them into ordinary differential equations. The chapter also discusses using the bvp4c function in MATLAB to compute numerical solutions, which helps understand how the governing parameters change.

Mathematical Model

$$f''' + ff'' - f'^2 - Mf' - A\left(f' + \frac{\eta}{2} f''\right) = 0$$

$$(1+R)\theta'' + \Pr\left(f\theta' - f'\theta\right) - \Pr A\left(\theta + \frac{\eta}{2}\theta'\right) + \Pr\left(Ec f''^2 + \gamma\theta\right) = 0$$

with the boundary conditions,

$$f(0) = 0,\ f'(0) = 1,\ \theta(0) = 1 \text{ at } \eta = 0,$$

$$f'(\eta) \to 0,\ \theta(\eta) \to 0 \text{ as } \eta \to \infty$$

Schematic Diagram

Figure 1.2: Velocity boundary layer development on a flat plate

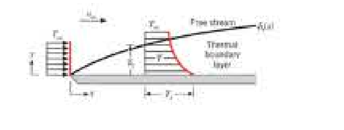


Figure 1.3: Thermal boundary layer development on an isothermal flat plate

Results

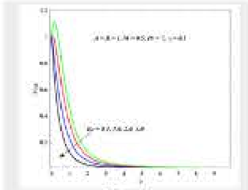


Figure 1:
Temperature profiles for different values of *Ec*.

Ec	$-f''(0)$	$-\theta'(0)$
0.5	2.8975998	3.1312837
1.0	2.8975998	1.8494563
1.5	2.8975998	0.5676289
2.0	2.8975998	-0.7141985
2.5	2.8975998	-1.9960259

- An increment in *Ec* infers that the viscous dissipation in the boundary layer becomes more significant. Viscous dissipation refers to the conversion of mechanical energy into thermal energy due to the internal friction within the fluid. Thus, the fluid temperature increases as *Ec* increases from 0.1 to 3.
- From the table, we can observe that when parameters are fixed at $M = 2$, $A = 7$, $R = 1$, $\gamma = 1$, as *Ec* increase, the local Nusselt number decreases as stated by Reddy et al. (2015).

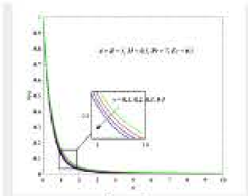


Figure 2:
Temperature profiles for different values of γ.

γ	$-f''(0)$	$-\theta'(0)$
-2	2.8975998	3.1206143
-1	2.8975998	2.7287770
1	2.8975998	1.8494563
2	2.8975998	1.3456238
3	2.8975998	0.7817206

- The heat source parameter represents the strength of the internal heat generation or absorption within the boundary layer. As the heat source parameter increases, the internal heat generation within the boundary layer also increases. Thus, the fluid temperature increases as γ increases from 0.1 to 0.4.
- It is observed that as the heat source parameter γ increases, the local Nusselt number decreases as proved by Reddy et al. (2015).
- The decreased temperature gradient between the stretching sheet and the fluid, caused by the increased fluid temperature, leads to a reduction in the convective heat transfer from the stretching sheet to the fluid.

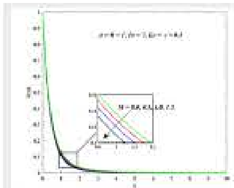


Figure 3:
Temperature profiles for different values of *M*.

M	$-f''(0)$	$-\theta'(0)$
1	2.7228832	2.0995194
2	2.8975998	1.8494563
3	3.0629214	1.6080302
4	3.2201620	1.3745288
5	3.3703640	1.1483043

- The magnetic parameter represents the strength of the applied magnetic field perpendicular to the flow direction. As the magnetic parameter increases, the Lorentz force acting on the electrically conductive fluid within the boundary layer also increases. The decreased fluid velocity within the boundary layer due to the increased Lorentz force leads to a reduction in the convective heat transfer from the stretching sheet to the fluid. Thus, when *M* increases the fluid temperature increase.
- We could see that as the magnetic parameter *M* increases, the skin friction coefficient increases while the local Nusselt number decreases.
- The Lorentz force introduced by the increased magnetic parameter acts to oppose the fluid motion within the boundary layer. This affect the wall shear stress to rise and results in an increase of $-f''(0)$.
- The decreased fluid velocity within the boundary layer due to the increased Lorentz force leads to a reduction in the convective heat transfer from the stretching sheet to the fluid. This affect the fluid temperature to rise and results in the decrease of $-\theta'(0)$.

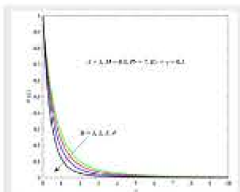


Figure 4:
Temperature profiles for different values of *R*.

R	$-f''(0)$	$-\theta'(0)$
1	2.8975998	1.8494563
2	2.8975998	1.7204523
3	2.8975998	1.6148481
4	2.8975998	1.5279230
5	2.8975998	1.4549626

- The radiation parameter represents the contribution of thermal radiation to the overall heat transfer in the boundary layer. The fluid within the boundary layer absorbs the incoming radiative heat, leading to an increase in the internal energy of the fluid. This additional heat absorbed by the fluid through radiative heat transfer causes the fluid temperature to rise. Thus, the fluid temperature increases as *R* increases from 1 to 4.
- We could see that as the radiation parameter *R* increases, the local Nusselt number decreases, as stated by Reddy et al. (2015).

Conclusion

- Temperature profiles decrease with an increase in the magnetic parameter, *M*.
- Temperature distribution increases as the radiation parameter, Eckert number, and the heat source parameter increase.
- The magnitude of the skin friction coefficient increases with an increase in magnetic parameter whereas the magnitude of the Nusselt number decreases.
- The heat transfer rate at the surface decreases with an increase in the radiation parameter, magnetic parameter, Eckert number and heat source parameter.
- The contribution of the present work has rectified the mistake which has been made by Reddy et al. (2015).

Acknowledgement

I would like to express my deepest gratitude to my supervisor, Dr. Kohilavani A/P Naganthran, for her unwavering support, guidance, and encouragement throughout my final year project (FYP). Her expertise and insights have been invaluable in shaping my research and academic growth. I am also immensely thankful to my family and friends for their constant love, understanding, and encouragement. Their unwavering support kept me motivated and focused, enabling me to successfully complete my FYP.

ASSESSMENTS OF MACHINE LEARNING REGRESSION MODELS FOR PREDICTING THE DEVELOPMENT OF THE MIDFACE IN SYNDROMIC CRANIOSYNOSTOSIS

HaoRan Xin S2106630

UNIVERSITY MALAYA

Introduction

Syndromic Craniosynostosis is a congenital condition in which the sutures of the skull fuse prematurely, resulting in abnormal head and face shape. In order to develop a targeted personalized treatment plan, the prediction of facial growth patterns is particularly important. By examining the accuracy of a specific regression model for the face measurement, it provides the basis for personalized and precise treatment programs.This study aimed to determine the effectiveness of three regression models-least square linear regression, support vector machine (SVM) regression, and random forest-in predicting the middle facial area of Syndromic Craniosynostosis. Identify the most reliable models for high-precision prediction in clinical applications.

Analysis Methodology

Least Squares Linear Regression

Least squares linear regression minimizes the sum of squared errors to fit a line through the data, simplifying the analysis when the variables have an almost linear relationship.

Support Vector Machine Regression

SVM regression uses kernel functions to map input features to higher dimensional Spaces, thus achieving linear separation and suitable for handling complex interactions

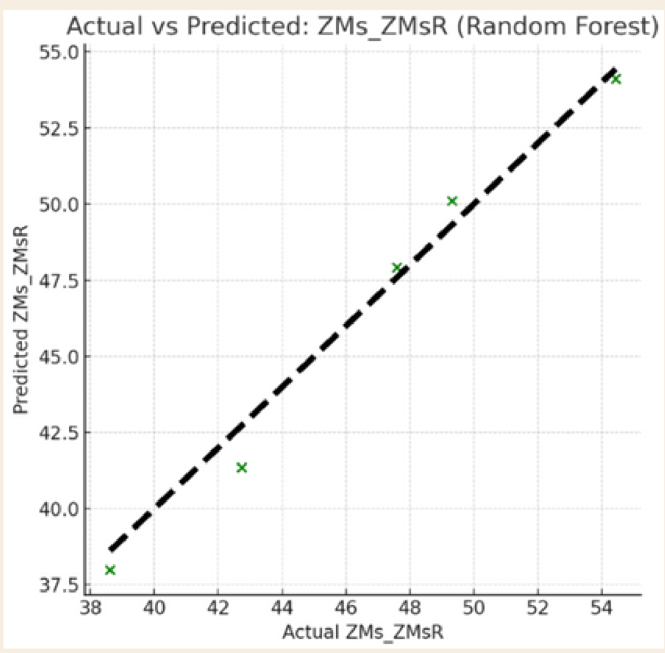
Random Forest

The random forest builds multiple decision trees and averages their predictions to reduce errors and improve confidence. Its efficient handling of nonlinear relationships and complex interactions makes it suitable for predicting high-precision midplane measurements.

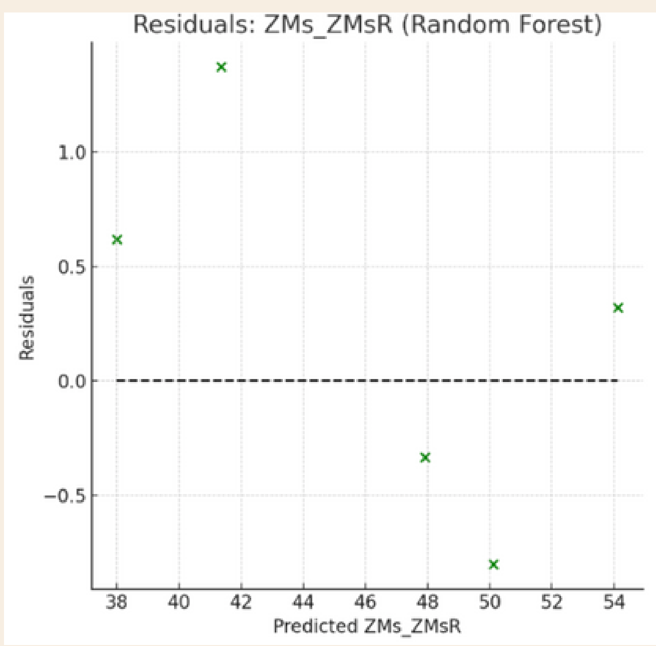
Random Forest Results

Variable	Mean Squared Error (MSE)	R-squared (R^2)
ZMs-ZMsR	6.698	0.877
N-ANS	6.042	0.992
ANS-PNS	10.123	0.836
ZF-ZFR	18.989	0.904
ZMs-ZTi	13.462	0.934

The outcome of Random Forest Regression indicates the sufficiently good seemingly in each dependent variable, with the range of R-squared (R^2) values fluctuation from 0.836 to 0.992.
The N-ANS factor is depicted as bearing a value of 0.992 at R^2
, which could mean that the
model explains 0.992. And the N-ANS variable MSE also comes out low in correlation



Random Forest Actual vs Predicted Values for ZMs_ZMsR



Random Forest Residuals for ZMs_ZMsR

In the least square linear regression model, ZMs-ZMsR is chosen as the best predictor. In order to compare the performance of different models on the same variable, we also use the ZMs-ZMsR variable for analysis in the random forest model.Although not as good as N-ANS, it still has higher explanatory power and lower prediction error.The random forest model shows high prediction accuracy when predicting the variable "ZMs_ZMsR".Most of the data points are close to the ideal 1:1 relationship line, indicating that the random forest model can fit the data well and provide accurate predictions.

Most of the residials are between -0.5 and 1.5, showing that the random forest model's predictions of the "ZMs_ZMsR" variable are very accurate with very little error.
While there are some positive and negative residuals near the predicted values of 40, 48, and 54, the values are very small, showing that the model's prediction error in these intervals is also small.The random forest model performs well in predicting the "ZMs_ZMsR" variable, with small residual values, relatively random error distribution, and no obvious systematic bias.

Least Squares Linear Regression Results

Variable	Mean Squared Error (MSE)	R-squared (R^2)
ZMs-ZMsR	6.484	0.906
N-ANS	12.840	0.888
ANS-PNS	8.815	0.860
ZF-ZFR	20.022	0.870
ZMs-ZTi	13.115	0.835

ZMs-ZMsR and ANS-PNS are the most suitable predictors because they have high R^2 values and relatively low MSE values, indicating that the model has strong explanatory power for these variables and small prediction errors.

ZMs-ZMsR Function:

$$\text{ZMs-ZMsR}=0.42(\text{S-N})-1.62(\text{N-SO})-0.46(\text{SO-BA})-0.15(\text{N-S-BA})+0.05(\text{S-SO-BA})-21.38$$

ANS-PNS Function:

$$\text{ANS-PNS}=0.12(\text{S-N})+1.74(\text{N-SO})+0.91(\text{SO-BA})+1.15(\text{N-S-BA})+1.24(\text{S-SO-BA})+45.01$$

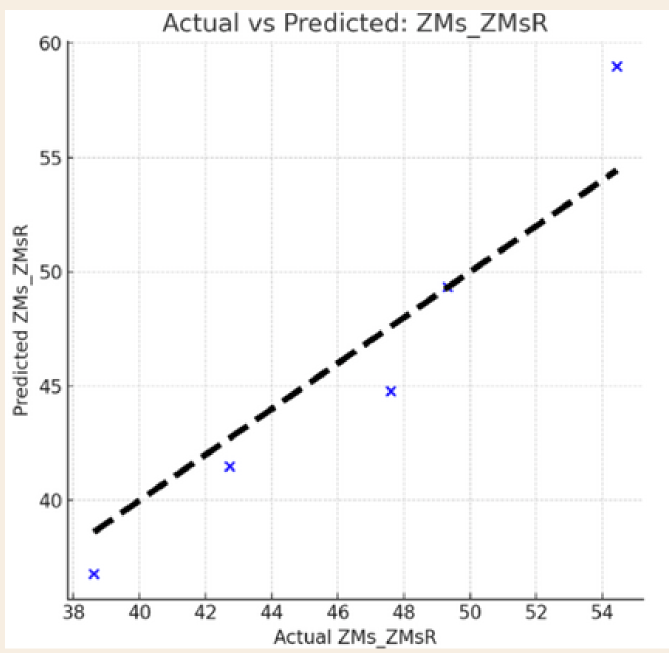
ZF-ZFR Function:

$$\text{ZF-ZFR}=0.15(\text{S-N})-0.42(\text{N-SO})-0.22(\text{SO-BA})+0.19(\text{N-S-BA})+0.24(\text{S-SO-BA})-0.54$$

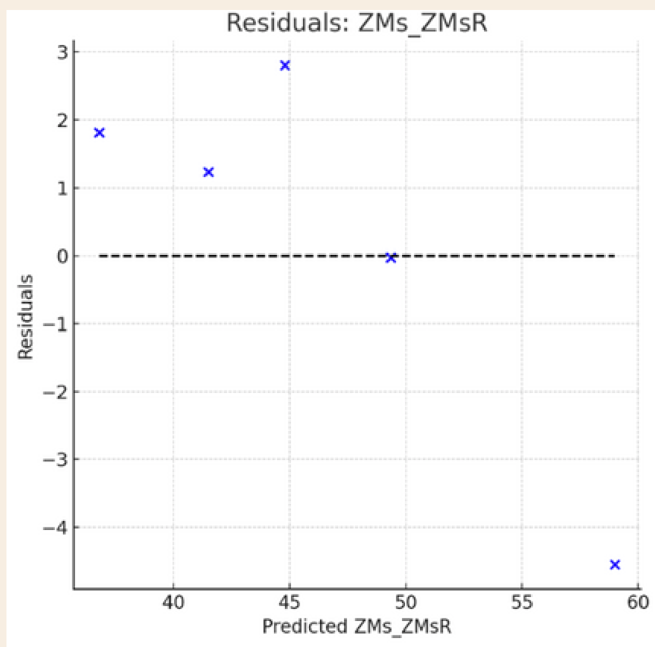
ZMs-ZTi Function:

$$\text{ZMs-ZTi}=-0.05(\text{S-N})-0.13(\text{N-SO})-0.04(\text{SO-BA})-0.06(\text{N-S-BA})+0.01(\text{S-SO-BA})-48.03$$

These formulas are derived by linear regression model, they describe the linear influence of different independent variables on the target variable, and can be used to predict the corresponding target variable value.



LEAST SQUARES LINEAR REGRESSION ACTUAL VS PREDICTED VALUES FOR ZMS_ZMSR



LEAST SQUARES LINEAR REGRESSION RESIDUALS FOR ZMS_ZMSR

At the point where the X-axis is close to 40, the predicted value is significantly lower than the actual value, and the positive residual is larger in the corresponding residual plot.
For points where the X-axis is close to 60, the predicted value is significantly higher than the actual value, and the negative residual is larger in the corresponding residual plot.The residual error in the middle region is small,indicating that the model is more accurate in this interval.

The residual distribution is relatively random, and there is no obvious systematic bias, indicating that the error of the model is random in most of the predicted values.
Large residuals appear at the extremes of the predicted values (near 45 and 60), suggesting that the model has a large error problem in these intervals.

Support Vector Machine (SVM)regression Result

Variable	Mean Squared Error (MSE)	R-squared (R^2)
ZMs-ZMsR	56.512	0.178
N-ANS	92.9762	0.191
ANS-PNS	45.671	0.274
ZF-ZFR	130.316	0.150
ZMs-ZTi	66.714	0.160

The mean square error (MSE) values of all variables are relatively high, which indicates that the SVM model has a large prediction error on these variables.
All variables had low r-squared (R^2) values, indicating that the SVM model had limited ability to explain these variables.
These conditions indicate that the model cannot effectively predict these mid-face measurement parameters in the current

Summary and Conclusion

Least square linear regression	Support vector machine regression (SVM)	Random forest regression
It performed well in predicting the "ZMs-ZMsR" variable with an MSE of 6.484 and an R^2 of 0.906, showing strong explanatory power and a small prediction error. However, it does not perform well on extreme values, and the residual is large.	It performed poorly in predicting the "ZMs-ZMsR" variable, with MSE of 56.512 and R^2 of only 0.178, showing large prediction error and weak explanatory power. It performs poorly on all variables and may require further tuning and optimization.	It performed well in predicting the "ZMs-ZMsR" variable, with MSE of 6.698 and R^2 of 0.877, showing good predictive performance and explanatory power. The prediction of extreme values is stable and the residual is small. It also performs well on other variables and is suitable for data with complex and non-linear relationships.

The prediction of extreme value by Random forest is stable and the residual is small. It also performed well on other variables. And Random forest model hasn obvious advantages in dealing with data complexity and multicollinearity, and is the best predictive model among the three models. Its the best choice for predicting facial development in patients with Syndromic Craniosynostosis

Future Research Directions

1. For future research, the parameters of the random forest model can be further optimized and more relevant variables can be included to further improve the prediction accuracy.
2. Removing features with high VIF values or dimensionality reduction through principal component analysis (PCA) can reduce the influence of multicollinearity on the model and improve the stability and prediction accuracy of the model.
3. Use regularization methods such as Ridge regression and Lasso regression to introduce penalty terms and reduce multicollinearity effects
4. Hyperparameter adjustment of SVM , such as adjusting regularization parameters, can significantly improve the model's prediction performance.
5. Data expansion: Expand and enrich the data set by collecting more patient data, especially patient data of different age groups and different disease severity.

References

1，Fearon, J. A. (2016). Evidence-based medicine: Craniosynostosis. Plastic and Reconstructive Surgery, 137(1), 289e-297e. https://doi.org/10.1097/PRS.0000000000001962
2，Jiang, F., Jiang, Y., Zhi, H., Dong, Y., Li, H., Ma, S., ... & Wang, Y. (2017). Artificial intelligence in healthcare: past, present and future. Stroke and Vascular Neurology, 2(4), 230-243. https://doi.org/10.1136/svn-2017-000101
3，Kearns, A. E., Campbell, C. A., Kinter, J., & Cunningham, M. L. (2017). Skeletal findings in syndromic craniosynostosis: A
radiographic review. American Journal of Medical Genetics Part A, 173(8), 2029-2044. https://doi.org/10.1002/ajmg.a.38312
4，Vallée, A., Vallee, J. H., & Matillon, Y. (2020). Artificial intelligence and predictive models in craniosynostosis. Journal of
Craniofacial Surgery, 31(6), e528-e532. https://doi.org/10.1097/SCS.0000000000006617

Predicting Midface Development in Syndromic Craniosynostosis Using Penalized Regression Models in Machine Learning

Authors
WANG JIALE

Affiliations
Institute of Mathematical Science, University of Malaya, Kuala Lumpur, Malaysia

Introduction

Syndromic craniosynostosis is a group of rare genetic disorders characterized by the premature fusion of one or more of the cranial sutures, the joints between the bones of the skull. This premature fusion can lead to an abnormal head shape and, in some cases, associated physical and cognitive impairments.Treatment for syndromic craniosynostosis typically involves a multidisciplinary approach, with input from neurosurgeons, craniofacial surgeons, geneticists, and other specialists. The primary goal is to allow for normal brain growth and development by surgically correcting the abnormal skull shape. Additional therapies may address other medical issues that arise.

age	ANS-PNS	N-ANS	ZMs-ZMs]	N-S-BA	N-SO	S-N	SO-BA	S-SO-BA	ZF-ZFR	ZMs-ZTi
15y 10m	52.11	54.87	54.44	142.09	78.62	66.65	22.13	137.83	98.26	49.56
11.3m	33.81	29.79	42.13	147.15	58.75	49.65	14.21	150.00	79.56	34.69
9y 6m	44.65	47.35	50.59	128.24	67.50	59.74	18.07	145.26	91.94	44.70
4y 11m	38.50	36.46	46.16	139.48	70.21	59.21	14.57	145.47	82.97	44.10
5y 10m	43.35	46.17	43.58	134.71	64.32	55.11	21.09	153.20	86.58	44.05
4y 3m	37.49	36.36	50.23	134.69	64.12	55.77	16.03	146.68	87.14	39.87
1y	34.03	28.77	40.79	137.81	57.91	49.23	13.67	160.10	78.54	36.59
1.75y	35.37	31.97	40.48	141.19	60.98	50.06	14.34	163.36	77.17	37.48
3y 3m	36.91	34.06	47.59	137.61	61.54	54.37	16.48	137.04	83.11	37.94
4.6m	30.92	28.93	35.66	149.76	51.24	41.93	13.10	156.28	73.66	32.54
3y 8m	38.83	39.69	49.22	144.05	68.46	57.99	17.64	146.36	87.47	45.47
4y 10m	40.66	41.67	49.31	140.74	70.10	58.29	17.61	165.75	89.82	44.70
13y 2m	44.46	48.02	55.52	141.97	71.99	60.19	20.24	134.89	101.31	52.05
9m	32.88	28.85	38.08	141.28	55.27	45.82	12.95	153.39	72.80	33.31
10y 1m	45.85	50.36	50.30	135.22	71.60	62.41	22.54	139.08	89.95	49.45
2y 5m	36.66	34.97	44.47	135.25	65.42	56.13	14.46	150.39	80.59	35.86
10.8m	33.32	29.37	38.62	141.16	54.39	44.97	12.60	152.75	71.11	32.29
17m 13d	35.00	32.19	37.20	138.85	57.92	47.85	13.08	153.13	75.13	35.76
1.2y	31.87	30.55	42.73	145.12	58.59	49.44	12.67	147.20	75.35	35.01
9y 4m	42.61	44.93	51.85	134.80	70.82	62.30	20.17	140.16	92.40	46.43
10y 3m	43.33	46.14	55.92	145.64	74.26	59.73	17.67	157.32	100.53	52.79
8.3m	33.16	28.15	38.74	144.95	55.75	46.79	13.96	153.75	73.88	34.35
6y 4m	41.63	43.29	46.33	138.96	63.99	55.26	20.68	148.36	91.63	51.14
15y 4m	47.13	57.86	56.70	135.50	84.57	63.02	17.02	169.01	98.92	50.07

The dataset for this study consisted of 24 patients, each representing a patient diagnosed with sydromic craniosynostosis. The response variables include 'ZMs-ZMsR', 'N-ANS', 'ANS-PNS', 'ZF-ZFR', 'ZMs-ZTi', likely derived from anatomical or physiological measurements, used for assessing conditions or anatomical relationships.

Objective

The primary objective of this study is to compare three regression models—Least squares linear Regression, Ridge Regression, and Lasso regression—for predicting midface region measurements based on cranial base parameters.

Methodology

Data Description
Multicollinearity Check (VIF Test)
Normality Test (Shapiro-Wilk Test)
Analysis Methodology
(1) Least Squares Linear Regression
(2) Ridge Regression
(3) Lasso Regression

Multicollinearity is a common problem in regression analysis where the predictors are highly correlated. It was to detect a degree of multicollinearity which is the correlation between independent variables.
We will perform the Shapiro-Wilk Test to check their normality. This determines whether the variable is suitable for linear regression analysis. The Shapiro-Wilk test is used to test the residuals of each model and to test the normality of the distribution, which is the basic assumption of linear regression analysis

Results

Variable	Mean Squared Error (MSE)	R-squared (R^2)
ZMs-ZMsR	6.48	0.91
N-ANS	12.84	0.89
ANS-PNS	8.82	0.86
ZF-ZFR	20.02	0.87
ZMs-ZTi	13.12	0.84

Least Squares Linear Regression

Variable	Mean Squared Error (MSE)	R-squared (R^2)
ZMs-ZMsR	5.81	0.80
N-ANS	5.90	0.93
ANS-PNS	6.23	0.88
ZF-ZFR	6.11	0.94
ZMs-ZTi	10.27	0.75

Ridge Regression

Variable	Mean Squared Error (MSE)	R-squared (R^2)
ZMs-ZMsR	1.95	0.93
N-ANS	2.76	0.97
ANS-PNS	9.08	0.83
ZF-ZFR	3.90	0.96
ZMs-ZTi	4.33	0.89

Lasso Regression

The Mean Squared Error (MSE) for the predictions is approximately 6.78.
MSE is a measure of the average squared difference between the estimated values and the actual value.
The R-squared (R^2) value is about 0.856.
This value represents the proportion of the variance for the dependent variable that's explained by the independent variables in the model. An R^2 value closer to 1 indicates a model that explains a large portion of the variance.

The mean square error (MSE) value represents the mean square error between the estimate and the actual value. The greater the MSE value, the greater the prediction error. In this analysis, the mean square error value ranges from 5.81 to 10.27, and there is a significant difference in prediction accuracy between different variables. The R Squared value represents the proportion of the difference in the dependent variable that can be predicted by the independent variable. The R Squared value is between 0.75 and 0.94.

The Lasso Regression results show that each dependent variable is superficially good enough, and the R Squared value fluctuates from 0.836 to 0.992. The N-ANS factor has a value of 0.97 at R Squared, which probably means that the model explains 97%. The correlation of the ZMs-ZMsR variable is also low (R^2=0.93). Although this value is reduced from the original 100%, the model still describes the proposed phenomenon fairly well.

Discussion

We evaluated the Least Squares Linear Regression, Ridge Regression and Lasso Regression, revealed large differences in the predictive power of facial measurements in patients with Sydromic Craniosynostosis. The Lasso Regression, which has unique features for handling multicollinearity problems and variable selection, has an R^2 of 0.97 for N-ANS, suggesting that the model can explain 97% of the variance. Under the least squares model, VIF values can indicate problems of high multicollinearity. Processing multicollinearity with Ridge Regression and Lasso Regression can reduce the effect of insignificant predictors, thereby improving the reliability of the model.

Conclusion

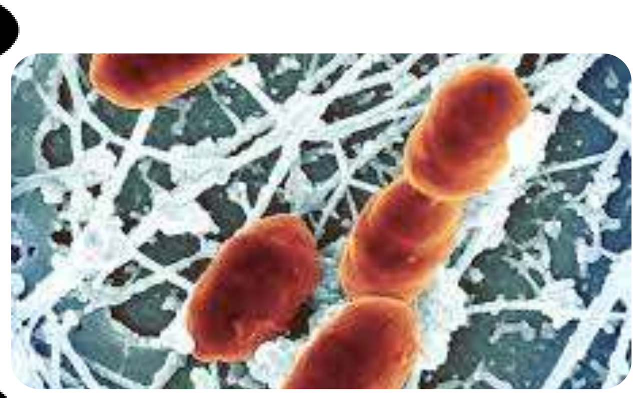
The primary objective of this study was to compare the performance of three regression models—Least Squares Linear Regression, Ridge Regression, and Lasso Regression—in predicting midface region measurements in patients with syndromic craniosynostosis. Our analysis demonstrates that the choice of regression model significantly impacts the accuracy and reliability of predictions, with Lasso Regression showing the most promising results. The lowest MSE and highest R-squared values obtained with Lasso Regression indicate its superior capability in predicting midface development in patients with syndromic craniosynostosis. This makes it an invaluable tool for clinicians involved in treatment planning and intervention strategies.

Forecasting Ricketsiosis Case Numbers in Malaysia: Analysis of Count Time Series Following Generalized Linear Model

By: WANG CAN S2037359

What are Rickettsia and Rickettsiosis?

Rickettsia are small, obligately intracellular Gram-negative bacilli. They are distributed among various hematophagous arthropod vectors and cause Rickettsioses, an acute undifferentiated febrile illness, and are often accompanied by headache, myalgias, and malaise.



Cases of Rickettsiosis in Malaysia

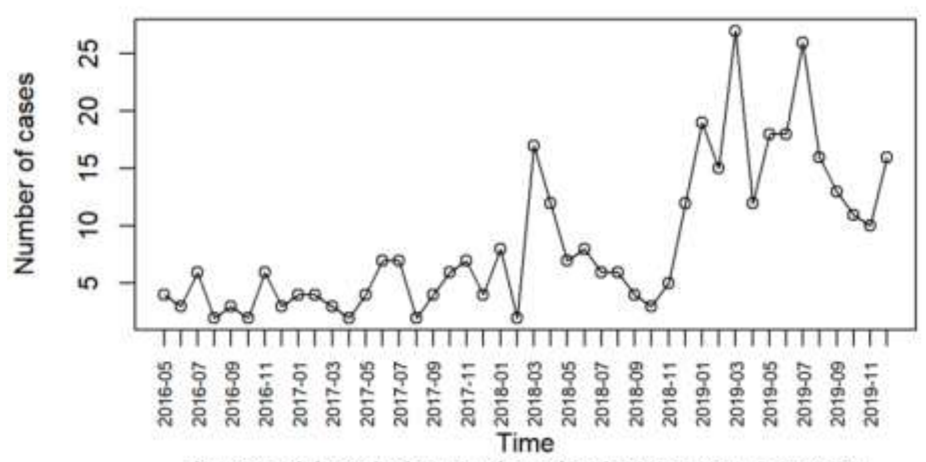


Figure 1. Rickettsiosis Cases in Malaysia (May 2016 - Dec 2019)

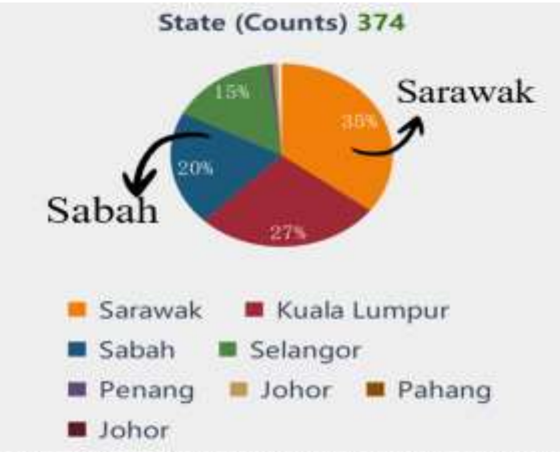


Figure 2: Distribution of Rickettsiosis Cases in Malaysia by State (May 2016 - December 2019)

As shown in Figure 1, we can see a clear trend of increasing Rickettsiosis cases over the years from May 2016 to December 2019.
Additionally, Figure 2 shows the distribution of cases by state. From this pie chart, we can see that a significant proportion of cases are concentrated in East Malaysia, particularly in the states of Sabah and Sarawak.

Objectives Of Project

Objective 01
To develop and train a **Seasonal ARIMA (SARMA)** model based on past data from 2016 to 2019, in order to forecast Rickettsiosis cases in Malaysia.

Objective 02
To develop and train a **generalized linear model(GLM)** based on past data from 2016 to 2019, in order to forecast Rickettsiosis cases in Malaysia.

Objective 03
Compare the performance of SARIMA model and generalised GLM in forecasting discrete time series data

Objective 04
Understand the results of the predictions, highlighting the information gained from the analyses and their potential impact on public health.

What is SARIMA Model?

- **The Seasonal Autoregressive Integrated Moving Average (SARIMA) model** is an extension of the ARIMA model that supports time series data with a seasonal component. It combines autoregressive (AR), differencing (I), and moving average (MA) components to capture both the seasonal and non-seasonal behavior of the time series.
- **Components of SARIMA:**
  The SARIMA model is represented as:

ARIMA (p, d, q) (P, D, Q) m

Lowercase letters — Uppercase letters

(p, d, q): These parameters represent the non-seasonal part of the model.
- ***p: Non-seasonal autoregressive order.***
  It indicates the number of lag observations included in the model.
- ***d: Non-seasonal differencing order.***
  It indicates the number of times the raw observations are differenced to make the time series stationary.
- ***q: Non-seasonal moving average order.***
  It indicates the size of the moving average window used to smooth the time series.

(P, D, Q): These parameters represent the seasonal part of the model.
- ***P: Seasonal autoregressive order.***
  It captures the relationship between a value and its seasonal lags.
- ***D: Seasonal differencing order.***
  It helps to remove seasonal trends in the data, making it easier to model.
- ***Q: Seasonal moving average order.***
  It captures the relationship between a value and past forecast errors from seasonal lags.

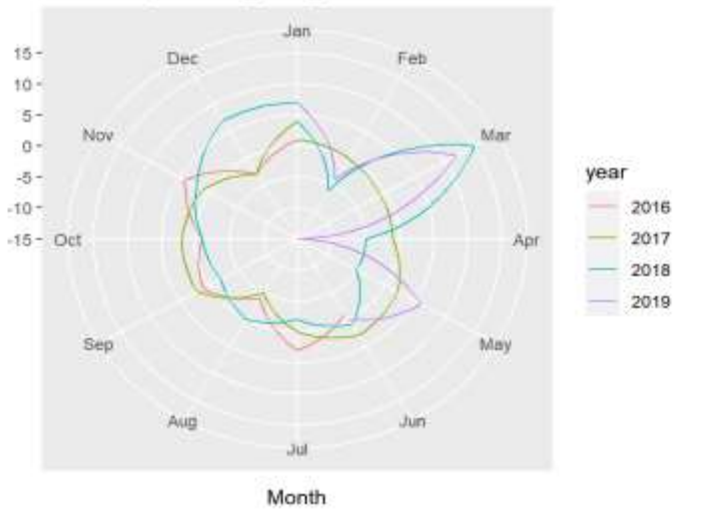
Data

Years	Month	Positive	Years	Month	Positive
2016	May-16	4		Mar-18	17
	Jun-16	3		Apr-18	12
	Jul-16	6		May-18	7
	Aug-16	2		Jun-18	8
	Sep-16	3		Jul-18	6
	Oct-16	2		Aug-18	6
	Nov-16	6		Sep-18	4
	Dec-16	3		Oct-18	3
2017	Jan-17	4		Nov-18	5
	Feb-17	4		Dec-18	12
	Mar-17	3	2019	Jan-19	19
	Apr-17	2		Feb-19	15
	May-17	4		Mar-19	27
	Jun-17	7		Apr-19	12
	Jul-17	7		May-19	18
	Aug-17	2		Jun-19	18
	Sep-17	4		Jul-19	26
	Oct-17	6		Aug-19	16
	Nov-17	7		Sep-19	13
	Dec-17	4		Oct-19	11
2018	Jan-18	8		Nov-19	10
	[illegible]	[illegible]		[illegible]	[illegible]

Monthly Rickettsiosis Case Numbers in Malaysia
- **Study Area** :
  Includes 8 states in Malaysia
- **Time Period** :
  May 2016 ~ December 2019
  (Total of 44 observations)
- **Data Type:**
  Discrete Time Series

Seasonal Check

The radar plot shows the monthly distribution of cases from 2016 to 2019.
Seasonality Detection: The plot indicates a recurring peak in March across multiple years, suggesting the presence of seasonal patterns in the data. Given the observed seasonality, we could train a SARIMA model to capture these seasonal effects for more accurate forecasting.



Fit SARIMA Model

1.Split train and test set:

Train Set: 2016 May ~ 2019 Jun   Test Set: 2019 Jul ~ 2019 Dec

2. Model Selection

To choose the best model, we used three information criterias: AIC, AICc, and BIC. They are information criteria to measure the goodness of fit of a statistical model. **Lower values mean that the model fits better.**

Table 1. Information Criteria

Model	AIC	AICc	BIC
SARIMA(0,1,1)(0,1,1)[12]	159	160	162
SARIMA(0,1,1)(0,0,1)[12]	219	220	224
SARIMA(1,1,0)(0,0,1)[12]	219	220	224

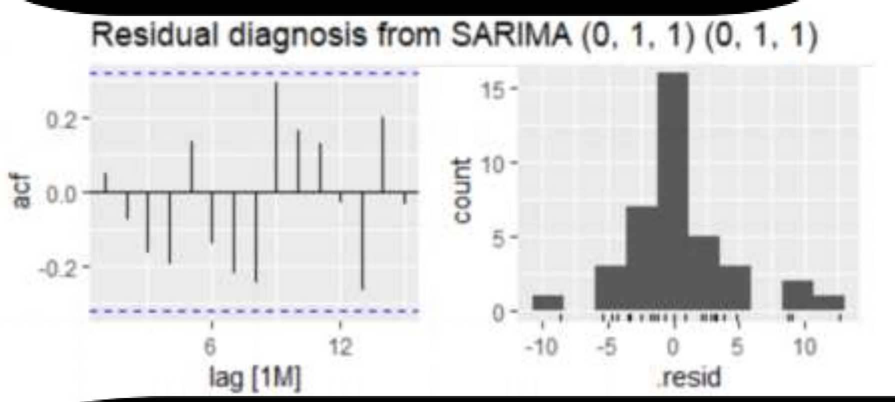
As shown in Table 1, our best model is SARIMA(0,1,1)(0,1,1)[12], which has the lowest AIC, AICc, and BIC scores.

3.SARIMA Model

(p, d, q) (P, D, Q) m

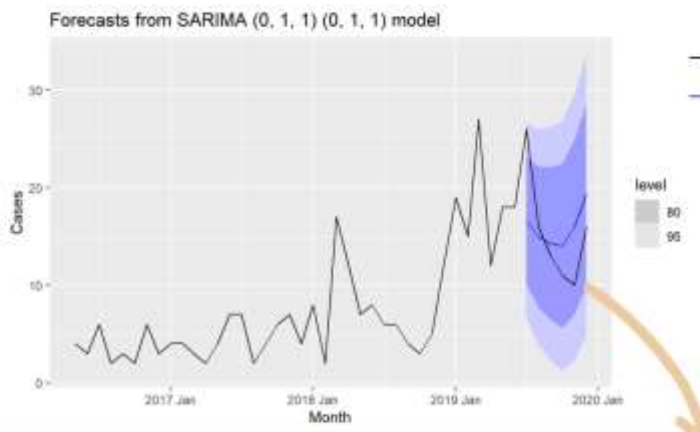
Model: **SARIMA(0,1,1)(0,1,1)[12]**
- No autoregressive terms, so p = 0 and P = 0.
- One differencing term (d = 1) and one seasonal differencing term (D = 1).
- One moving average term (q = 1) and one seasonal moving average term (Q = 1).
- A seasonal period of 12 months (m = 12).

Residual Diagnostics (SARIMA)



The ACF plot indicates that most residuals fall within the confidence intervals, suggesting no autocorrelation.
The histogram reveals that the residuals are approximately normally distributed but exhibit a slight skew and a few extreme values, indicating the presence of outliers.
The residual diagnostics suggest that the SARIMA(0,1,1)(0,1,1)[12] model provides a reasonable fit to the data.

Forecast Value Vs. Actual Value (SARIMA)



—Actual Value: 2016 May ~ 2019 Dec
— Forecast Value: 2019 Jul ~ 2019 Dec

- ***The SARIMA model provides forecasted values generally follow the trend of the actual values, indicating a reasonable model fit.***
- ***Almost all actual values fall within the 95% confidence interval, suggesting the model's predictions are reliable.***
- ***Some divergence between forecasted and actual values indicates potential areas for model refinement.***

80% Confidence Interval (Light Blue Shading)
95% Confidence Interval (Dark Blue Shading)
The shading areas show the confidence intervals.
Indicates the range within which the actual values are expected to fall with 80% and 95% probabilities.

Why would we want to develop a GLM?

The monthly cases of Rickettsiosis form a count time series with low counts, ranging from 2 to 27. As noted by Stephan (2016), such series are not well-suited for methods designed for continuous distributions. Traditional SARIMA methods, which assume a continuous sample space, are therefore inappropriate for this type of data. Given the absence of a generalized approach, we decided to investigate a specific model for the data. Our analysis focuses on the Generalized Linear Model (GLM), which is particularly effective for this purpose, and we compared its performance against SARIMA models.

GLM

- Generalized Linear Models (GLM):
  is the most common model for analyzing count data
- **Compared to the usual GLM, the innovative GLM we used** from Tobias Liboschik **considers the empirical autocorrelation of the count time series data.**
- **It can be defined in the following form:**

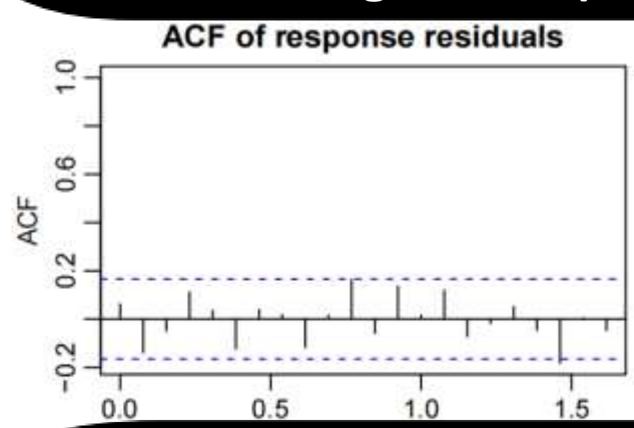
$$g(\lambda_t) = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}(Y_{t-i_k}) + \sum_{l=1}^{q} \alpha_l g(\lambda_{t-j_l}) + \eta^T X_t$$

Fit GLM Model

```
R > summary(fit)
Call:
tsglm(ts = train_timeseries, model = list(past_obs = c(1, 12)),
   xreg = train_regressors, link = "log", distr = "nbinom")
```

Capture short-term serial dependencies through a first-order autoregressive term and annual seasonality through a twelfth-order autoregressive term, both specified in the model parameter list element named 'past_obs'.
Due to the more severe infections in East Malaysia, we include the number of cases in East Malaysia as a covariate. For distribution, we chose the negative binomial. Due to our data being discrete and not satisfying the Poisson distribution assumption (variance equals mean).

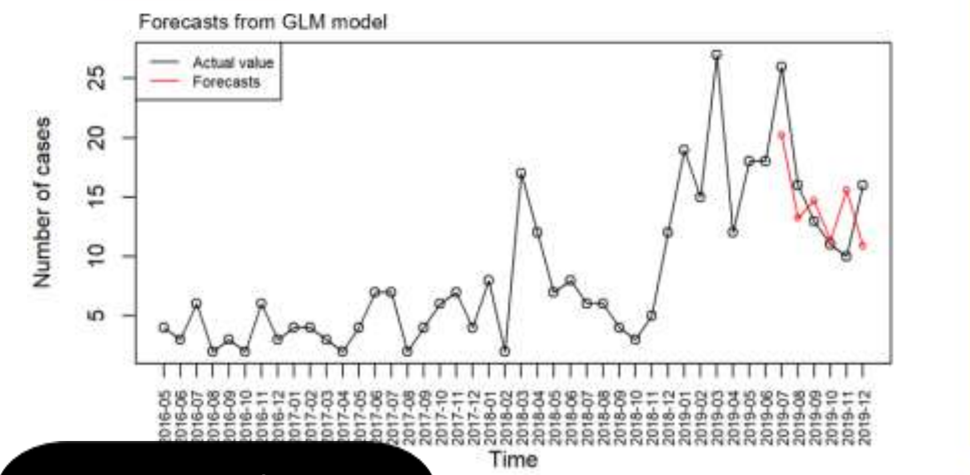
Residual Diagnostics (GLM)



The ACF plot helps identify any correlations between residuals over different time lags. All autocorrelations lie within the 95% confidence bands (blue dashed lines), suggesting that the residuals do not exhibit significant autocorrelation. It shows that does not exhibit any autocorrelation or seasonality which has not been taken into account by the model. This is a desirable property indicating a good model fit.

Forecast Value Vs. Actual Value (SARIMA)

—Actual Value: 2016 May ~ 2019 Dec
—Forecast Value: 2019 Jul ~ 2019 Dec



The forecasts (red lines) have generally trended in line with the actuals and in most months the forecasts have almost accurately predicted the number of cases, e.g. September and October 2019. But there are still months that deviate significantly from the actual values, which is common because our prediction horizon, i.e. the test set, is too short. We can look forward to the predictive performance of this model over longer time periods.

Comparison

Table 2. Forecast Accuracy Comparison of SARIMA and GLM for Rickettsiosis Case Numbers

Model	RMSE	MAE	MPE	MAPE
SARIMA	4.945685	4.015738	-12.67336	26.86456
GLM	4.103619	3.543842	-0.1469728	23.89788

RMSE, MAE, MPE, and MAPE are all measures of forecast accuracy. Lower values indicate that the model performs better.

Conclusion:
- **From the table 2, The GLM outperforms the SARIMA model in all measures.**
- Recommendation:
- **For forecasting Rickettsiosis case numbers in Malaysia, the GLM is preferred due to its higher accuracy and its suitability for discrete data.**

- Accuracy: GLM model shows lower **RMSE and MAE**, indicating higher accuracy compared to the SARIMA model.
- Prediction Bias: GLM's **MPE** is closer to zero, suggesting it has less prediction bias.
- Error Percentage: GLM's lower **MAPE** demonstrates better performance in percentage error terms.

Conclusions

01
Both SARIMA and GLM models capture the overall trend in rickettsiosis case numbers in Malaysia .

02
For forecasting Rickettsiosis case numbers in Malaysia, the GLM is preferred due to its higher accuracy and its suitability for discrete data.

03
The public health department can use statistical models to develop more precise prevention and control plans for rickettsiosis.

Comparative Analysis of ARIMA Model and Generalized Linear Model for Predicting MRSA Infection Cases in Malaysia

WANG CUI
S2100923

Introduction

Methicillin-resistant *Staphylococcus aureus* (MRSA) bacterium is a significant hospital-acquired pathogen. It resists many antibiotics that treat regular staph infections, posing a serious public health risk . *Staphylococcus* skin infection usually begin as swollen, painful red bumps that may resemble pimples or spider bites. Sometimes the bacteria stay on the skin, but they can also go deeper and cause serious infections in bones, joints, surgical wounds, the bloodstream, heart valves, and lungs.

Data Background

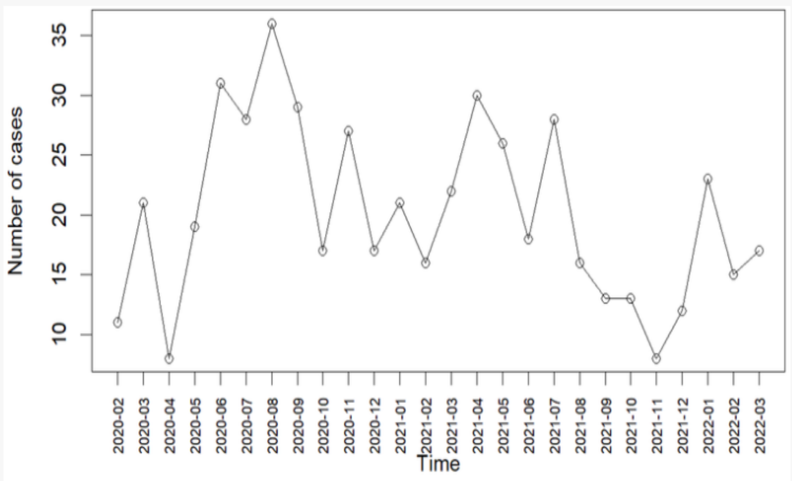


Figure 2. Monthly Case Numbers of MRSA in the two hospitals Malaysia from February 2020 to March 2022.

In this research , we study the monthly cases of MRSA infections in two major hospitals in Klang Valley from February 2020 to March 2022.The trend in the number of cases over these 26 months, as shown in figure 2, is typically between 7 to 36.

The monthly number of MRSA infections is a count time series . The variance of our data is 59.26, exceeds the mean, 20.15, indicating overdispersion.

Motivation for the Study

The monthly infections number of MRSA is a typically count time series with low counts. According to Stephan (2016),This kind series can no longer be dealt with using the approximative methods that are appropriate for continuous distributions. Given this requirement, it is clear that traditional ARIMA methods are inappropriate as they assume a continuous sample space for the data. In the absence of a generalized methodology, we are interested in considering some specific classes of models and noting the suitability of particular models for the count data. As a result, the Generalized Linear Model (GLM), which is the most common model for analyzing count data, became the model we wanted to analyze and compare with ARIMA.

Objective

01.Fitting and forecasting with Auto Regression Integrated Moving Average (ARIMA) Model

02.
Fitting and forecasting with Generalised Linear Model (GLM)

03.
Compare the performance of the ARIMA model and Generalized Linear Models (GLM) in forecasting MRSA Infection Number.

Methodology

ARIMA Model

(Auto Regression Integrated Moving Average)

- Is a statistical tool used for analyzing and forecasting time series data.
- It combines differencing with autoregressive(AR) and moving average(MA) components.

The model is defined as:

$$X_t = c + \sum_{i=1}^{p} \phi_i X_{t-i} + \varepsilon_t + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j}$$

where

- p:order of the AR part
- q:order of the MA part
- $\phi i,\theta j$ are coefficients
- c is a constant
- ϵt is the error term

GLM

(Generalized Linear Model)

- The Generalised Linear Model (GLM), which is the most common model for analyzing count data

The model is defined as:

$$g_{(\lambda_t)} = \beta_0 + \sum_{k=1}^{p} \beta_k \tilde{g}(Y_{t-ik}) + \sum_{l=1}^{q} \alpha_l g(\lambda_{t-jl}) + \eta^T X_t$$

where

- { $Y_t : t \in N$ } stands for a count time series
- { $X_t : t \in N$ } is a time-varying r-dimensional covariate vector
- $\lambda_t = E(Y_t|F_{t-1})$ is the conditional mean of the count time series.
- The distributions we are interested in are the Poisson distribution and the negative binomial distribution.

Similar to ARIMA

- p:order of the AR part
- q:order of the MA part
- α,β are coefficients

Result and Discussion

Step1:Fitting the ARIMA model

Objective 01

1.Split train and test set

Train Set: 2020 February ~ 2021 October

Test Set: 2021 November~ 2022 March

(The training set and test set account for 80% and 20% of the total data，respectively)

2.Stationary check

- ADF Test:(Null hypothesis-Nonstationary)

```
        Augmented Dickey-Fuller Test

data:  train.ts_adj
Dickey-Fuller = -2.5332, Lag order = 2, p-value = 0.3692
alternative hypothesis: stationary
```

ADF Test Result of d=0

```
        Augmented Dickey-Fuller Test

data:  difference_train.ts_adj
Dickey-Fuller = -4.0548, Lag order = 2, p-value = 0.02134
alternative hypothesis: stationary
```

ADF Test Result of d=1

```
        Augmented Dickey-Fuller Test

data:  difference_train.ts_adj
Dickey-Fuller = -4.0548, Lag order = 2, p-value = 0.02134
alternative hypothesis: stationary
```

ADF Test Result of d=2

we chose the difference order d=2, the p-value is less than 0.05, means that the series is stationary.

3. Model Selection

Table 1. Information Criteria

Model Name	AIC	AICc	BIC
ARIMA(1,2,1)	143.6286	145.2286	146.4619
ARIMA(0,2,1)	144.6893	145.4393	146.5782
ARIMA(1,2,0)	151.1609	151.9109	153.0498
ARIMA(0,2,0)	159.0685	159.3038	160.0129

ARIMA (p=1,d=2,q=1)

Parameters P=1,Q=1 in this model means that the model uses 1 lagged data and one prediction error to make predictions.
d=2: minimum number of differencing needed to make the series stationary.

Step2:Residual Diagnosis

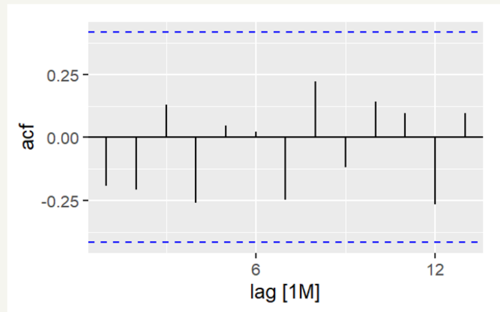


Figure 3: ACF Plot of ARIMA(1,2,1) models residual

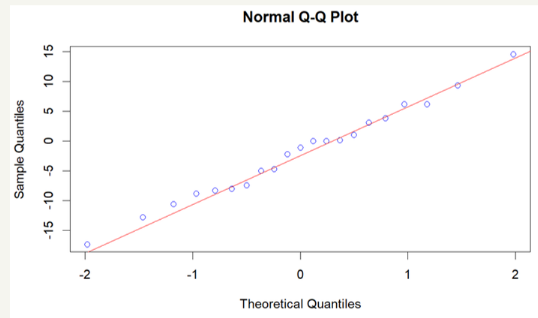


Figure 4: Normal Q-Q Plot of ARIMA(1,2,1) models residual

- The ACF plot shows that most of the residuals are within the confidence interval, indicating little autocorrelation.
- The Q-Q plot of residual shows normally distributed since the points appears as roughly a straight line.

The residual diagnostics suggest that the ARIMA(1,2,1) model provides a reasonable fit to the data.

Objective 02

Model Selection

1.AIC & QIC method

- Select the model with the smallest AIC and QIC,the result obtained is model with parameters **p=1 and q=1**

Negative binomial with p=1,q=1

2.Compare the Scoring values

Model name	logarithmic	quadratic	spherical	rankprob	dawseb	normsq	sqerror
Poisson GLM	3.582194	-0.0240271	-0.1715512	4.227328	5.294492	2.20975	48.38627
Negative-Binomial GLM	3.367235	-0.0373987	-0.1933294	4.020282	4.891814	0.85000	48.38627

Table 2: Scoring values of Poisson and Negative Binomial

Similar to ARIMA, the parameters P=1,Q=1 in this model means that the model uses 1 lagged data and one prediction error to make predictionsThe selected distribution is Negative Binomial, it means that the model can be applied to overdispersion data.

Residual Diagnosis

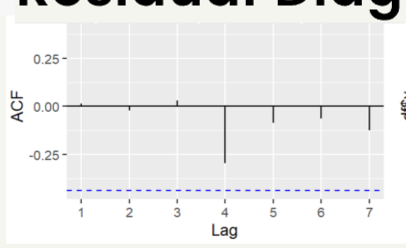


Figure 6: ACF Plot of GLM models residual

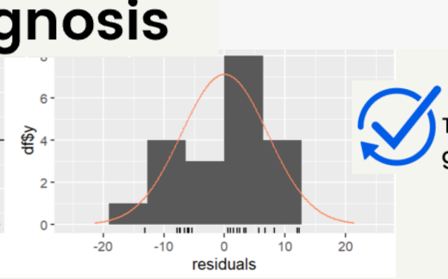


Figure 7: Histogram Plot of GLM models residual

The residual diagnostics suggest that the Negative Binomial generalized linear model provides a reasonable fit to the data.

Objective 03

Comparison

1.Compare Forecasting Plots

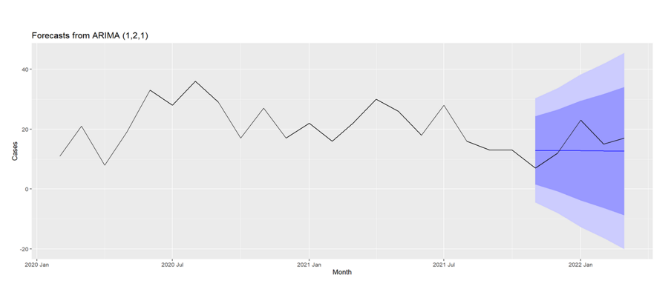


Figure 9.Plot of Forecasting Results of ARIMA Model from 2021 Nov~2022 Mar

VS

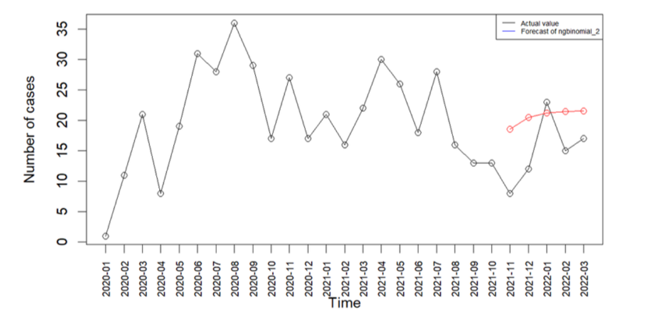


Figure 10: Plot of Forecasting Results of GLM Model from 2021 Nov~2022 Mar

- ARIMA model and generalized linear model (GLM) both roughly capture the overall trend of the actual values
- Generalized Linear Model (GLM) seems better captures the upward trend of the actual

2.Compare Accuracy Values

Model name	RMSE	MAE	MAPE
ARIMA(1,2,1)	5.724014	4.716344	35.32365
Negative-Binomial GLM	7.758253	6.855503	65.89999

Table 4 Forecast Accuracy Valus of ARIMA and GLM for MRSA Infection Cases

Conclusion

1. Both ARIMA and generalized linear model (GLM) capture the overall trend in MRSA infection numbers in two major hospitals in Klang Valley.
2. In comparison , generalized linear model (GLM) is more appropriate for forecasting MRSA infection number.
3. Continuous data collection on MRSA infection cases in Malaysia is essential. Addressing the data's limitations is crucial for developing and analyzing more robust models in the future.

ELECTRON SCATTERING FROM ZINC ATOM

SIM3021 MATHEMATICAL SCIENCE PROJECT
RANJITHAA A/P VASU (U2103400)

UNIVERSITI MALAYA

Supervisor : Dr. Mohd Zahurin Mohamed Kamali
Co-Supervisor : Dr.Shahizat Bin Amir
Institute of Mathematical Sciences (ISM), Faculty Science

Abstract

This study investigates electron scattering from zinc atoms, emphasizing the calculation of phase shifts and differential cross-sections. The research employs the Schrödinger equation to model electron interactions, converting complex equations into a solvable format using a similarity transformation. Numerical solutions are obtained via the Runge-Kutta method solved using R studio software. Results indicate that higher electron energy increases both phase shifts and differential cross-section values, revealing significant changes in scattering amplitude and cross-section distribution. Additionally, the influence of potential energy affected by exchange and polarization effects becomes more prominent with increased electron energy, highlighting the physical parameters' critical role in scattering dynamics. These insights have implications for material analysis, radiation therapy, and semiconductor technology, providing a basis for optimizing electron scattering conditions.

Objective

a)To compute elastic scattering amplitude and differential cross-sections for electron-Zn atom interactions spanning an energy range from 25 eV to 80eV.
b)To validate the computational approach by comparing the computed differential cross-sections with theoretical models
c)To develop numerical algorithms for each formulated problems and programming for computations in generating numerical solutions via R Studio software.

Literature Review

1. Electron Scattering

- The study of electron scattering using cathode ray was initially founded by Sir J.J. Thomson in 1897, which is also the groundbreaking experiment in the late 19th century.
- Garcia and Wang (2024) developed a sophisticated theoretical model to mimic electron scattering in intricate molecular systems by combining quantum mechanics with machine learning algorithms.

2. Elastic and inelastic scattering

- The first investigations by Sir J.J. Thomson in 1897, which showed that cathode rays were made of negatively charged particles subsequently dubbed electrons, laid the groundwork for our understanding of electron scattering.

3. Cross-section Analysis

- In 1984, Trajmar et al. published a comprehensive compilation and critical assessment of available cross-sections for electron-molecule scattering.
- Lee and Park (2017) defined total cross-sections as the likelihood that an electron will experience any kind of scattering event, providing a fundamental understanding of these processes.
- Smith and Jones (2019) investigated differential cross-sections, shedding light on how scattering probabilities change with energy and direction

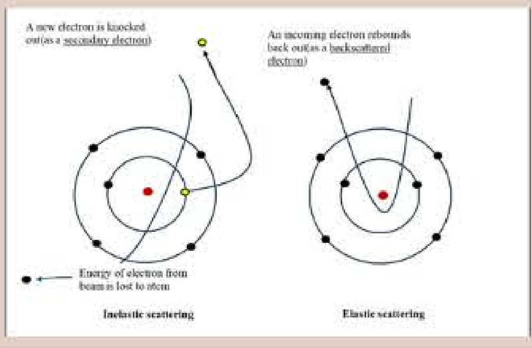
4. Electron-Zinc (Zn) scattering processes

- The optical excitation functions of Zn for many transitions have been measured in Shpenik et al. (1973) and Souter et al. (1974).
- Zn represents a quasi-two-electron atom, like helium, beryllium, and magnesium, making it valuable for testing theoretical models, especially where relativistic effects might be important (Marinković et al., 2019).

Methodology

1. Static field approximation, which simplifies the electron-atom interaction by assuming a static, time-independent potential.
2. To solve the Schrödinger equation, the study employs the Frobenius series expansion method.
3. Using Runge-Kutta, the R studio software used to numerically solve the reduced form of mathematical model.
4. The results are presented graphically, illustrating the relationship between electron energy and phase shifts, as well as the variation in differential cross-sections.

Figure 1: Types of scattering



Mathematical Model

The numerical method used to solve the second order radial differential Schrödinger equation in the Runge-Kutta method.

$$\frac{d^2u_l(r)}{dr^2}+\left(k^2-\frac{l(l+1)}{r^2}-U_0(r)\right)u_l(r)=0$$

where,

$$U_0(r)=2V_0(r)$$

$$k=\left(\frac{\sqrt{2m(E-E_L)}}{h}\right)$$

$$U_0(r)=+2\left(1+\frac{1}{r}\right)e^{-2r}$$

Optical potential,

$$V_o=V_s+V_p+V_x$$

The static potential is given by Salvat and Parellada which is,

$$V_s=-\left(\frac{Z}{r}\right)\phi(r)$$

The polarization and exchange potential is given by,

$$V_p(r)=-\left[\frac{(\alpha_p r^2)}{2(r^2+d^2)^3}\right]$$

$$V_x(r)=\frac{1}{2}\left\{\frac{1}{2}k^2-V_s(r)-\left[\left(\frac{1}{2}k^2-V_s(r)\right)^2-8\pi\rho(r)\right]^{\frac{1}{2}}\right\}$$

Results

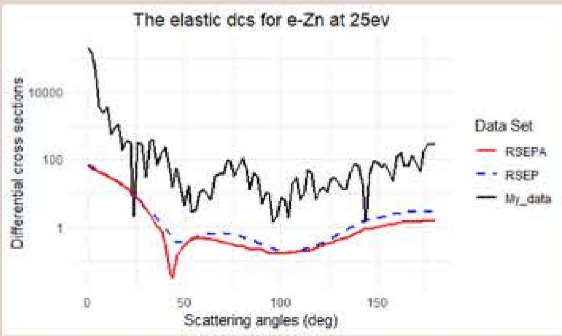


Figure 2. Differential cross section for electrons incident at 25 eV on bound zinc atom in the SPX approximation.

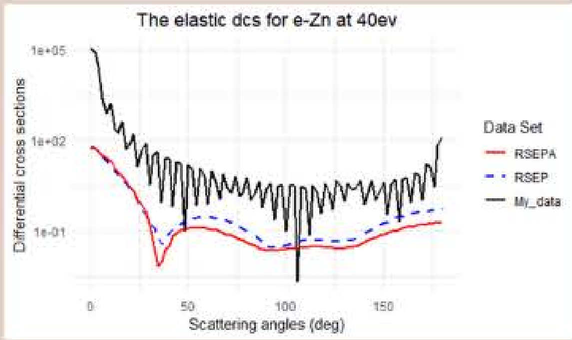


Figure 3. Differential cross section for electrons incident at 40 eV on bound zinc atom in the SPX approximation

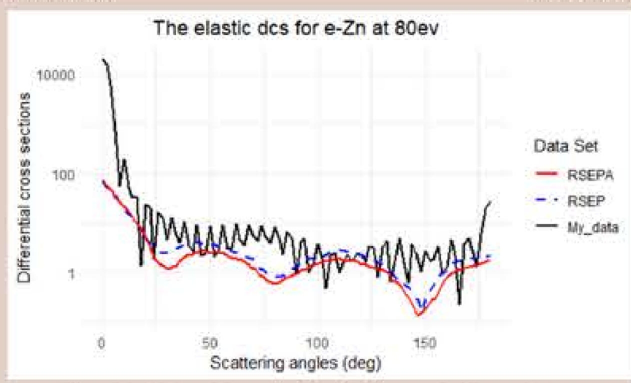


Figure 5.3. Differential cross section for electrons incident at 80 eV on bound zinc atom in the SPX approximation.

- The significant differences between the DCS results obtained in my study using the SPX approximation and Kelemen et al (2023) is be largely attributed to the simplifications inherent in the SPX model.

The significant differences observed can be attributed to several factors:

- The SPX model employs a straightforward combination of static, polarization, and exchange potentials. This contrasts with the detailed and parameterized potentials in the RSEPA and RSEP models, which include contributions from spin-orbit interactions and relativistic corrections.
- Kelemen et al (2023) include absorption potentials that account for inelastic scattering processes. The SPX model, however, is purely elastic and does not include such effects.
- The Kelemen et al (2023) model predicts detailed patterns in forward and backward scattering, including the impact of various partial waves and the inelastic contributions. The SPX model, with its simplified approach, may not fully replicate these detailed patterns, resulting in differences in the observed local maxima and minima, as well as variations in the intensity of scattering at extreme angles.
- Kelemen et al (2023) include detailed treatments of exchange and correlation effects through various potentials, such as the Furness-McCarthy (FM) potential and its modifications.
- The accuracy of the scattering computations is strongly dependent on the choice of angular momentum, especially at higher angles when the contribution of higher-order partial waves becomes greater.

Conclusion

Overall, this work serves an extension study of Kelemen et al (2023) who use complex model but in my study we used simple model. The significant differences between the DCS results obtained in my study using the SPX approximation and Kelemen et al (2023) can be largely attributed to the simplifications inherent in the SPX model. These simplifications, while beneficial for computational efficiency, result in the exclusion of complex interactions and inelastic effects that are critical for a detailed understanding of electron scattering processes. This highlights the necessity for careful consideration when selecting a model for DCS calculations, ensuring that the chosen methodology aligns with the specific goals and accuracy requirements. In conclusion, this work of study demonstrated the fulfillment of the two objectives outline at the beginning of the study.

Acknowledgement

I sincerely thank my supervisor, Dr. Mohd Zahurin Mohamed Kamali and Dr.Shahizat Bin Amir in completing this project and their valuable insights and constructive criticism. I would also like to thank my family and friends for their unwavering support and encouragement throughout this academic journey.

References

1. Kelemen, V. I., Demes, S., & Remeta, E. Y. (2023). Theoretical study of elastic electron scattering by zinc atoms in the framework of relativistic optical potential model. Journal of Electron Spectroscopy and Related Phenomena, 266, 147365. https://doi.org/10.1016/j.elspec.2023.147365.
2. Meredith, R. J., Williamson, W., Montemayor, V. J., Öztürk, N., & Antolak, A. J. (1990). Low-energy elastic scattering of electrons by bound silicon and germanium atoms. Journal of Applied Physics, 68(10), 4937–4941. https://doi.org/10.1063/1.347078

A Review on General Relativity and its Application to Cosmology

Sia Chen Yi

Universiti Malaya, Institut Sains Matematik, Malaysia

UNIVERSITI MALAYA

Abstract

General relativity is a crucial topic that serves as a cornerstone in the revolution of our understanding of spacetime and gravity. Concepts such as dark matter, dark energy, and black holes are fundamentally based on the principles of general relativity.

Objectives

The main goals for our research are:

- Explain the derivation of Einstein's Field Equation in general relativity;
- Apply Einstein's Field Equation to construct some basic cosmological models;
- Briefly explain the origin of cosmological constant and the notion of dark energy.

Historical Overview

The story behind the Albert Einstein's work on general relativity:

- In 1905, Albert Einstein published his paper entitled "On the Electrodynamics of Moving Bodies", which introduced his theory of special relativity;
- Albert Einstein tried to seek for the natural extension of special relativity which included the effects of gravity;
- He comes out the Einstein's Equivalence Principle and realized that the space is curved;
- He sought out his classmate at Zurich's University, Marcel Grossmann to teach him tensor calculus and Riemann geometry, which describe the movement of objects in space and time and also describe the dynamic frame of reference;
- Early development of tensor calculus are done by Gregorio Ricci-Curbastro, Tulio Levi-Civita, and Professor Luigi Bianchi;
- Einstein's student David Hilbert also attempted to derive Einstein's Field Equation before Einstein. In the end, both of them managed to derive the same equation by using different methods by November 1915;
- Einstein derived the equation by realizing Second Bianchi Identity is proportional to Noether's theorem. While Hilbert derived the equation using Lagrangian approach.

Tensor Calculus

- Covariant Tensor,

$$A_{i_1 i_2 \dots i_m} = \bar{A}_{\alpha_1 \alpha_2 \dots \alpha_m} \frac{\partial \bar{x}^{\alpha_1}}{\partial x^{i_1}} \frac{\partial \bar{x}^{\alpha_2}}{\partial x^{i_2}} \dots \frac{\partial \bar{x}^{\alpha_m}}{\partial x^{i_m}}. \tag{1}$$

- Contravariant Tensor,

$$A^{j_1 j_2 \dots j_n} = \bar{A}^{\beta_1 \beta_2 \dots \beta_n} \frac{\partial x^{j_1}}{\partial \bar{x}^{\beta_1}} \frac{\partial x^{j_2}}{\partial \bar{x}^{\beta_2}} \dots \frac{\partial x^{j_n}}{\partial \bar{x}^{\beta_n}}. \tag{2}$$

- Connection,

$$\Gamma^h_{\ ij} = x^h_{\ \lambda}(\bar{x}^\lambda_{\ ji} + \bar{\Gamma}^\lambda_{\ \mu\nu} \bar{x}^\mu_{\ i} \bar{x}^\nu_{\ j}). \tag{3}$$

- Christoffel Symbol,

$$\mathring{\Gamma}^\lambda_{\ \mu\nu} = \frac{1}{2} g^{\lambda\alpha}(\partial_\nu g_{\alpha\mu} + \partial_\mu g_{\alpha\nu} - \partial_\alpha g_{\mu\nu}). \tag{4}$$

- Covariant Derivative for Covariant Vector,

$$\nabla_j A_i = \partial_j A_i - \mathring{\Gamma}^l_{\ ij} A_l. \tag{5}$$

- Covariant Derivative for Contravariant Vector,

$$\nabla_j A^i = \partial_j A^i + \mathring{\Gamma}^i_{\ lj} A^l. \tag{6}$$

- Riemann-Christoffel Curvature Tensor,

$$R^\lambda_{\ \mu\alpha\nu} = \partial_\alpha \mathring{\Gamma}^\lambda_{\ \mu\nu} - \partial_\nu \mathring{\Gamma}^\lambda_{\ \mu\alpha} + \mathring{\Gamma}^\lambda_{\ \sigma\alpha} \mathring{\Gamma}^\sigma_{\ \mu\nu} - \mathring{\Gamma}^\lambda_{\ \sigma\nu} \mathring{\Gamma}^\sigma_{\ \mu\alpha}. \tag{7}$$

- Ricci Tensor,

$$R_{\mu\nu} = R^\alpha_{\ \mu\alpha\nu}, \tag{8}$$
$$= R_{\beta\mu\alpha\nu} g^{\alpha\beta}. \tag{9}$$

- Ricci Scalar,

$$R = R^\nu_{\ \nu}, \tag{10}$$
$$= R_{\mu\nu} g^{\mu\nu}. \tag{11}$$

Einstein's Field Equation

Einstein's Approach

From Bianchi Second Identity,

$$\nabla_\mu (R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R) = 0, \tag{12}$$
$$G^{\mu\nu} = R^{\mu\nu} - \frac{1}{2} g^{\mu\nu} R. \tag{13}$$

Since $\nabla_\mu G^{\mu\nu} = 0$ and we assume $\nabla_\mu T^{\mu\nu} = 0$,

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = \kappa T_{\mu\nu}. \tag{14}$$

After we solving for κ,

$$R_{\mu\nu} - \frac{1}{2} g_{\mu\nu} R = 8\pi G T_{\mu\nu}. \tag{15}$$

Hilbert's Approach

Our total action gravity is as follows:

$$S = \gamma S_g + \kappa S_m, \tag{16}$$
$$S = \int \mathcal{L}(\Phi^i, \nabla_\mu \Phi^i) d^n x. \tag{17}$$

We can separate S_g into three parts:

$$\delta S_g = \delta S_1 + \delta S_2 + \delta S_3, \tag{18}$$
$$\delta S_1 = \int d^n x \sqrt{-g} g^{\mu\nu} \delta R_{\mu\nu}, \tag{19}$$
$$\delta S_2 = \int d^n x \sqrt{-g} R_{\mu\nu} \delta g^{\mu\nu}, \tag{20}$$
$$\delta S_3 = \int d^n x (-\frac{R}{2\sqrt{-g}} \frac{\delta g}{\delta g^{\mu\nu}}) \delta g^{\mu\nu}. \tag{21}$$

δS_1 will equals to 0 by Stokes' Theorem. After that we will get the following:

$$\delta S_g = \int ([R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu}] \sqrt{-g} \delta g^{\mu\nu}) d^n x. \tag{22}$$

After we get δS_g, then we substitute it into (16) to get Einstein's Field Equation,

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} = 8\pi G T_{\mu\nu}. \tag{23}$$

Cosmological Constant

We take $S = \gamma(S_g + S_p) + \kappa S_m$,

$$R_{\mu\nu} - \frac{1}{2} R g_{\mu\nu} + \Lambda g_{\mu\nu} = 8\pi G T_{\mu\nu}. \tag{24}$$

Cosmos with only matter and radiation

- Ricci scalar of FLRW Metrics,

$$R = 6[\frac{\ddot{a}}{a} + (\frac{\dot{a}}{a})^2 + \frac{\kappa}{a^2}]. \tag{25}$$

- Friedmann Equation,

$$(\frac{\dot{a}}{a})^2 + \frac{\kappa}{a^2} = \frac{8\pi G \rho}{3} + \frac{\Lambda}{3}. \tag{26}$$

- Acceleration Equation,

$$\frac{\ddot{a}}{a} = \frac{\Lambda}{3} - \frac{4\pi G}{3}(\rho + 3p). \tag{27}$$

- Energy Conservation Equation,

$$\dot{\rho} = -3\frac{\dot{a}}{a}(\rho + p). \tag{28}$$

By using Energy Conservation Equation, we do able to express energy density in terms of ω.

$$\rho = \rho_0 (\frac{a_0}{a})^{3(1+\omega)}. \tag{29}$$

- Einstein-de Sitter Universe,

$$a(t) = a_0 (\frac{t}{t_0})^{\frac{2}{3}}, \tag{30}$$
$$t_0 = \frac{2}{3H_0}. \tag{31}$$

- Radiation Dominated Universe,

$$a(t) = a_0 (\frac{t}{t_0})^{\frac{1}{2}}, \tag{32}$$
$$t_0 = \frac{1}{2H_0}. \tag{33}$$

The Universe with Cosmological Constant

- Einstein's Static Universe,

$$a_0 = \frac{1}{\Lambda}. \tag{34}$$

- De Sitter Universe with $\kappa = -1$,

$$a(t) = \sqrt{\frac{3}{\Lambda}} \sinh(\sqrt{\frac{\Lambda}{3}} t). \tag{35}$$

- De Sitter Universe with $\kappa = 0$,

$$a(t) = a_0 e^{\sqrt{\frac{\Lambda}{3}} t}. \tag{36}$$

- De Sitter Universe with $\kappa = 1$,

$$a(t) = \sqrt{\frac{3}{\Lambda}} \cosh(\sqrt{\frac{\Lambda}{3}} t). \tag{37}$$

Now we are looking at some tools which help us to relate the non-observable parameters with observable parameters.

- Redshift,

$$a(t_e) = \frac{1}{1+z}. \tag{38}$$

- Instantaneous Physical Distance,

$$d_p(t) = a(t) \int_0^r \frac{dr}{\sqrt{1 - \kappa r^2}}. \tag{39}$$

- Luminosity Distance Formula,

$$d_L = (1+z) a(t_0) r. \tag{40}$$

Conclusions

In conclusion, we have successfully derived Einstein's Field Equations by employing the Einstein-Hilbert action, a pivotal step in our journey through the realms of theoretical physics. These equations have served as the foundation for deriving several other fundamental equations, including the Friedmann Equation, Acceleration Equation, and Energy Conservation Equation, among others.

Furthermore, our exploration extended to various cosmological solutions within the framework of general relativity. We delved into the origin of the cosmological constant in Einstein's Field Equations and its reinterpretation as dark energy, based on empirical evidence such as supernova measurements. Additionally, we discussed how cosmological models are verified through the measurement of key cosmological parameters, redshift, instantaneous physical distance, and also luminosity distance formula on the intricate workings of our universe.

References

[1] Peter Challis Adam G. Riess, Alexei V. Filippenko. Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological Constant. *Astronomial Journal*, 1998.

[2] Sean M. Carroll. *Spacetime and Geometry: An Introduction to General Relativity*. Cambridge University Press, 2023.

[3] F Atrio J Martin, E Ruiz. *Relativity and Gravitation in General*. World Scientific Publishing Co. Pte. Ltd, 1999.

[4] Horst Eckardt Myron W Evans, Stephen J Crothers. *Criticisms of the Einstein Field Equation*. Cambridge International Science Publishing, 2011.

[5] Oliver Piattella. *Lecture Notes in Cosmology*. Springer, 2018.

[6] Bernd G. Schmidt. *Einstein's Field Equations and Their Physical Implications*. Springer, 2000.

[7] Bernard F. Schutz. *A First Course in General Relativity*. Cambridge University Press, 2009.

[8] Joydeep Sengupta U.C.De, Absos Ali Shaikh. *Tensor Calculus Second Edition*. Alpha Science International LTD., 2008.

[9] S. Walters. How Einstein Got His Field Equations. *In Commemoration of General Relativity's centennial*, 2016.

IDENTIFYING SIGNIFICANT FACTORS IN HOSPITAL AND COMMUNITY-ACQUIRED MRSA USING LOGISTIC REGRESSION MODELLING

UNIVERSITI MALAYA

MUHAMMAD AFIQ BIN ZAINIDAR
U2103651

SUPERVISORS: DR NORLI ANIDA ABDULLAH
DR NUR ANISAH MOHAMED

INTRODUCTION

Methicillin-Resistant Staphylococcus aureus **(MRSA) infection originates from a group of *Staphylococcus* bacteria that has become resistant to many antibiotics typically used to treat standard *Staphylococcus* infections. Most MRSA infections occur in people who have been hospitalized or have had contact with healthcare settings such as hospitals and dialysis centers, and are known as hospital-acquired MRSA (HA-MRSA). MRSA is also found in community clusters, such as among child care workers, athletes, and people who live in close proximity. This form is known as community-acquired MRSA (CA-MRSA).**

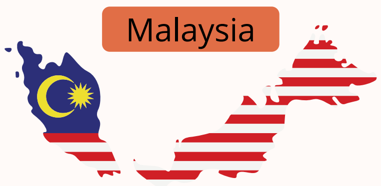
OBJECTIVES

1. **To investigate the association between type of gene and infection type using chi-squared test.**
2. **To model the dataset for the identification of significant factors in HA-MRSA and CA-MRSA using logistic regression (LR) model.**
3. **To evaluate the performance of the model using testing dataset.**

METHODOLOGY

STUDY AREA AND SOFTWARE

Malaysia



RStudio

DATA SOURCE

NATIONAL INSTITUTES OF HEALTH
MINISTRY OF HEALTH MALAYSIA

CHI-SQUARE TEST

Examined the association between MRSA type (HA or CA) and the presence of 16 virulence genes.

LR MODEL

- **The dataset was partitioned into two subsets**
- **90% for training and 10% for testing.**
- **Analyzed data to identify key predictors of HA and CA-MRSA, focusing on virulence genes, age, and infection source.**

MULTICOLLINEARITY

VIF (Variance Inflation Factor) was used to detect multicollinearity in our model, helping to identify which predictor variables are correlated with each other.

BACKWARD ELMINATION METHOD

Improved the model by gradually removing the least important predictors.

MODEL PERFORMANCE EVALUATION

- **The performance of the model was evaluated using accuracy and AUC (Area Under the ROC Curve).**
- **Accuracy measures the proportion of correctly classified instances, while AUC assesses the model's ability to distinguish between classes.**

RESULTS AND DISCUSSIONS

		HA	CA
	sea	38	17
	seb	17	7
	sec	15	13
	sed	5	5
	see	44	18
	pvl	32	33
	hlα	370	153
	hlβ	354	146
	fnbA	46	32
	fnbB	28	11
	tsst-1	82	33
	sccmecI	13	2
	sccmecIII	5	3
	sccmecIV	342	139
TYPE OF GENE	sccmecV	20	15

The contingency table shows the distribution of virulence genes in HA and CA. The p-value of the chi-square test for the association between MRSA type and the presence of these 16 virulence genes is 0.01366, which is below the conventional significance level of 0.05. This indicates that there is significant evidence to suggest an association between MRSA type (HA or CA) and the presence of various virulence genes.

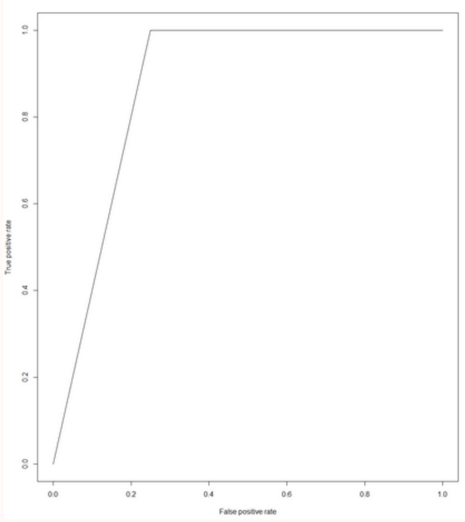
In Model 1, all predictors were included, yet none exhibited a satisfactory p-value, indicating insignificance. Subsequently, multicollinearity assessment using VIF revealed severe multicollinearity with SCCmecIV, registering a coefficient of 11.74. Thus, in Model 2, SCCmecIV was omitted. In Model 2, none of the predictors are statistically significant, and the chi-square goodness-of-fit test yields a very low p-value, indicating a poor model fit. No multicollinearity was detected in Model 2. Consequently, the backward elimination method was employed to improve the model's accuracy and interpretability in model 3.

```
Call:
glm(formula = HaCa ~ pvl, family = binomial(link = "logit"),
    data = train)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.9050     0.1060   8.538   <2e-16 ***
pvl          -0.6173     0.2901  -2.128   0.0333 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 602.00  on 489  degrees of freedom
Residual deviance: 597.61  on 488  degrees of freedom
AIC: 601.61

Number of Fisher Scoring iterations: 4
```

There is only one significant predictor variable in model 3. The odds ratio for *pvl* (*Panton-Valentine leukocidin*) is 0.5394. This indicates the odds of *pvl* being detected in CA-MRSA are 53.94% higher than in HA-MRSA.

	0 (Predicted)	1 (Predicted)
0 (Actual)	9	3
1 (Actual)	0	42

From the confusion matrix, the model's accuracy is calculated to be 94.44%, indicating that the model is highly accurate.



The AUC for the ROC curve shown is 0.875, which means that Model 3 has an 87.5% probability of correctly distinguishing between HA-MRSA and CA-MRSA.

CONCLUSION

In conclusion, this study identifies *pvl* as a key significant factor in CA-MRSA, consistent with existing scientific research. It provides robust statistical evidence supporting the significant role of *pvl* using logistic regression model . This underscores the need for targeted *pvl*-specific interventions to improve the prevention and management of CA-MRSA infections.

REFERENCES

Classification of Abelian Group and Related Problems

Pong Ching Hin

Universiti Malaya

Introduction

The fundamental theorem of finitely generated abelian group is frequently used to classify and study the different structure of abelian groups. It is a generalization of the fundamental theorem of finite abelian group proven by Leopard Kronecker (1870) and who defined a arbitrary set of elements with an operation satisfying certain laws which now known as finite abelian groups.

F.G. Frobenius and L. Stickelbelger (1879) proved the fundamental theorem of finite abelian groups with a group theoretical approach and formulated the finite abelian group theory to modern mathematical views. Later, Poincare (1900) and Emmy Noether (1926) gave a proof for the fundamental theorem of finitely generated abelian groups which does not have the restriction on finite abelian groups.

We can use the theorem to generalize the structures of finitely generated modules over a principle ideal domain and use the theorem for other applications such as study the structure of the groups and finding isomorphic abelian groups of an given group.

Fundamental Theorem of Finitely Generated Abelian Groups

There are two forms for the theorem namely:

Invariant Factors Decomposition Every finitely generated abelian group G is isomorphic to a finite direct sum of cyclic groups in which the finite cyclic groups are in the order $m_1, m_2, \cdots, m_t$ where $m_1 > 1$ and $m_1 | m_2 | \cdots | m_t$.

$$G \cong \mathbb{Z}_{m_1} \oplus \mathbb{Z}_{m_2} \oplus \cdots \oplus \mathbb{Z}_{m_t} \oplus F$$

where F is a free abelian group.

Elementary Divisors Decomposition Every finitely generated abelian group is isomorphic to a finite direct sum of cyclic groups, each of which is either infinite or of order a power of a prime.

$$G \cong \mathbb{Z}_{p_1^{n_1}} \oplus \mathbb{Z}_{p_2^{n_2}} \oplus \cdots \oplus \mathbb{Z}_{p_t^{n_t}} \oplus F$$

where F is a free abelian group, p_i are primes and n_i are positive integers.

Remark on the Theorem

- For a finite abelian group, the decomposition of the group does not contains the free abelian group part.
- If the group G is a free abelian group, then it is isomorphic to the group F stated in the theorem.

List of Nonisomorphic Groups

Using the theorem, we can get all the possible combinations of nonisomorphic abelian group of order n.
For invariant factors:

- Find the prime factorization of n and list out all the possible m_t.
- For each m_t, work out all the possible combination of $m_1, m_2, \ldots, m_t$ such that $m_1 | m_2 | \ldots | m_t$.

For elementary divisors:

- Find the prime factorization of n and find the partition of each power of the prime factors.
- The direct sum of the p^k where k is the partition of the power of p is the nonisomorphic list for the prime factor. Last we just pair each combination of a prime factor to other factors to obtain the list of the abelian group of order n.

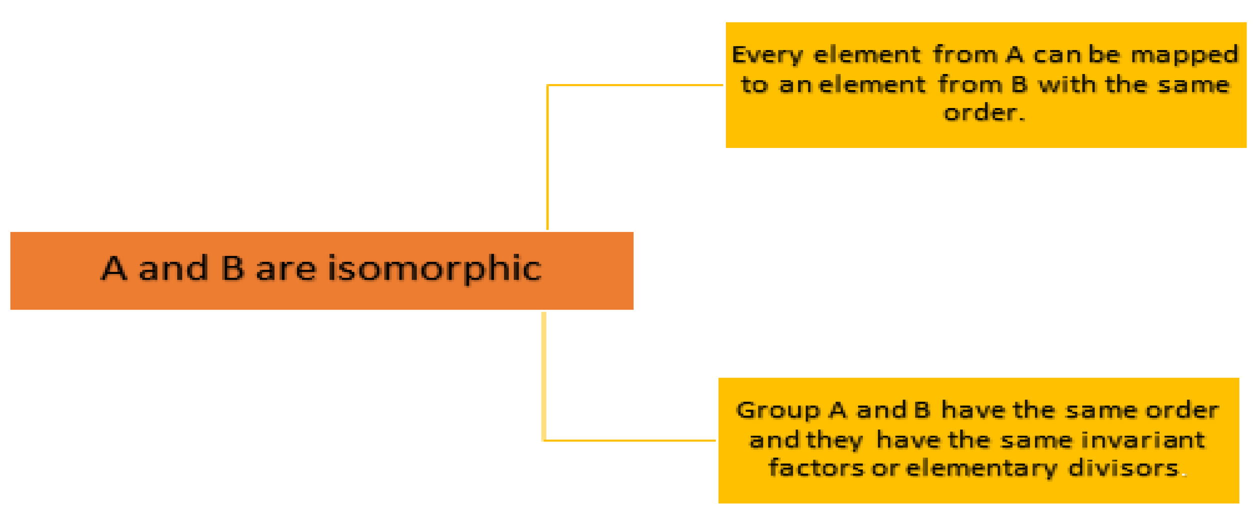
For a group of order $n = p_1^{\alpha_1} p_2^{\alpha_2} \cdots p_k^{\alpha_k}$, the number of nonisomorphic abelian group is the product of the number of partition for each α_i.

Maximum Order of Elements in an Abelian Group

The maximum order of elements in an abelian group can be determine by using the theorem and some tricks. If we know the decomposition of the group, we can find the maximum order of elements from the group by finding the lowest common multiple of the order of each cyclic group.

For example, the element in the group $\mathbb{Z}_4$ has element of order 4 but the group $\mathbb{Z}_2 \oplus \mathbb{Z}_2$ only has element of maximum order 2 even though both group are of order 4. This is a useful tools that we can use to determine the isomorphism.

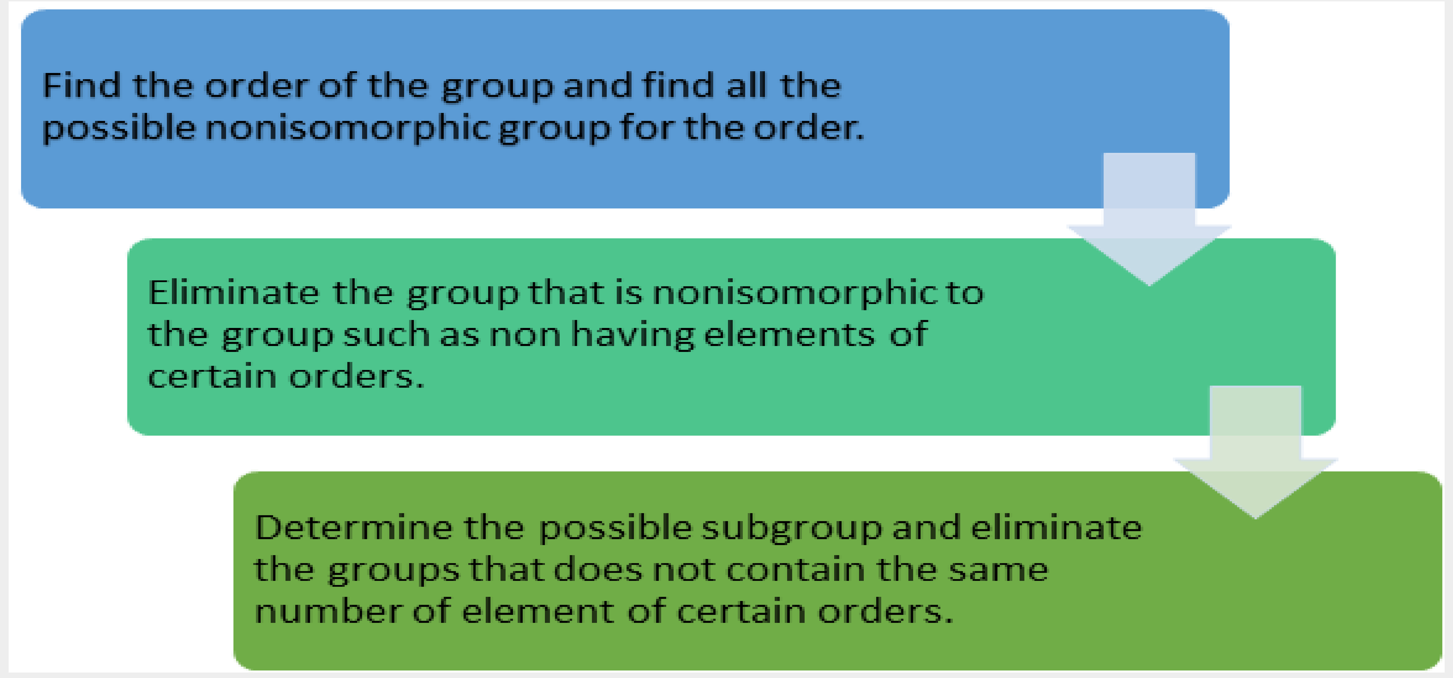
Isomorphism Between 2 Abelian Groups



We can show that two groups are isomorphic or not by using the conditions above. As mention in the previous part, we can conclude that $\mathbb{Z}_4 \ncong \mathbb{Z}_2 \oplus \mathbb{Z}_2$ because they exist an element of order 4 in one group that cannot be mapped into the other group. Using this we can also match up the pair of invariant factors and elementary divisors from the list for order n.

Classification of Abelian Groups

By using the statement from before we can now determine the isomorphic group of any given abelian group by the following steps:



Note that for a group $\mathbb{Z}_{p^m} \oplus G$ is a subgroup of $\mathbb{Z}_{p^n} \oplus G$ for all integer $0 \leq m \leq n$ and G is a group.

Other Applications

- Finding the number of subgroups for a given finitely generated abelian group.
- Finding number of elements with certain order for a given finitely generated abelian group.
- Investigate and classify the structure of the factors group of a finitely generated abelian group.

References

Frobenius, F. G., & Stickelberger, L. (1879).Ueber gruppen von vertauschbaren elementen..

Kronecker, L. (1870). *Auseinandersetzung einiger eigenschaften der klassenzahl idealer complexer zahlen.*

Noether, E. (1926).Der endlichkeitssatz der invarianten endlicher linearer gruppen der charakteristik p. *Nachrichten von der Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-Physikalische Klasse*, *1926*, 28–35.

ANALYZING URBAN DEVELOPMENT AND ACCESSIBILITY THROUGH FRACTAL DIMENSIONS OF ROAD NETWORKS IN PETALING JAYA, SELANGOR, MALAYSIA

NUR DANIA BINTI NOR AZMI
Supervised by: Dr. Shahizat bin Amir and
Dr. Mohd Zahurin Mohamed Kamali

UNIVERSITI MALAYA

INTRODUCTION

Urbanization, a hallmark of modern society, necessitates sustainable planning as cities grow. In Petaling Jaya (PJ), Malaysia, a city of over 600,000 residents spanning 97.2 square kilometers, managing urban expansion and accessibility is crucial. Traditional analysis methods, focusing on metrics like road length and population density, often miss urban complexity. This study proposes using fractal dimensions to analyze Petaling Jaya's road networks, uncovering patterns and structures of urbanization. Fractal analysis offers enhanced connectivity insights and a visual understanding of urban form, aiding better infrastructure development and planning for improved accessibility and quality of life.

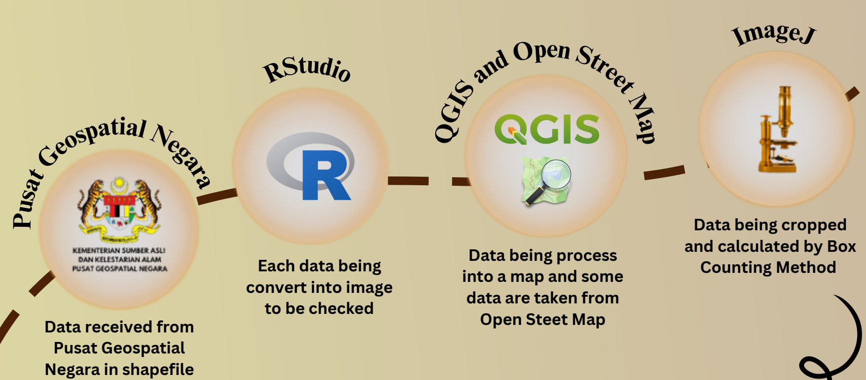
PROBLEM STATEMENT

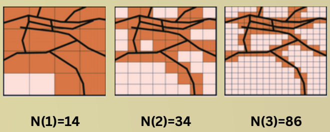
Petaling Jaya city has complex connections of its road networks and growth patterns. Accessibility of a city is a crucial thing to its development, thus the road networks contribute a major role to determine the correlation. However, we need to use conventional methods to observe and statistically analyze road networks, its accessibility and the urban development

OBJECTIVES

- To investigate and observe road networks of Petaling Jaya and its accessibility through fractal dimensions.
- To investigate and observe how accessibility correlates to urban development through fractal dimensions.

METHODOLOGY





Box Counting Method

This method entails placing a grid of boxes with different sizes over a fractal pattern and tallying the number of boxes that intersect with the pattern. This procedure is iterated with 2,3,4,6,8,12,16,32,64 boxes for this research, and the correlation between the box count and their sizes is leveraged to compute the fractal dimension

RESULTS

	PJU (North)	SS (West)	Seksyen (East)	PJS (South)
Transportation	1.5677	1.6267	1.5255	1.4990
Build Environment	1.7015	1.7125	1.5968	1.4523
Overlap Both	1.7469	1.7545	1.6473	1.5530

DISCUSSIONS

Petaling Jaya and Its Section Analysis

- Petaling Jaya has an overall fractal dimension of 1.8204, indicating high urbanization. All city sections have dimensions exceeding 1.5000, reflecting favorable urban conditions.
- Sungei-Way Subang (SS) has the highest complexity that is 1.7545, dedicating SS as most urbanized section.
- Petaling Jaya Utara (PJU) is the second highest of urbanization that is 1.7469 which highly accessible and well-developed.
- Seksyen has a moderate complexity with fractal dimension of 1.6473 and have increasing attractiveness and residential growth.
- Petaling Jaya Selatan (PJS) has the lowest complexity of that is 1.5530 which likely to be less accessible and urbanize but has the potential for future development.

Accessibility

- Overall road network fractal dimension 1.7008 that is high in complexity
- PJS has the lowest road network dimension with 1.4990, indicating less efficient infrastructure of road.
- SS is the most accessible of all section, with highest connectivity to other sections.

Built Environment

- Overall fractal dimensions is 1.7871 with SS and PJU that are most well-developed with numerous amenities.
- Seksyen is close to optimal urbanization with 1.5968.
- PJS has lower complexity but steady population growth and potential for future development.

CONCLUSION

Analyzing Petaling Jaya's road networks through fractal dimensions reveals insights into accessibility and urbanization, identifying areas for potential planning improvements. This study highlights the complexity of the PJ's road structure and suggests integrating fractal dimensions into urban planning to enhance mobility, land use, and infrastructure investments. These findings support sustainable urban growth and improve quality of life for the future

ACKNOWLEDGEMENT & REFERENCES

KEMENTERIAN SUMBER ASLI DAN KELESTARIAN ALAM PUSAT GEOSPATIAL NEGARA
MBPJ

FORECASTING THE NUMBER OF MRSA CASES IN MALAYSIA USING TIME SERIES MODEL

MUHAMAD IKLIL HAKIMI BIN HASAN (U2103286)

SUPERVISOR: DR ANISAH BINTI MOHAMED @ RAHMAN
DR NORLI ANIDA BINTI ABDULLAH

NATIONAL INSTITUTES OF HEALTH
MINISTRY OF HEALTH MALAYSIA

UNIVERSITY OF MALAYA
The Leader in Research & Innovation

UM-CDACS

INSTITUT STATISTIK MALAYSIA (ISM)

INTRODUCTION

Methicillin-resistant Staphylococcus aureus (MRSA) infections are rapidly spreading in Malaysia, and a particularly high number of cases are occurring. Forecasting of MRSA growth is essential for healthcare authorities to implement timely preventive measures.

This study focuses on predicting the future trend of MRSA infections in Malaysia using time series models, and it finds that the ARIMA model is the most suitable one. Through detailed analysis, ARIMA (1,2,1) was identified as the most effective model for this purpose.

The findings demonstrate that the ARIMA model can reliably forecast future MRSA cases, providing valuable insights for public health authorities to enhance their strategic planning and response efforts

OBJECTIVE

- Justify the selection of the ARIMA model based on its fit to the data compared to other models.
- Train the ARIMA model using MRSA case data from 2020 to 2022.
- Forecast future MRSA cases in Malaysia using the trained ARIMA model.
- Highlight key insights gained from the forecast.

METHODOLOGY

1

- Split the data into training and testing sets.
- Transform the data to reduce variance and differentiate it to achieve stationarity.
- Verify stationarity using the Augmented Dickey-Fuller (ADF) test.
- Ensure the p-value from the ADF test is less than 0.05. If the p-value is greater than 0.05, apply additional differencing to the data.

2

- Analyze ACF and PACF plots to identify potential models based on data patterns.
- Use ACF and PACF to assess the suitability of AR (p) and MA (q) models and identify possible candidate models.

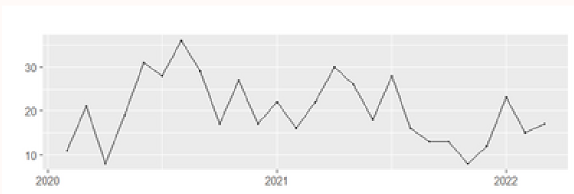
3

- Determine the parameters for potential models and select the most appropriate ones.
- Choose the best model based on the lowest AIC or BIC value.
- Verify that the residuals' ACF and PACF indicate white noise.
- Use the p-value from the Ljung-Box test or a normal probability plot to check if the residuals are normally distributed.

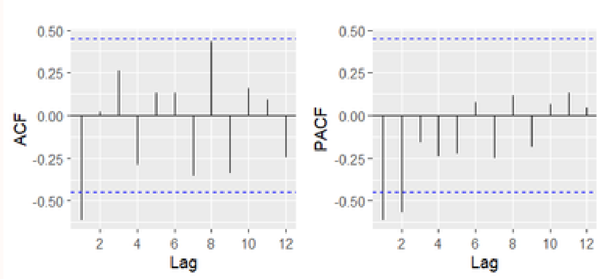
4

- Use the selected model to forecast the time series.
- Verify forecasting accuracy by comparing forecasted values with actual values.
- After validating the model, refit it using the entire dataset.

RESULT & DISCUSSION



Utilizing historical MRSA case data collected by the National Institute of Health from February 2020 to March 2020, a time series plot was generated, comprising 26 observations. The training set includes monthly cases from February 2020 to October 2021, while the testing set spans from November 2019 to March 2022. The plot illustrates that MRSA infections in Malaysia exhibited a non-stationary trend and demonstrated an upward trajectory over the years.

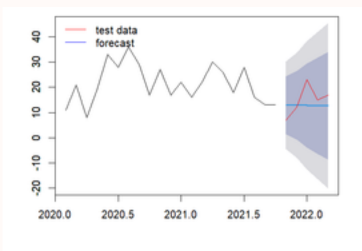


We applied first-order differencing as the initial data was non-stationary, yet the data remained non-stationary. Consequently, we proceeded with second-order differencing. Upon reviewing the ACF and PACF plots for the second-order differencing, we observed no decay, indicating stationarity. Notably, the partial autocorrelation function displayed significant spikes at lags 1 and 2, suggesting a potential AR model order of 1 or 2. In contrast, the autocorrelation function exhibited a significant spike at lag 1, implying a possible MA model order of 1.

MODEL	AIC	AICc	BIC
1,2,1	143.63	145.23	146.46
0,2,1	144.69	145.44	146.58
1,2,0	151.16	151.91	153.05
2,2,1	145.05	147.9	148.82
0,2,0	159.07	159.3	160.01

These are the following 5 models that have been estimated from the ACF and PACF plots. Based on the table, it is observed that ARIMA (1, 2, 1) shows the minimum values of these measures.

It is observed that the number of MRSA cases in Malaysia from October 2021 to March are within the 80% and 95% confidence interval, which shows that the forecasting values are acceptable. TheME and MSE of the forecast values are -1.00 and 0.62, respectively, which arerelatively low. In general, the forecastresults are acceptable.



Month	forecast	actual	Lo 80	Hi 80	Lo 95	Hi 95
Nov-21	12.9	7	1.53	24.29	-4.49	30.31
Dec-21	12.9	12	0.74	26.46	-7.95	33.66
Jan-22	12.8	23	-3.83	29.41	-12.62	38.21
Feb-22	12.7	15	-6.28	31.74	-16.35	41.81
Mar-22	12.7	17	-8.75	34.09	-20.09	45.42

CONCLUSION

Data Limitations:

Short Data Length:

- With only 26 data points, capturing long-term patterns and trends becomes challenging, which may impact the model's ability to make accurate predictions.

Potential Seasonal Patterns:

- The short length of the dataset may obscure potential seasonal patterns that could be better understood with a longer dataset.

Model Performance:

Accuracy Metrics:

- The low ME and MASE values suggest the model performs reasonably well regarding scaled error. Yet, the large deviations in forecast values indicate that further improvements are necessary.

Future Improvements:

Data Enhancement:

- Increasing the dataset length by including more months or incorporating relevant external data (e.g., hospital policies and seasonal trends) can improve model accuracy.

Model Complexity:

- Exploring more sophisticated models such as SARIMA, Prophet, or neural networks could potentially yield better forecasting results, especially if there are underlying complex patterns not captured by simpler models.

REFERENCE:

Symmetric Teleparallel Geometries and its Applications to Cosmology

Kok Ying Hao*

Institute of Mathematical Sciences, Universiti Malaya, Malaysia

yinghao3301@gmail.com

UNIVERSITI MALAYA

1. Abstract and Objectives

In an attempt to obtain exact solutions to Einstein's field equations, we derive the most general metric and affine connection that respect various spacetime symmetries. Furthermore, we consider the case of symmetric teleparallel geometries by imposing vanishing torsion and curvature. Lastly, we obtained the general modified Friedmann equations for various $f(\mathbb{Q})$ gravity theories with the FLRW metric and a perfect fluid universe.

Objectives

- To determine the most general metric and connections of spacetimes with certain symmetries.
- To determine a special case where the torsion and non-metricity tensor vanishes (known as symmetric teleparallel geometries)
- To obtain the modified Friedmann equations in symmetric teleparallel FLRW universes using the $f(\mathbb{Q})$ field equations.

2. Background

Einstein's General Relativity (GR), published in 1915, is the current best description of gravity. However, GR is not the full story of gravity.

Notable issues include:

- What causes the expansion of the universe?
- Does dark matter and dark energy exist?

Potential modifications of GR include:

- $f(\mathbb{Q})$ gravity
- Loop quantum gravity and string theory
- Conformal field theories

3. Tensor Calculus

The metric tensor $g_{\mu\nu}$ are the components of the distance formula:

$$ds^2 = g_{\mu\nu} dx^\mu dx^\nu.$$

An affine connection $\Gamma^\lambda{}_{\mu\nu}$ satisfies the transformation rule:

$$\Gamma^\lambda{}_{\mu\nu} = \frac{\partial x^\lambda}{\partial \bar{x}^k}\left(\frac{\partial^2 \bar{x}^k}{\partial x^\nu \partial x^\mu} + \bar{\Gamma}^k{}_{ij}\frac{\partial \bar{x}^i}{\partial x^\mu}\frac{\partial \bar{x}^j}{\partial x^\nu}\right).$$

The torsion, non-metricity and Riemann curvature tensor are defined respectively as follows:

$$\begin{aligned} T^\lambda{}_{\mu\nu} &= \Gamma^\lambda{}_{\mu\nu} - \Gamma^\lambda{}_{\nu\mu} \\ Q_{\mu\nu\rho} &= \partial_\mu g_{\nu\rho} - \Gamma^\lambda{}_{\mu\nu} g_{\lambda\rho} - \Gamma^\lambda{}_{\mu\rho} g_{\nu\lambda} \\ R^\lambda{}_{\mu\alpha\nu} &= \partial_\alpha \Gamma^\lambda{}_{\mu\nu} - \partial_\nu \Gamma^\lambda{}_{\mu\alpha} + \Gamma^\lambda{}_{\sigma\alpha}\Gamma^\sigma{}_{\mu\nu} - \Gamma^\lambda{}_{\sigma\nu}\Gamma^\sigma{}_{\mu\alpha}. \end{aligned}$$

4. Symmetric Teleparallel Geometries

Symmetric Teleparallel Geometries refer to the class of spacetimes with zero torsion and curvature, which can be found by setting $T^\lambda{}_{\mu\nu} = R^\lambda{}_{\mu\alpha\nu} = 0$.

To determine the most general metric and connection under a given set of symmetries, we set the Lie derivative of the metric and connection to be 0 for each Killing vectors X associated with the spacetime symmetries: $0 = (\mathcal{L}_X g)_{\mu\nu} = (\mathcal{L}_X \Gamma)^\lambda{}_{\mu\nu}$.

For a spacetime with spherical symmetry, the most general symmetric teleparallel metric is:

$$g_{\mu\nu} = \begin{bmatrix} g_{tt} & g_{tr} & 0 & 0 \\ g_{rt} & g_{rr} & 0 & 0 \\ 0 & 0 & g_{\theta\theta} & 0 \\ 0 & 0 & 0 & g_{\theta\theta}\sin^2\theta \end{bmatrix}$$

while the most general symmetric teleparallel connection is in terms of 12 functions of t and r:

$$\Gamma^t{}_{\mu\nu} = \begin{bmatrix} C_1 & C_2 & 0 & 0 \\ C_2 & C_3 & 0 & 0 \\ 0 & 0 & C_7 & 0 \\ 0 & 0 & 0 & C_7\sin^2\theta \end{bmatrix} \quad \Gamma^\theta{}_{\mu\nu} = \begin{bmatrix} 0 & 0 & C_9 & -C_{11}\sin\theta \\ 0 & 0 & C_{10} & -C_{12}\sin\theta \\ C_9 & C_{10} & 0 & 0 \\ -C_{11}\sin\theta & -C_{12}\sin\theta & 0 & -\sin\theta\cos\theta \end{bmatrix}$$

$$\Gamma^r{}_{\mu\nu} = \begin{bmatrix} C_4 & C_5 & 0 & 0 \\ C_5 & C_6 & 0 & 0 \\ 0 & 0 & C_8 & 0 \\ 0 & 0 & 0 & C_8\sin^2\theta \end{bmatrix} \quad \Gamma^\phi{}_{\mu\nu} = \begin{bmatrix} 0 & 0 & C_{11}\csc\theta & C_9 \\ 0 & 0 & C_{12}\csc\theta & C_{10} \\ C_{11}\csc\theta & C_{12}\csc\theta & 0 & \cot\theta \\ C_9 & C_{10} & \cot\theta & 0 \end{bmatrix}.$$

For spacetimes with spherical and translational symmetry, we obtained the FLRW metric for scale factor $A(t)$ and lapse function $N(t)$:

$$ds^2 = -N^2(t)\,dt^2 + \frac{A^2(t)}{1-kr^2}\,dr^2 + A^2(t)\,r^2\,d\theta^2 + A^2(t)\,r^2\sin^2\theta\,d\phi^2.$$

For spacetimes with spherical, translational and time symmetry, we obtained the since disproven Einstein's static universe metric.

$$ds^2 = g_1 dt^2 + g_2\left(\frac{dr}{1-kr^2} + r^2 d\theta^2 + r^2\sin^2\theta d\phi^2\right).$$

5. Applications to Cosmology

$f(\mathbb{Q})$ gravity is a modified theory of gravity using the non-metricity scalar $\mathbb{Q}$ and an aribtrary function $f(\mathbb{Q})$, while requiring vanishing torsion and curvature.

Assume a perfect fluid universe with mass density ρ and pressure p with the FLRW metric,

- $f(\mathbb{Q})$ field equations (with $8\pi G = 1$ and $c = 1$):

$$\frac{df}{d\mathbb{Q}}G_{\mu\nu} + \frac{1}{2}g_{\mu\nu}\left(\mathbb{Q}\frac{df}{d\mathbb{Q}} - f\right) + 2\frac{d^2 f}{d\mathbb{Q}^2}(\nabla_\alpha \mathbb{Q})P^\alpha_{\mu\nu} = T_{\mu\nu}.$$

- Non-metricity scalar $\mathbb{Q}$

$$\mathbb{Q} = Q_{\alpha\mu\nu}P^{\alpha\mu\nu} = -6H^2 + 9HK_3 + 3K_1K_3 - 3K_3^2 + \frac{3HK_2}{A^2} - \frac{3K_1K_2}{A^2} - \frac{3K_2K_3}{A^2}$$

  where K_1, K_2, K_3 are functions of time, $H = \frac{\dot{A}(t)}{A(t)}$ is the Hubble parameter and $P^\alpha_{\mu\nu}$ is the non-metricity conjugate tensor.
- There are 4 possible cases, depending on the value of k. One such case is $k = 0$:

$$\begin{aligned} \mathbb{Q} &= -6H^2 \\ \rho &= 3H^2\frac{df}{d\mathbb{Q}} - \frac{\mathbb{Q}}{2}\frac{df}{d\mathbb{Q}} + \frac{1}{2}f \\ p &= -3H^2\frac{df}{d\mathbb{Q}} - 2H\frac{d^2 f}{d\mathbb{Q}^2}\frac{d\mathbb{Q}}{dt} + \frac{\mathbb{Q}}{2}\frac{df}{d\mathbb{Q}} - \frac{1}{2}f - 2\frac{df}{d\mathbb{Q}}\frac{dH}{dt}. \end{aligned}$$

6. Conclusions

As a conclusion, we explored the history of Einstein's theory of General Relativity and its modifications.

To find solutions to Einstein's field equations, we determined the most general symmetric teleparallel metric and connection satisfying some desired symmetries, deriving the FLRW metric as a consequence.

Lastly, using the $f(\mathbb{Q})$ field equations, we obtained some modified Friedmann equations in the case of symmetric teleparallel FLRW universes.

7. References

[1] Manuel Hohmann. Metric-affine Geometries with Spherical Symmetry. *Symmetry*, 12(3), 2020.

[2] Lavinia Heisenberg. Review on $f(\mathbb{Q})$ Gravity. *Physics Reports*, 1066:1–78, 2024.

[3] Sean M. Carroll. *Spacetime and Geometry: An Introduction to General Relativity*. Cambridge University Press, 2023.

UNIVERSITI MALAYA

SPATIAL ANALYSIS OF HOUSING PRICES USING R: INSIGHTS FROM CASE STUDIES

KISHAANNTH BALA CHANDRAN U2103226/1
SUPERVISED BY DR. DHARINI A/P PATHMANATHAN

ABSTRACT

The spatial variability of housing market determinants leads to differences in market activity and in real estate prices and values across different areas. This study analyses housing prices through spatial analysis in two different case studies: the suburbs of Perth and the towns of Boston, aiming to determine the existence of spatial autocorrelation. The Perth dataset is analysed using both non-spatial and spatial models, specifically the ordinary least squares (OLS) regression and the Manski model, respectively. The low rank spatial lag model (LSLM) with the selected spatial variables from the dataset was fitted to the Boston dataset.. The results from the first case study, Perth, showed a weak spatial autocorrelation and the data was proven to be a poor fit for the Manski model. However, the second case study, Boston house price data, showed a moderately strong spatial autocorrelation, and the low rank spatial lag model proved to be a good fit for the data. Several spatial variables were identified to have a direct impact on housing prices in the region.

INTRODUCTION

Spatial analysis has been used to study, understand and derive the insights from the spatial data either geometrically or geographically.One of the crucial uses of spatial analysis is to predict the housing prices in a region. In earlier times, the method for analysing the data through a perceived spatial lens is by modelling the data onto the ordinary least squares (OLS) regression. As years went by, spatial models such as the Manski model and the low rank spatial lag model were developed. This study aims to conduct a spatial analysis on housing prices with two different case studies, which are the suburbs of Perth and the towns of Boston. This study also aims to compare the two case studies to determine the existence of spatial autocorrelation in the housing prices and spatial features which could be the factors that affect the housing prices in Perth and Boston.

DATA DESCRIPTION

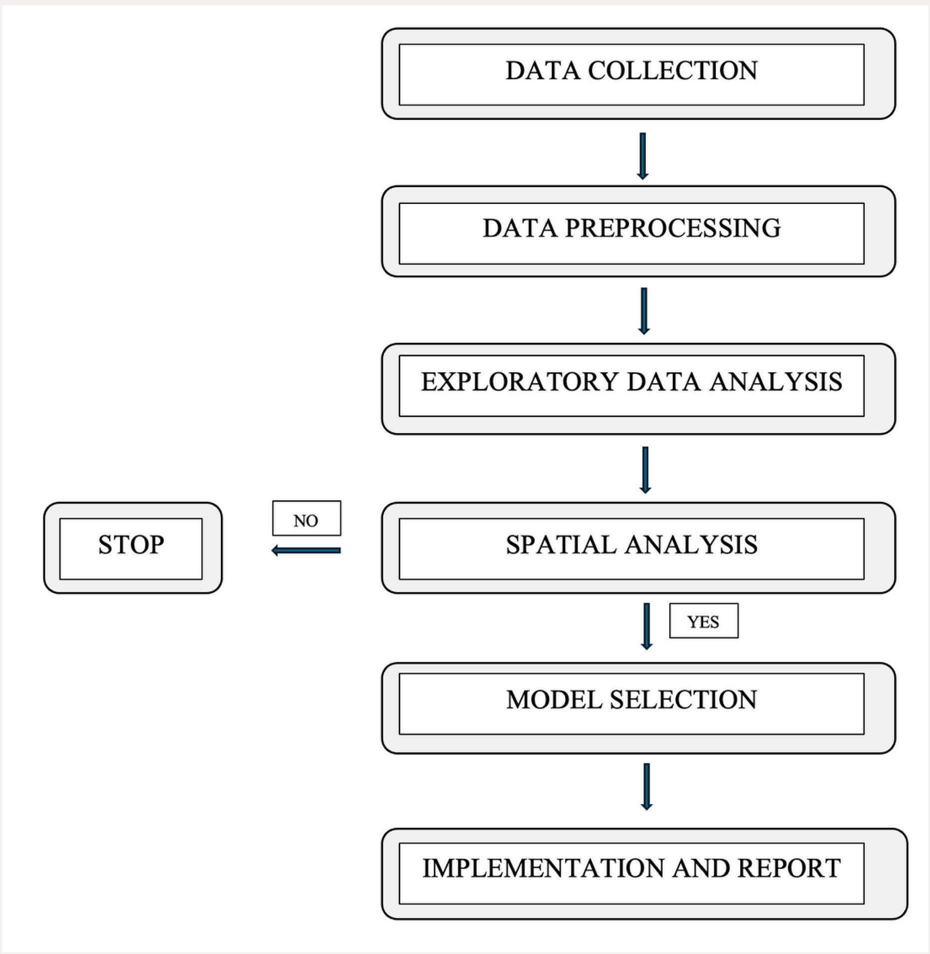
Perth

- Housing price dataset obtained from 2021
- Consists of 19 columns

Boston

- Housing price dataset obtained from the 1978 census
- Well known dataset for conducting spatial analysis
- Consists of 14 columns

METHODOLOGY



Workflow of the study

- Moran's I Statistic

$$I = \frac{N}{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}} \times \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(x_i-\mu)(x_j-\mu)}{\sum_{i=1}^{n}(x_i-\mu)^2}, \text{for } i \neq j,$$

  - Estimated based on covariance
  - Index values ranges from -1 to 1
  - Values above 0 give a positive spatial autocorrelation
  - Values below 0 give a negative spatial autocorrelation

- Ordinary Least Squares (OLS) Regression
  - Provides the impact of the spatial variables to the dependent variable
  - Calculates adjusted R-squared value which gives the percentage of variability in the data due to the selected spatial variables

- The Manski Model

$$Y = \rho WY + a_{l_N} + X\beta + WX\theta + u,$$

$$u = \lambda Wu + \varepsilon,$$

- The Low Rank Spatial Lag Model

$$y_i = \beta_0 + \epsilon_i + z_i,$$

$$\epsilon_i \sim N(0,\sigma^2) z_i = \rho \sum_{i \neq j}^{N} w_{i,j} z_j + \sum_{k \neq 1}^{K} x_{i,k}\beta_k + u_i,$$

Spatial Models

- To calculate the impact of the spatial variables onto the dependent variable
- Provides the direct, indirect and total impact onto the dependent variable

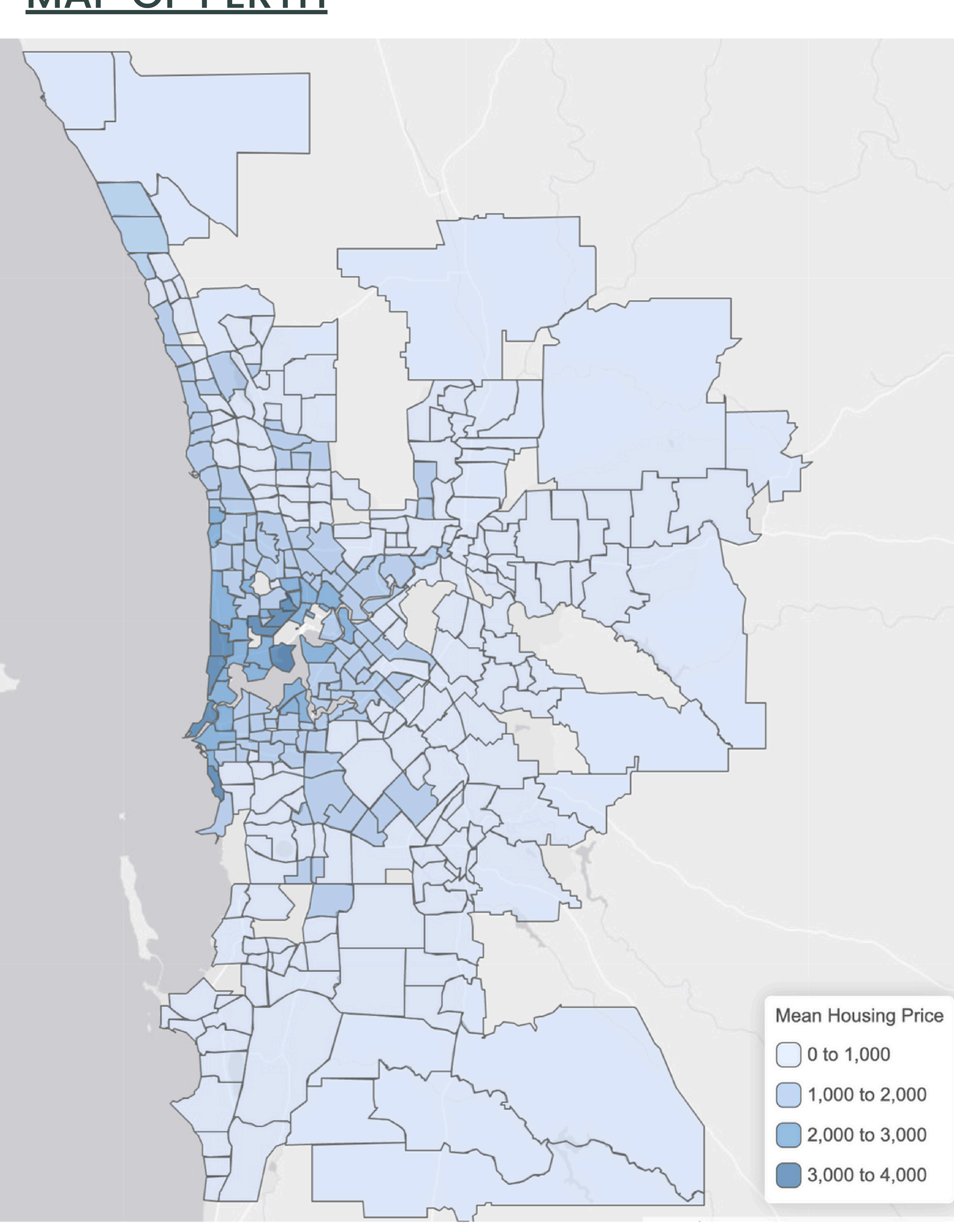
The spatial weights matrices generated for these models are the Queen contiguity weights which are a combination of Rook contiguity (only common vertices) and Bishop contiguity (only comon vertices).

$$w_{ij} = \begin{cases} 1 & \ell_{ij} > 0 \\ 0 & \ell_{ij} = 0 \end{cases},$$

APPLICATION OF DATA TO CASE STUDIES

1) PERTH

MAP OF PERTH



OLS REGRESSION

Predictors	Estimates	Confidence Interval	p-value
(Intercept)	1833.08	1812.86 – 1853.30	**<0.001**
LAND AREA	-0.00	0.00 – 0.00	**<0.001**
NEAREST STN DIST	-0.04	-0.04 – -0.04	**<0.001**
CBD DIST	-0.03	-0.03 – -0.03	**<0.001**
NEAREST SCH DIST	8.01	0.31 – 15.70	**0.041**

- LAND AREA shows a negligible negative impact
- CBD DIST & NEAREST STN DIST shows a small negative impact
- NEAREST SCH DIST shows a significant positive impact
- Adjusted R-squared value is 0.241

THE MANSKI MODEL

	DIRECT	INDIRECT	TOTAL
CBD_DIST	0.018167695	-0.05751353	-0.03934583
NEAREST_STN_DIST	-0.016319871	-0.07852185	-0.09484173
LAND_AREA	-0.007572612	-0.02167637	-0.02924898
NEAREST_SCH_DIST	-6.113631294	358.42221512	352.30858383

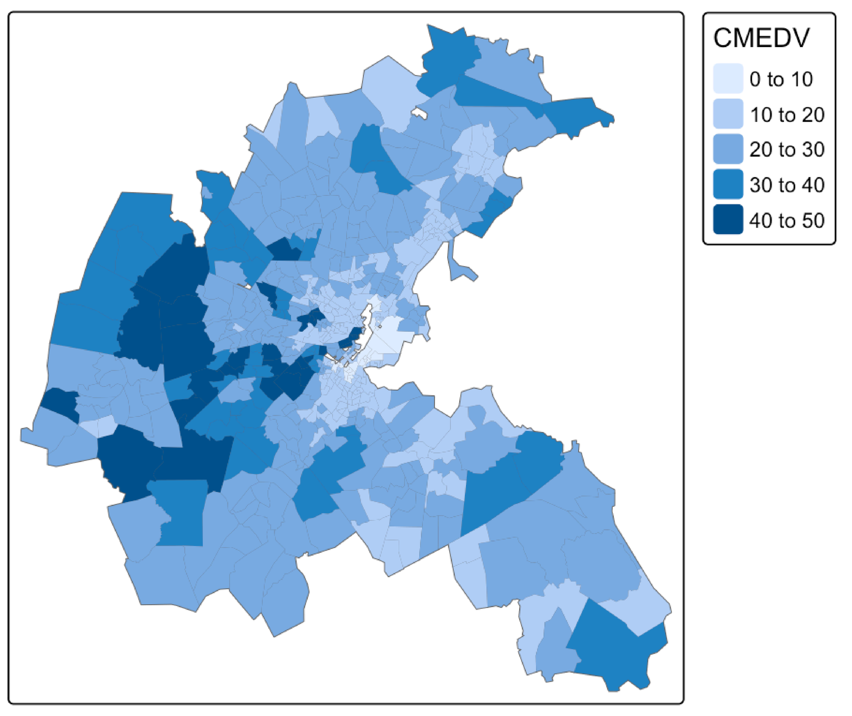
- CBD_DIST, NEAREST_STN_DIST & LAND_AREA show a negative and low effect on the price per square meter
- NEAREST_SCH_DIST shows a significant positive effect on the price per square meter
- The Manski model proves to be a poor fit for the data

2) BOSTON

LOW RANK SPATIAL LAG MODEL

- The crime rate (CRIM) and proportion of owner-occupied buildings prior to 1940 (AGE) significant negative effect towards the median value prices
- The Charles River dummy variable (CHAS) and the rooms per property (RM) show a significant positive effect towards the median value prices
- The nitrous oxide concentrations (NOX) and the proportion of non-retail business acres (INDUS) show a lack of effect on the median value prices

MAP OF BOSTON



Variable	Estimate	Standard Error (SE)	t-value	p-value
(Intercept)	-14.6152	2.7742	-5.2682	$2.0972e^{-7}$
CRIM	-0.1073	0.0284	-3.7711	$1.8315e^{-4}$
CHAS	1.8942	0.8897	2.1290	0.0338
RM	6.9288	0.3320	20.8720	$0.0000e^{0}$
AGE	-0.0414	0.0121	-3.4088	$7.0827e^{-4}$
INDUS	-0.1012	0.0604	-1.6749	0.0946
NOX	-5.1642	3.8309	-1.3481	0.1783

CONCLUSION

The Boston house price data, which is used as benchmark data in this work, is a classic spatial data set and is comprehensive. Hence, suitable spatial models can be employed to proceed with spatial predictive modelling. However, the Perth house price data needs to be addressed further by implementing more spatial variables to proceed with spatial modelling. Future studies involving more variables that project strong spatial autocorrelation can be investigated to improve the existing spatial predictive model accuracy. It is also of interest to study the effects of various spatial weight matrices to quantify the spatial relationships that exist among the features in the dataset of interest.

SELECTED REFERENCES

Bivand, R. (2022). R packages for analyzing spatial data: A comparative case study with areal data. Geographical Analysis, 54(3), 488–518. Retrieved from 10.1111/gean.12319

Des Rosiers, F., Theriault, M. and Villineuve, P., (2000). Sorting out Access and Neighbourhood Factors in Hedonic Price Modelling. Journal of Property Investment & Finance, 18(3): 291–315. Retrieved from 10.1108/14635780010338245

Pebesma, E., & Bivand, R. (2023). Spatial Data Science: With Applications in R. Chapman and Hall/CRC. Retrieved from 10.1201/9780429459016.

FAKULTI SAINS
Faculty of Science

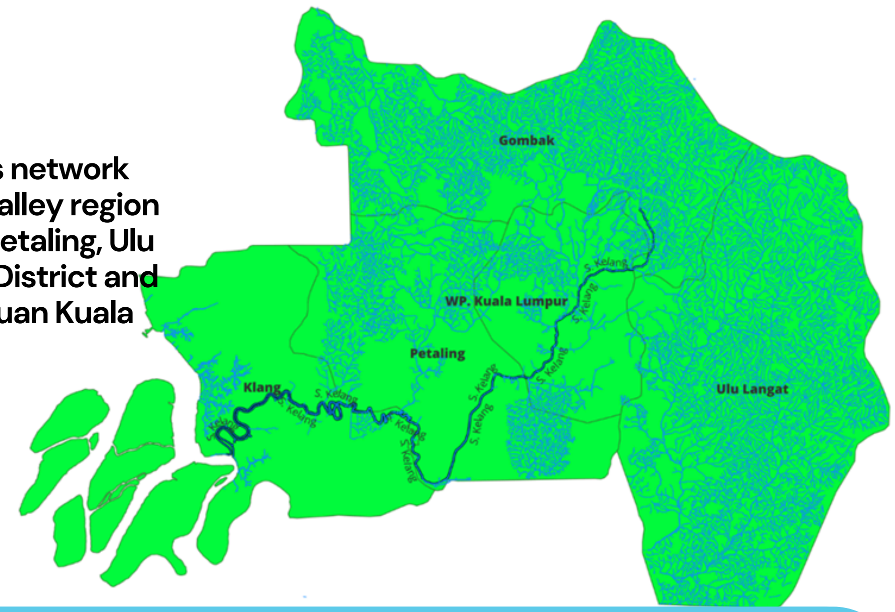
Name: Ikhwan bin Mohd Nizam
Matric No.: U2001245
Supervisor: Dr. Shahizat bin Amir & Dr. Mohd. Zahurin bin Mohamed Kamali

for SIM3021: Mathematical Science Project

Fractal Dimension of Klang River Network: A Box-counting Method

Study Area

Klang River and its network system in Klang Valley region consist of Klang, Petaling, Ulu Langat, Gombak District and Wilayah Persekutuan Kuala Lumpur



Introduction

- Fractal dimension was coined by Benoit B. Mandelbrot in 1975.
- The dimension is not integer, but rather a fraction unlike Euclidean dimensions.
- Fractals are geometry figures & their characteristics:
  - Self similar
  - Recursive
- A fractal dimension is a ratio providing a statistical index of complexity.

Problem Statement:

The Klang River network in the Klang Valley region lacks a comprehensive assessment of its fractal properties. Understanding its fractal dimension is crucial for insights into its structure, evolution, and behaviour. This study aims to evaluate the strengths and limitations of using the box-counting method integrated with GIS to estimate the fractal dimension, providing valuable insights for environmental management, flood control, and urban planning.

Objectives:

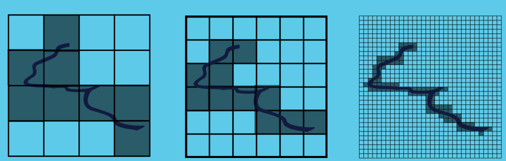
- Understanding the fractal complexity in Klang River network in Klang Valley area using spatial analysis and mathematical computation.
- Investigate the applicability of Geographic Information System (GIS) with the box counting method for estimating the fractal dimension of Klang River network.

Methodology

$$N(L) \propto L^{-D} \quad (1) \qquad D = \frac{\log N(L)}{\log 1/L} \quad (2)$$

where:

- D represents the fractal dimension.
- N(L) is the number of grid boxes of side length L needed to cover the network.
- L is the grid boxes with side length L.



$N(L1) = 8$,
$N(L2) = 11$,
$N(L3) = 101$

This power law corresponds to a linear relationship of **Equation 1** on a logarithmic scale. A log-log plot of N(L) against 1/L is constructed with the plotted points, and the slope of the resulting line gives an estimate of the fractal dimension, D by fitting these points with linear regression using **Equation 2**.

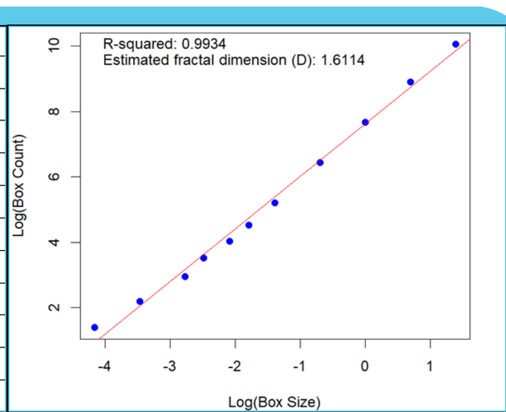
QGIS + R python™

GIS & Spatial Analysis **Mathematical Computation**

Obtain high resolution spatial map of river network from *Pusat Geospatial Negara* → Add the river network shapefile as a layer in QGIS 3.34 → Crop the network within Klang Valley Area only. → Superimposed with grid of square shapes varies in sizes with Vector processing toolbox in QGIS. → Obtain the number of grid boxes needed to cover the network using Python. → Calculate Fractal Dimension using R

Results and Discussion

Grid size, L (km)	N(L)	Log (1/L)	Log (N(L))
0.25	23580	0.6021	4.3725
0.50	7436	0.3010	3.8713
1.00	2148	0	3.3320
2.00	622	-0.3010	2.7938
4.00	181	-0.6021	2.2577
6.00	93	-0.7782	1.9685
8.00	56	-0.9031	1.7482
12.00	34	-1.0792	1.5315
16.00	19	-1.2041	1.2788
32.00	9	-1.5051	0.9542
64.00	4	-1.8062	0.6021



Interpretation of Results:

- The table lists grid sizes and corresponding box counts, showing an inverse relationship.
- The log-log plot of box size and box count indicates a power-law distribution, characteristic of fractals.
- The fractal dimension of 1.6114 suggests a complex river network, impacting hydrological processes like water flow and flood dynamics.

Box Counting Method Using GIS:

- QGIS automates the box-counting process, ensuring accuracy and efficiency.
- This method helps manage large datasets and provides insights into river network complexity.

Urban River Network and Impact of Urbanization:

- Urbanization can simplify river networks, reducing their capacity to manage heavy rainfall, as observed in the Taihu Region, China.
- Understanding the Klang Valley river network's fractal properties is vital for effective water management and flood mitigation, particularly after the severe December 2021 floods.
- Sustainable urban planning should balance development with ecological preservation, using advanced technologies and community engagement.

Conclusion

The box-counting method effectively measures the complexity of the Klang River network, revealing a fractal dimension of 1.6114. This high degree of branching and connectivity is crucial for hydrological and ecological processes. Limited data accessibility highlights the need for more comprehensive studies and institutional support. The project also demonstrates the utility of GIS in enhancing the box-counting method, facilitating efficient and accurate spatial analysis. Further studies on Malaysian river systems are recommended to understand the impacts of urbanization. Policymakers should use GIS technologies for better river conservation, flood management, and urban planning.

Acknowledgement and References

Data:

KEMENTERIAN SUMBER ASLI DAN KELESTARIAN ALAM
PUSAT GEOSPATIAL NEGARA

Scholar: