ISSN 1391 4395 September - December 2022

IASSL NEWSLETTER

Institute of Applied Statistics Sri Lanka
The Professional Center
275/75
Prof. Stanley Wijesundara Mawatha
Colombo 07
Sri Lanka
+94 11 2588291
appstatsl@gmail.com/ editor@iassl.lk
http://www.iassl.lk
http://www.facebook.com/iassl2020/
https://www.linkedin.com/company/iassl/

> "Statistical thinking will one day be as necessary for efficient citizenship as the ability to read and write."
>
> H. G. Wells

Featured Segments

SPATIAL STATISTICS IN A NUTSHELL

"Any observation with a space tag can be considered as a spatial datum. "

LEAST SQUARES: CONNECTING LINEAR ALGEBRA, CALCULUS AND GEOMETRY

"In this brief article, I review the connection between the quasi-solution to a linear system via least squares and its underlying geometry. "

NEWS IN BRIEF

SIMPLEMAN'S MEAN (ONE-ACT-PLAY)

"Simple man has a problem with the concept of the average. He believes that the average is always based on the arithmetic mean."

IDENTIFICATION OF REASONS FOR CULPABLE HOMICIDES AND ATTEMPTED MURDERS

"A Case Study for the Kelaniya Police Division."

ANNOUNCEMENTS

SUDOKU PUZZLE COMPETITION
UPCOMING EVENTS
UPCOMING COURSES

IASSL President's Message

Dr. D.C. Wickramarachchi
President/IASSL

It's the end of year 2022. The right moment to look back and assess our progress. Why not for IASSL? In this brief message I would like to glance through the year 2022 and report the progression of the Institute. Welcome to the final Newsletter of IASSL for the year 2022.

The general objective of IASSL is to promote and assist the advancement of applied statistics by furtherance of research, development, education, training and extension. As the president, I would like to proudly announce that the activities that IASSL carried out throughout the year 2022 were very successful and thereby the IASSL was able to perform as intended. It is also important to note that IASSL's success would not have been possible without the dedication of our strong council including the sub-committees, as well as our members and the hard work of our small, but efficient staff at the office.

Sri Lanka Journal of Applied Statistics (SLJAS) is now up-to-date. A Newsletter is published tri-annually without any lag. The newly designed IASSL website is now fully functioning on a local host. Every event and activity planned/carried out will be/has been extensively publicized before and after completion of each event/activity via social media such as Facebook, LinkedIn, Instagram etc. Many short courses were successfully completed and many more have been planned for the next few months. Diploma and Higher Diploma in Applied Statistics are functioning well and hence getting wider community attraction. Best research award-2021 competition was successfully held and the same for the year 2022 has already been advertised. The webinars that were conducted on various topics gained vast public appreciations. One of the key activities that is worth mentioning is the National Statistics Olympiad that was conducted in December 2022 by the IASSL with the collaboration of Department of Statistics, University of Sri Jayewardenepura. The selected candidates are eligible for the International Statistics Olympiad which will be held in January 2023.

The coming year will be an exciting year for IASSL. In the recently concluded council meeting it was discussed in length as to how IASSL should contribute in large scale to various national projects/surveys. Further, founding steps will be laid in becoming the sole connecting hub of all Sri Lankan Statisticians who are working locally as well as internationally across the globe. Although the triennial International Statistics conference also provides such a platform, an annual event which is less academically oriented would pay the way to gather all Statisticians and Students/novice in Statistics to one place for better networking.

Thank you for your continued dedication and loyalty to IASSL. It is a great honor for me to be leading the IASSL and I look forward to a very successful year ahead with continued efforts to promote and assist the advancement of applied statistics. Let us all march into 2023 with pride of what we have accomplished and the drive to make the next year even more amazing.

I wish you a Merry Christmas and Happy New Year!

Editorial

Dr. Vasana Chandrasekara
Editor/IASSL

It is indeed a great honour to be the Editor of IASSL and it is an immense pleasure to launch this third issue of the newsletter for the year 2022. In this issue, we will recount various events, projects and activities in which IASSL members were actively volved from the 1st of September 2022 until the 31st of December 2022. Basically this issue contains articles from senior academics and industry professionals in the field of Statistics, One-Act-Play article from an emirates professor, Stat Undergrad coloumn with articles from undergraduates, articles from IASSL members, News in brief which cover all events of IASSL during the considered period of this newsletter. As usual, the puzzle completion is included for all readers to relish and win prizes and the winners of the puzzle competition of the last issue are announced in this issue.Finally, the upcoming events of IASSL are listed for your information.

A huge thank you to all the professors, industry professionals, IASSL members and undergraduates who contributed to writing the valuable articles for this issue. Moreover, I appreciate the support extended by the president, secretory, all subcommittee chairpersons and executive council members of IASSL in providing information relating to the events conducted by them during the period September to December 2022.

Last but not least, I would like to thank the editorial board members and especially the associate editor and the editorial assistant of IASSL for their immense support throughout the creation of this issue of the IASSL newsletter.

I invite all readers to submit articles and news to be consider in the next issue of the IASSL newsletter (editor@iassl.lk) and hope you all will enjoy reading this issue.

CONTACT INFORMATION

Institute of Applied Statistics Sri Lanka
The Professional Center
275/75
Prof. Stanley Wijesundara Mawatha
Colombo07
Sri Lanka

☎ +94 11 2588291

facebook.com/iassl2020

linkedin.com/company/iassl/

@ appstatsl@gmail.com/
editor@iassl.lk

ONE-ACT-PLAY:
Simpleman's Mean

Professor Emeritus R.O. Thattil
Founder President
Applied Statistics Association of Sri Lanka (ASASL)

The following dialogue is between a Statistician (SS) and a novice to Statistics - Simple man (SM). Simple man has a problem with the concept of the average. He believes that the average is always based on the arithmetic mean.

SM : I can't understand why people are talking about different types of means, when the arithmetic mean can be used under all circumstances.

SS : The arithmetic mean is useful only for problems involving linear behavior. When it comes to population growth and speed problems which do not possess linear behavior, the arithmetic mean does not measure the average.

SM : Why not?

SS : I will ask you a simple question regarding population growth. Consider the following example, which appeared in the Scientific American journal. A microbe is placed in a jar at 3.00 p.m. The microbe divides into 2 every minute. The new microbes also divides into 2 in the next minute and so on until the jar is full of microbes at 4.00 p.m. When was the jar became half full of microbes?

SM : Obviously at 3.30 p.m.!

SS : This is the answer given by many who have no idea about non-linear growth.

SM : So, what is the actual answer?

SS : One minute before 4.00 p.m., since in the next minute the half becomes full.

SM : My God! I didn’t realize it.

SS : People who think only about linear behavior have no idea that the mean for population growth is based on the geometric mean. Here is another question that brings in the concept of the geometric mean. A population of a small village was 10 in 1970, while it was 40 in 1990. What was the population in 1980 assuming a constant rate of population growth?

SM : That is easy. In 1980 the population would be the average of 10 and 40 which is 25.

SS : You have again having the arithmetic mean in your head. The actual answer is the geometric mean which is the square root of the product 10 x 40, which works out $\sqrt{10 \text{ x } 40} = \sqrt{400}$ = 20. Therefore, in 1980 the population was 20 and not 25. In general the geometric mean of a set of n numbers = n th root of the product of then numbers. For 2 numbers it is therefore the square root of the product, for 3 numbers it is the cube root of the product of the 3 numbers, and so on.

SM : Are there other situations where a different type of mean is used?

SS : Of course! I will frame another question for you. The distance between 2 villages is 60 kms. A person driving at an average speed of 60 km per hour will take only 1 hour to cover the distance. Suppose he does 30km per hour in the first 30 minutes. How fast should he go in the second half to complete the distance in 1 hour.

SM : Definitely at 90 km per hour.

SS : You are again thinking of the arithmetic mean. The answer is he will never complete the distance in 1 hour, since he has already finished 1 hour to cover half the distance at 30 km per hour!

SM : What kind of mean can be used for speed problems?

SS : It is the harmonic mean. I will this time pose a similar question. A person does half the distance at 60 km per hour and the second half at 30 km per hour. What is the average speed over the entire distance?

SM : I think it is the average of 60 and 30, which will be 45 km per hour.

SS : Wrong again. This time you have to use the harmonic mean (H). The harmonic mean is given by the formula

$1/H = (1/X_1 + 1/X_2) / 2$ where X_1 and X_2 are the 2 speeds.

$\therefore 1/H = (1/30 + 1/60) / 2, = (\frac{2+1}{60})/2 ,= (3/60) / 2$

$= 3/120 , =1/40$

$\therefore H = 40$

Thus, the average speed is 40 km per hour and not 45 km per hour.The harmonic mean is used as the average for speed problems.

SM : I now understand the use of different types of means.

SS : This is not the end of the story. There are also the median as the measure of the average, which is used for small data sets, which contain outliers. I should not burden you with ideas of trimmed means which is also used for data sets with outliers. I hope you can see how different means are used under different situations.

SM : Can you give me an idea about trimmed means?

SS : When the data set has outliers, we can leave out the extreme values from both sides of the data set after arranging it in order (ascending or descending order) and then calculate the mean from the remaining data.

SM : Can you give me an example of its use?

SS : In diving competitions, 7 judges give marks. The highest and lowest marks are discarded before the average is computed. This will avoid bias or favouration for a diver.

SM : Thank you Sir. That will be enough for the present. You have really enlightened me.

Spatial Statistics in a Nutshell

Dr. Ravindra Lokupitiya
Senior Lecturer, Department of Statistics,
Faculty of Applied Sciences,
University of Sri Jayewardenepura.

Introduction:

The main purpose of this article is to give a brief overview of the spatial data analysis. The history of the field and available spatial data analysis methods are briefly discussed. Some theoretical aspects and an application in the Sri Lankan context are given. Finally, advances in the field and available statistical software for spatial data analysis are briefly discussed.

Historically, spatial data analysis emerged in the mining industry. The main objective was to estimate the true mean value of ore (a natural rock which contains minerals) in a mining block (area) accurately with a limited number of sampling points. In the early 1950s, D. G. Krige, a South African mining engineer, initially developed an empirical method for that purpose. Later, in 1960s, Georges Matheron, a French mathematician, developed the statistical framework for the same method, which is called "kriging" — named after D. G. Krige.

Loosely speaking, any observation with a space tag can be considered as a spatial datum. Therefore, one or many attributes measured at different locations belong to spatial data. For example, some weather parameters such as rainfall, temperature, or relative humidity measured at different weather stations can be considered as spatial data. These types of data tend to be correlated over the space. For instance, there is greater likelihood of rainfall measured at two locations 1 km apart being similar than that measured at two locations 50 km apart. Incorporating spatial information in the estimation process is called the spatial data analysis. On most occasions, spatial data are analyzed considering only the trend component, which is often modeled by using either linear or non-linear regression approaches as a function of the predictors such as (x, y) coordinates of the location, neglecting the additional information contained over the space. On such occasions, the assumption of independent error terms in regression is also violated since they are spatially correlated.

Data types and analyses:

There are mainly three types of spatial data — spatial point patterns, spatially continuous data, and areal data. In spatial point patterns, the event of interest is the location, not the attributes measured at the location. Locations of a certain invasive plant, birds' nests, or households with a pandemic disease can be considered as examples of spatial point patterns. In spatial point pattern analysis, the focus is to find whether the locations or points show any clustering, or a regular, or random pattern in a given specified region.

Spatially continuous data analysis, also known as geostatistics, is the most commonly used technique. For example, the analysis of weather parameters, pollution levels, and precious mineral concentrations measured at different locations belong to this category. Here the attributes vary continuously between the locations. These attributes could be spatially correlated and their covariance structure can be modeled using various covariance models. When predicting the value of a certain attribute at a given location, these correlations are accounted in addition to the mean structure of the field. The estimated value at a given location is found as a linear combination of the observed values; the weights of the linear combination are determined so that the properties of the estimator such as unbiasedness and minimum variance are satisfied. Hence the estimator is a best linear unbiased estimator (BLUE). This estimation technique is known as "Kriging".

In areal data analysis, attribute values measured are assigned to a certain region or area rather than assigning to a point location. For example, the number of dengue patients in a GN (Grama Niladari) division; here the attribute value is assigned to the whole GN division. One could analyze it using geostatistical methods, assigning attribute values to the centroids of the GN divisions. However, this approach is not valid because the attribute values do not vary continuously between the centroids, which violates the basic assumptions of the spatially continuous data analysis. In areal data analysis, the values in the neighboring regions of a given region are more likely to be similar than those in faraway regions. These associations are defined using a neighborhood matrix (say W). The simplest way to define the elements of the neighborhood matrix is to assign $w_{ij} = 1$, if regions i and j share a common boundary, and $w_{ij} = 0$ otherwise. Once the neighborhood matrix is defined, the modeling is done by specifying the covariance structure indirectly using an interaction scheme as in autoregressive models. There are two types of such autoregressive models used in areal data analysis, namely, simultaneous autoregressive (SAR) models and conditional autoregressive (CAR) models.

Some theoretical aspects:

A brief introduction to the theory of kriging method is given below. There are mainly three types of kriging, namely, simple kriging, ordinary kriging, and universal kriging, which are defined according to the mean structure. If the mean structure is known a-priori, it can be subtracted from the observations and analysis is done based on the remainder, which is called the simple kriging. In ordinary kriging, the mean structure is assumed to be an unknown constant. In universal kriging, a global trend is assumed instead of a constant mean.

Due to the limited space, only universal kriging procedure is briefly discussed here. Consider the model, $\underline{Z} = X\underline{\beta} + \underline{\gamma} + \underline{\varepsilon}$, where $\underline{Z} = \left[z(\underline{s}_1), z(\underline{s}_2), \dots, z(\underline{s}_n)\right]^T$ is the observation vector; $X\underline{\beta}$ is the trend component, where X is the matrix of the predictors; $\underline{\gamma} \sim N(\underline{0}, C)$ is the spatial component, where C is the spatial covariance matrix; and $\underline{\varepsilon} \sim N(\underline{0}, \sigma_\epsilon^2 I)$ is the usual random error term in the regression model. The most challenging part in kriging is to estimate the matrix C, where $(i,j)^{th}$element is the $Cov(z(\underline{s}_i), z(\underline{s}_j))$. When the covariance structure is modeled, the following assumptions are made regarding the spatial random process: the covariance between any two locations depends on the vector distance between the locations and is independent of their absolute locations, which is called the *weak or second-order stationarity*. In addition to this, it is assumed that the covariance depends solely on the distance between locations, i.e. it does not depend on the direction, which is called the *isotropy*.

Hence, if the process is *stationary*,

$$Cov\left(z(\underline{s}_i), z(\underline{s}_j)\right) = C(\underline{h}), \text{where } \underline{h} = \underline{s}_i - \underline{s}_j,$$

and if the process is *isotropic*,

$$Cov\left(z(\underline{s}_i), z(\underline{s}_j)\right) = C(h), \text{where } h = |\underline{h}|.$$

Usually, multiple observations are not available at the locations to estimate the spatial covariance structure and hence a technique called variogram approach is used. The variogram is defined as

$$2\gamma(\underline{h}) = Var[Z(\underline{s}_i) - Z(\underline{s}_j)].$$

$\gamma(\underline{h})$ is called the semi-variogram, which is half the variogram. Most geostatistical software use semi-variogram in their analysis and the word "variogram" is synonymously used for semi-variogram. The natural estimator for semi-variogram based on the method of moments is

$$\hat{\gamma}(\underline{h}) = \frac{1}{2|N(\underline{h})|}\sum_{N(\underline{h})}\left(Z(\underline{s}_i) - Z(\underline{s}_j)\right)^2, \text{where } N(\underline{h}) = \{(\underline{s}_i, \underline{s}_j)|\underline{s}_i - \underline{s}_j = \underline{h}; i, j = 1, 2, \dots, n\}.$$

And $|N(h)|$ is number of distinct pairs.

Semi-variogram is estimated under the assumption of isotropy; i.e. it depends solely on the distance between the locations and independent on the direction. However, generally, spatial random processes are anistropic (i.e. they depend on the direction), which could be detected by the directional variograms; i.e. by estimating variograms separately for selected directions. A common choice is the four directions, north, north-east, east, and south-east. If the varigram depends on the direction, then spatial random process is considered as anisotropic. In such cases, transformation has to be done to make it isotropic.

There are three parameters, which explain a structure of a variogram, — namely, nugget effect, range, and sill. *Nugget effect* represents the micro-scale variations or measurement errors at the sampling locations. As the separation between the locations increases, the corresponding variogram value also generally increases; eventually, it reaches a plateau; the distance at which the variogram reaches this plateau is called the *range*. The plateau that the variogram reaches at the range is called the *sill*. Figure 1 depicts a semi-variogram along with these parameters, where solid dots represent the sample variogram and the solid curve represents a fitted variogram model.

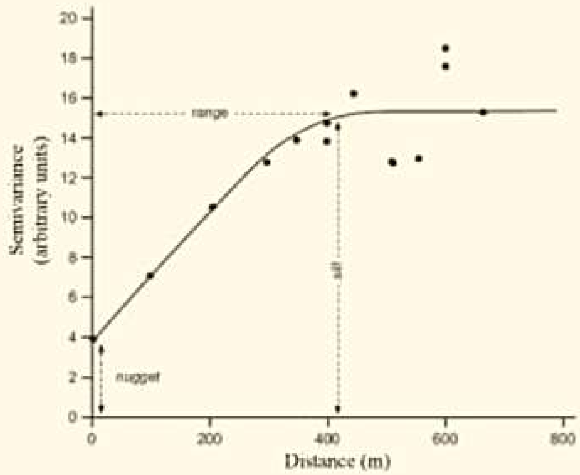


Figure 1. Semi-variogram with corresponding parameters (source: accessed 11 October 2022, <https://www.aspexit.com/variogram-and-spatial-autocorrelation>)

Once the sample variogram is estimated, a parametric model, which explains the covariance structure with a smooth continuous curve, will be fitted. There are many variogram models and out of them spherical, exponential, and Gaussian models are the most commonly used ones. The best fitted model is selected based on the Sum of Square Errors (SSE) and it corresponds to the minimum SSE. For example, the exponential model for an isotropic field is given by,

$$\gamma(h) = \begin{cases} 0 & , \quad h = 0 \\ c_0 + \sigma^2 \left[1 - e^{-\frac{h}{r}}\right], & \quad h \neq 0, \end{cases}$$

where c_0, $c_0 + \sigma^2$, and r correspond to the nugget effect, sill, and range of the semi-variogram, respectively.

The covariance function can be found using the relationship between the semi-variogram and covariance function,

$$\gamma(h) = \sigma^2 - C(h), \text{where } \sigma^2 = C(0)$$

Once the covariance function is specified, the predicted value at a given location $\underline{s}_0$ can be found as a linear combination of the observations given by,

$$\hat{Z}(\underline{s}_0) = \sum_{i=1}^{n} w_i Z(\underline{s}_i),$$

where the weights, w_i's, are determined such that the estimator is unbiased and has the minimum variance.

Applicability in Sri Lankan context:

As an example of the applicability of kriging in Sri Lanka, monthly totals of rainfall collected at 100 weather stations in May, 2011 have been considered. The analysis has been done in the log scale due to high variability in the rainfall measurements. The trend component was modeled using a linear regression model considering latitudes and longitudes as predictors.

Figure 2 shows the estimated variogram along with the fitted model. Here the exponential model, which has the minimum SSE, is fitted. It is assumed that the spatial random field is isotropic.

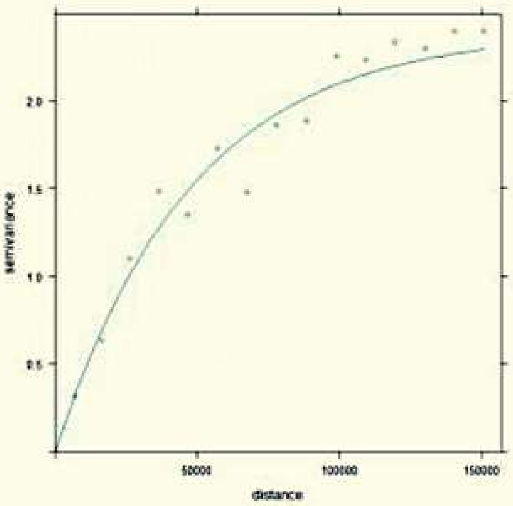


Figure 2. Semi-variogram (open circles) and fitted exponential model (solid line) for the rainfall in log scale.

Figure 3 shows the rainfall estimated using the universal kriging method. The left panel of the figure shows the predicted rainfall in log scale; the cold (blue) colors indicate the high precipitation areas. The right panel of the figure shows the corresponding (kriging) errors in estimation. According to the figure, kriging errors are minimum at the sampling locations (indicated by "+" sign) and surrounding areas (yellow color regions), whereas they are fairly large in sparsely observed (blue color) regions.

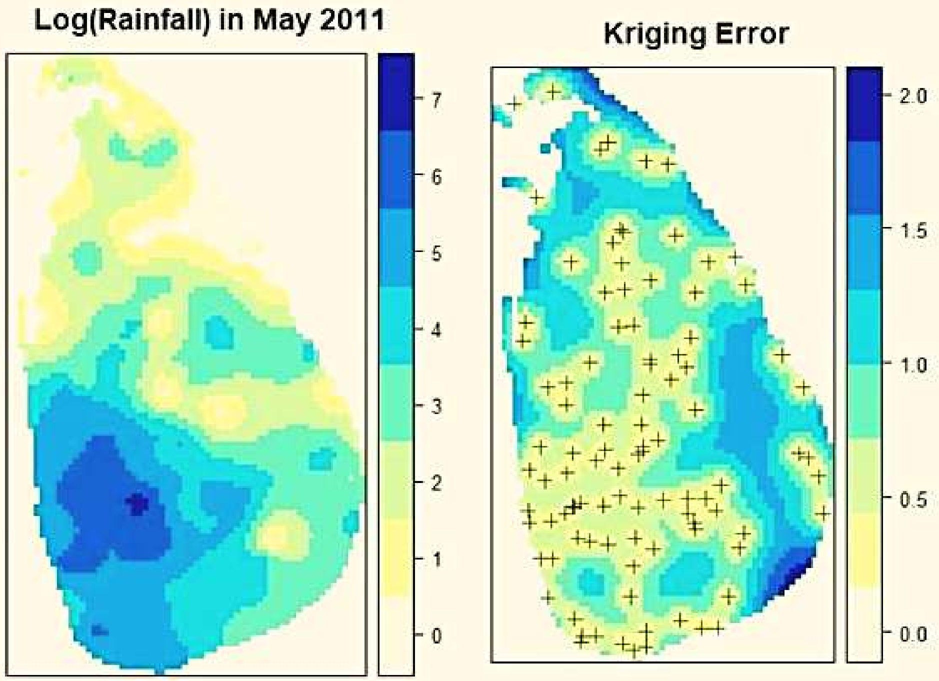


Figure 3. Predicted rainfall in log scale (left panel) along with the kriging error (right panel).

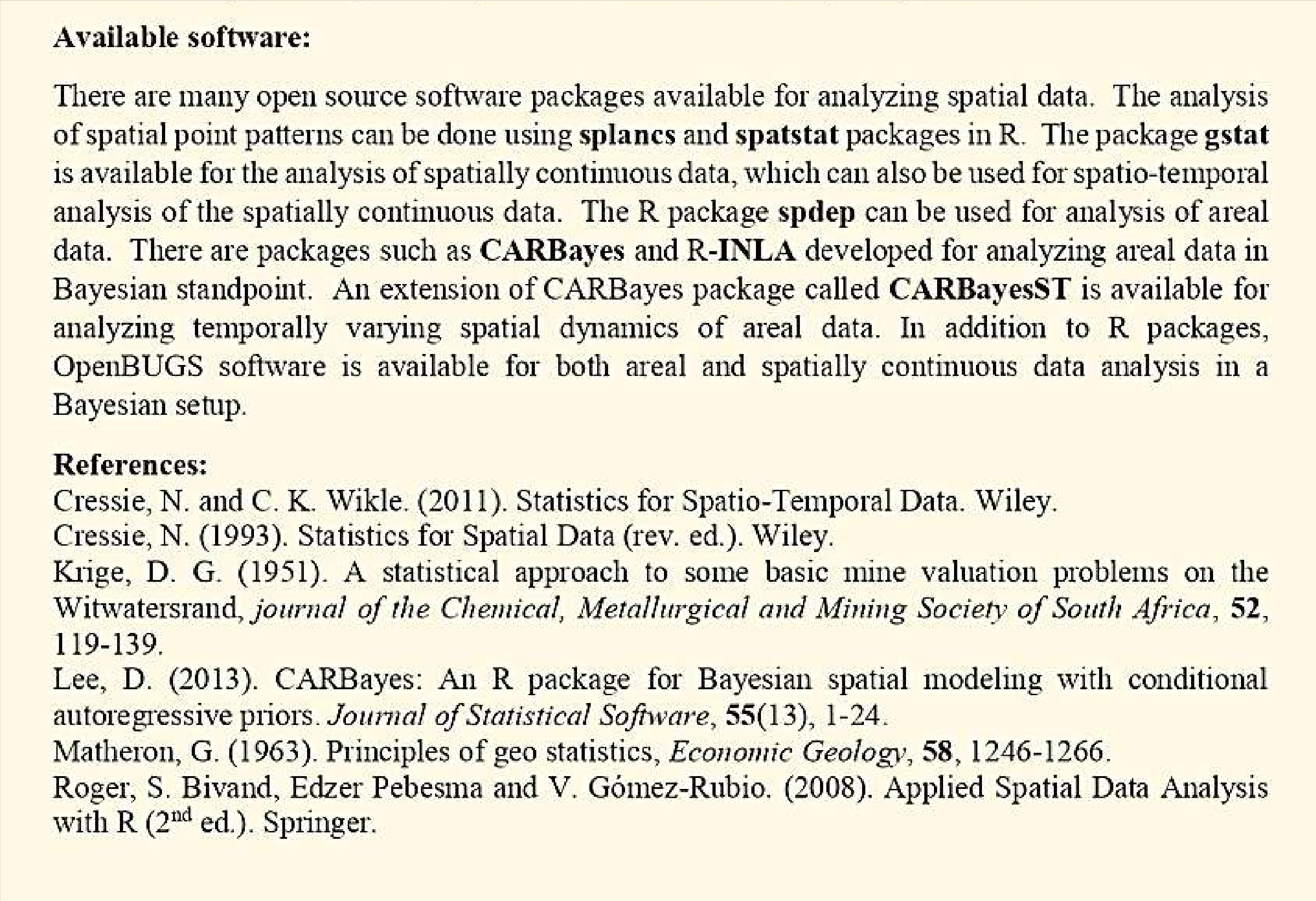
Advances in the field:

Spatial data analysis has evolved considerably during last few decades since its origin in the late 1960s. There are many kriging techniques, which have been developed to serve different purposes.

Cokriging is used when the primary variable of interest is cross-correlated with a secondary variable. For example, in ore sampling, the primary mineral of interest could be contaminated with some other minerals. In such cases, cokriging could be employed, which accounts for the cross-correlations between the variables in the estimation process.

Kriging can be considered in Bayesian aspect as well. The parameters of the kriging process such as slope parameters of the trend term and sill, range, and nugget effect of the covariance function are treated as random variables and prior distributions of them are assumed. Using the Baye's' theorem, where the likelihood function is combined with the prior distribution, an inverse calculation of the conditional probability of the parameters given the data, which is called the posterior distribution, can be derived. However, in general, a closed form for the posterior distribution cannot be derived; hence Marcov Chain Monte Carlo (MCMC) methods are used to simulate the posterior distribution.

The latest development in the field is space-time analysis, where the spatial correlations are combined with the temporal correlations. The advantage of space-time modeling is that future values can be predicted given the past data in addition to the spatial prediction.



Available software:

There are many open source software packages available for analyzing spatial data. The analysis of spatial point patterns can be done using **splancs** and **spatstat** packages in R. The package **gstat** is available for the analysis of spatially continuous data, which can also be used for spatio-temporal analysis of the spatially continuous data. The R package **spdep** can be used for analysis of areal data. There are packages such as **CARBayes** and R-**INLA** developed for analyzing areal data in Bayesian standpoint. An extension of CARBayes package called **CARBayesST** is available for analyzing temporally varying spatial dynamics of areal data. In addition to R packages, OpenBUGS software is available for both areal and spatially continuous data analysis in a Bayesian setup.

References:

Cressie, N. and C. K. Wikle. (2011). Statistics for Spatio-Temporal Data. Wiley.
Cressie, N. (1993). Statistics for Spatial Data (rev. ed.). Wiley.
Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand, *journal of the Chemical, Metallurgical and Mining Society of South Africa*, **52**, 119-139.
Lee, D. (2013). CARBayes: An R package for Bayesian spatial modeling with conditional autoregressive priors. *Journal of Statistical Software*, **55**(13), 1-24.
Matheron, G. (1963). Principles of geo statistics, *Economic Geology*, **58**, 1246-1266.
Roger, S. Bivand, Edzer Pebesma and V. Gómez-Rubio. (2008). Applied Spatial Data Analysis with R (2nd ed.). Springer.

Book Donation on behalf of Dr. Dhanapali Kottachchi

There was a book donation to the IASSL Library, on behalf of Dr. Dhanapali Kottachchi, a life member of IASSL, who passed away in November 2022. The books were handed over to the President of IASSL Dr. Chitraka Wickramarachchi by the family members of Dr. Dhanapali Kottachchi on the 28th November 2022 at the Palitha Sarukkali Memorial Library, in Institute of Applied Statistics Sri Lanka.

We would like to sincerely thank Dr. Kottachchi's family members for choosing IASSL library for this great donation. Further, we hope to offer these valuable books to IASSL members and to the students for enhancing their knowledge.

Least Squares: Connecting Linear Algebra, Calculus and Geometry

Dr. Ranjiva Munasinghe
CEO, MIND Analytics & Management
Assistant Professor, SLIIT

Introduction

Data Science and Artificial Intelligence are major fields of interest at present in both academia and industry. In particular, students and practitioners are encouraged to learn about coding and algorithms. In the past, programmers and computer scientists had to have a strong background in mathematics. The emphasis has shifted more to practical skills and there seems to be a diluting of the earlier prerequisite mathematical knowledge.

In this brief article, I review the connection between the quasi-solution to a linear system via least squares and its underlying geometry. The least squares “solution” can also be derived by calculus. We conclude with a well-known application of this method.

Overdetermined Linear Systems

Let us now consider the linear system of equations:

$$\mathbf{Ax} = \mathbf{b} \qquad (1)$$

We note that in equation (1), **A** is an M x N matrix, **x** is a N x 1 column vector to be determined and **b** is a M x 1 column vector. The entries of **A** and **b** are both known and **x** is to be determined. If M = N, then provided that **A** has a non-zero determinant then we may write the solution as

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} \qquad (2)$$

We work with the case M>N, i.e., there are more equations than unknowns. This is also known as an *overdetermined linear system of equations*. There are trivial cases of over deter mined systems with i) a unique solution and ii) infinitely many solutions, however for the most part these types of linear systems have no solution. We also recall that a linear system with a solution means that **b** is in the column span of **A** (the space of all linear combinations of the column vectors of **A**).

Least Squares Solution via Orthogonal Projection

To find an approximate solution to the overdetermined linear system we start with the assumption that the matrix **A** is of *full rank*. This will ensure the N-dimensional square matrix $\mathbf{A}^T\mathbf{A}$ is invertible, a result that we will use shortly. We know that **b** is not in the *column span of* **A**, so we look for the point in the column span of **A** that is closest to **b** – this will be our quasi-solution. This point is given by the *orthogonal projection of* **b** onto the column space of **A**. An orthogonal projection **P** is a linear map such that

i. $\mathbf{P}^2 = \mathbf{P}$ (Projection definition)
ii. $\mathbf{P}^T = \mathbf{P}$ (Orthogonality definition)

For our problem

$$\mathbf{P} = \mathbf{A}\,[\mathbf{A}^T\mathbf{A}]^{-1}\,\mathbf{A}^T \qquad (3)$$

The keen student should check that the formula above does indeed satisfy the two defining properties of an orthogonal projection. We write the quasi-solution as

$$\mathbf{x}^* = [\mathbf{A}^T\mathbf{A}]^{-1}\,\mathbf{A}^T\,\mathbf{b} \qquad (4)$$

The point in the column span of A that is closest to b is denoted **Ax*** and is given by

$$\mathbf{Ax}^* = \mathbf{A}[\,\mathbf{A}^T\ \mathbf{A}]^{-1}\ \mathbf{A}^T\ \ \mathbf{b} \qquad (5)$$

This solution is also termed the least squares solution - the proof of its existence can be shown via some basic principles of Linear Algebra.

Least Square Solution via Calculus

We now turn to calculus to derive the least squares solution. Define the error vector e as follows

$$\mathbf{e = Ax - b}$$

We look to minimize the Euclidean norm of the error vector with respect to the unknown vector **x** – the norm is given by

$$||\mathbf{e}||_2^2 = \sum_{i=1}^{m} \mathbf{e}_i^2 = \mathbf{e}^T \mathbf{e} = (\mathbf{Ax} - \mathbf{b})^T (\mathbf{Ax} - \mathbf{b})$$

We use matrix calculus to differentiate the expression above and obtain

$$\frac{\partial ||e||_2^2}{\partial X} = -2\mathbf{A}^T\mathbf{b} + 2\mathbf{A}^T\mathbf{Ax}$$

We now set the derivate to zero to find the unique minimum as before in equation (5):

$$\mathbf{x}^* = [\mathbf{A}^T\mathbf{A}]^{-1}\mathbf{A}^T\mathbf{b}$$

We note that in order to confirm this is indeed a minimum, we need to look at the second derivative which is given by the Hessian matrix of 2nd order derivatives (and equal to) and check that it is positive definite. To keep this note brief we omit the details and leave it as an exercise for the diligent student.

The other way to understand the existence and uniqueness is by noting the Euclidean norm of the error vector is a positive multivariate quadratic in the components of **x**, hence it must have a unique minimum.

We have now derived the least squares solution to an overdetermined linear system using i) geometry with linear algebra and ii) calculus.

A Familiar Application & Concluding Remarks

Let us formulate the problem of *multiple linear regression* in matrix notation

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

Note the vector **y** represents the observations of the dependent variable, the data matrix **X** contains the realizations of the independent variables, the vector of parameters β is to be determined and the vector ε represents the residuals.

We also note that in regression problems we typically deal with more observations than variables – i.e., there are more equations than unknowns. We choose our β parameters to minimize the sum of the squared residuals – this is termed *Ordinary Least Squares (OLS) regression*. OLS is exactly the same as finding the least squares solution for an overdetermined system:

$$\hat{\beta} = [x^T x]^{-1} x^T y$$

The predictions are given by

$$\hat{y} = x\hat{\beta} = x[x^T x]^{-1} x^T y$$

In other the words the vector of predictions given by $\hat{y}$ is the orthogonal projection of the vector of observables **y** onto the vector space spanned by the independent variables, i.e., column space of the data matrix **X** .

Understanding the problem of linear regression in the context highlighted is the starting point in understanding algorithms which are used to solve for the regression parameters. We have to keep in mind for large systems, the 'simple' matrix equation with operations of finding a matrix inverse might be cumbersome. We then have to resort to algorithms, such as QR *decomposition, Gradient Descent etc.*, in order to solve for the parameters. In turn these algorithms serve as a stepping stone for learning about more complex algorithms.

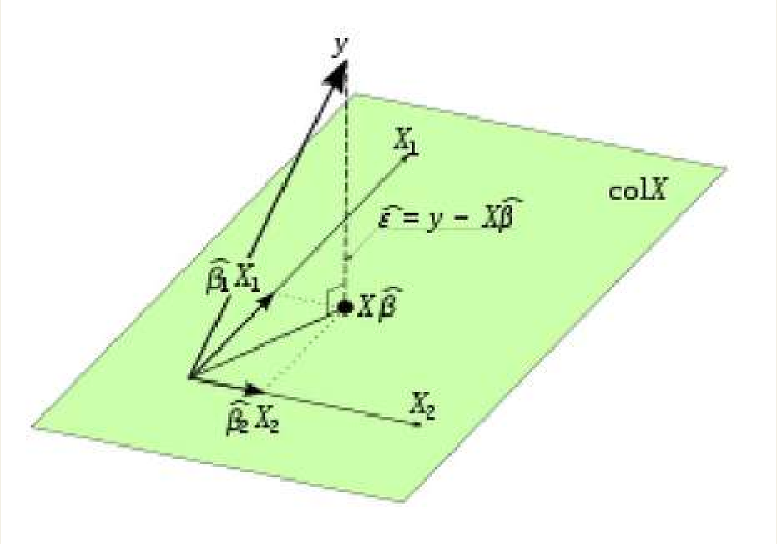


Figure 1: OLS estimation is equivalent to the projection onto the linear space spanned by the regressor variables.

Source: Wikipedia
"Ordinary Least Squares"

Identification of Reasons for Culpable Homicides and Attempted Murders: A Case Study for the Kelaniya Police Division.

Ms. H. Lakni A. Weerakoon
BSc (Honours in Statistics)
University of Kelaniya

Culpable Homicides, attempted murders are ultimate crimes that could create ripple effects on a society which could go far beyond the original loss of human life. In Sri Lanka, the intentional homicide rates have dropped considerably within the last three decades after the Civil war, but an increasing trend is slightly observed since 2017 (Figure 1).

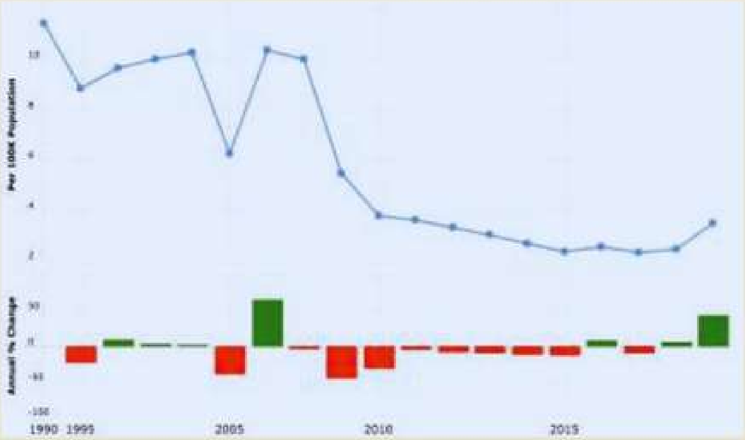


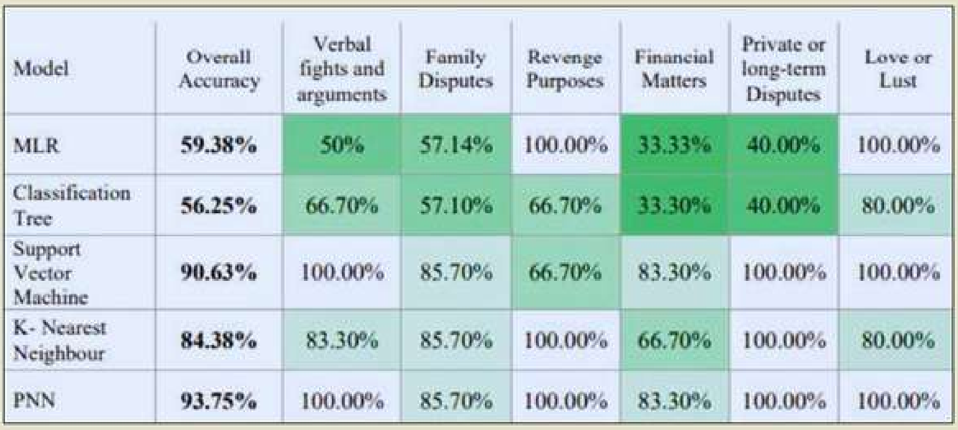
Figure 1 : Sri Lanka Murder/Homicide Rate 1990-2019

Owing to the unpredictable nature of such crimes that require complex investigations, most research were restricted to descriptive studies related to this domain. Furthermore, Multinomial logistic regression model was the often-employed strategy in such circumstances because the majority of the predictors were nominal categorical variables. Hence the objective of the study was to make an attempt to come up with an appropriate model with higher accuracy and predictive ability, to identify the reason for culpable homicide or attempted murder as a decision support tool for relevant authorities. The novelty of the research lies in the domain that is being addressed in the research as well as the concept of using a data mining model as a decision support tool in a complex research area such as crimes.

In Sri Lanka, case-by-case homicide and attempted murder data are preliminarily recorded in Grave Crime Registers (GCR) and Grave Crime Information Book (GCIB) at each police station. This study uses data from 320 cases recorded in aforementioned books in 12 Police Stations in Kelaniya Police Division related to the perpetrator, victim as well as the case, from 2010 to 2020.

The cases reported were observed to be categorised into 6 major categories when collecting data. This was further confirmed by discussions with domain experts and the categories were identified as Verbal fights and arguments, Family Disputes, Revenge Purposes, Financial Matters, Private or long-term Disputes and Love or Lust issues with closest proximity. Pearson Chi-square test was used in identifying the influential explanatory variables. Out of the 18 variables considered based on the literature and the discussions had with domain experts, 8 predictors including Weapon used, Relationship between victim and perpetrator, Location, Number of accusers, Civil Status of the perpetrator, Mental health status of the perpetrator, Gender of victim and Civil Status of the victim were statistically associated with the reason for culpable homicide or attempted murder at 5% level of significance.

Multinomial logistic regression (MLR) followed by four machine learning models including classification tree, support vector machine (SVM), k-nearest neighbour (KNN), and probabilistic neural network (PNN) were fitted initially with a training and testing set which was randomly selected in the ratio 90:10. The 4 machine learning models were then fitted separately by using the bagging technique for assuring robustness. The accuracies were compared using the confusion matrixes and rates of misclassifications of the critical classes. Considering the fitted models, the PNN model easily outperformed other models with the highest accuracy of 93.75% (Figure 2). Hence it is clear that the model was capable of capturing the hidden links among variables related to culpable homicides and attempted murders successfully.



Model	Overall Accuracy	Verbal fights and arguments	Family Disputes	Revenge Purposes	Financial Matters	Private or long-term Disputes	Love or Lust
MLR	**59.38%**	50%	57.14%	100.00%	33.33%	40.00%	100.00%
Classification Tree	**56.25%**	66.70%	57.10%	66.70%	33.30%	40.00%	80.00%
Support Vector Machine	**90.63%**	100.00%	85.70%	66.70%	83.30%	100.00%	100.00%
K- Nearest Neighbour	**84.38%**	83.30%	85.70%	100.00%	66.70%	100.00%	80.00%
PNN	**93.75%**	100.00%	85.70%	100.00%	83.30%	100.00%	100.00%

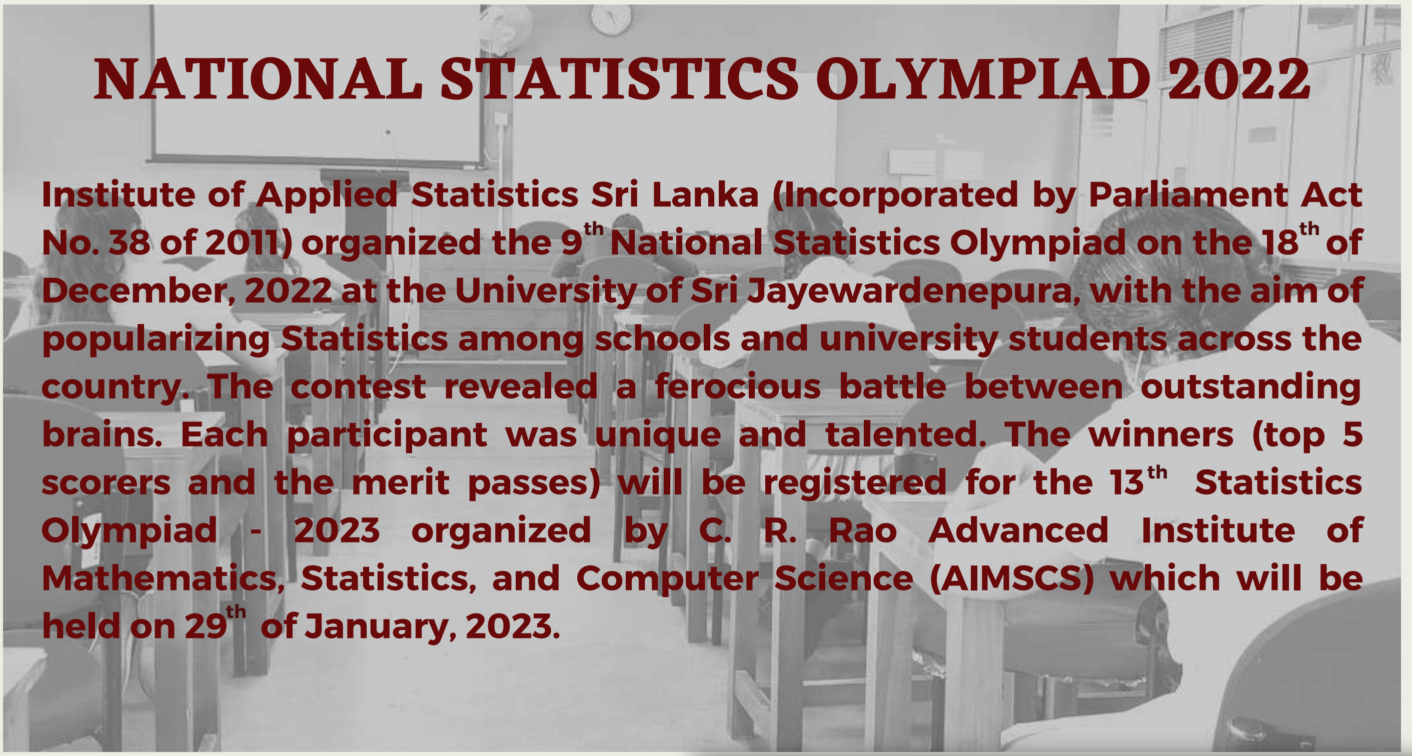
Figure 2: Correct classification accuracies of fitted models

Additionally, some eye-opening information was disclosed as a result of the descriptive analysis that was undertaken along with the model fitting. The majority of the cases were reported due to verbal fights and arguments (21.3%) as well as Private or long-term Disputes (20%). 58.75% of the attempted murders and culpable homicides have taken place during the night light which is from 6.00 pm to 6.00 am and the majority of the cases have taken place in public places or streets/roads (35.31%). Regarding the educational level of the offenders in the research domain, more than half of them (60%) have only completed up to Grade 5,10 or have never attended a school. Sharp or pointed weapons are the key armament used in committing attempted murders or culpable homicides in Kelaniya Police Division which is around 46%.

The assailant/perpetrator was an acquaintance or a stranger to the victim 67.19% of the time. The motive was either a family disagreement or Love/Lust in situations where the perpetrator and victim were intimate partners (15.94%). Most of the victims (17.50%) were above the age of 40, whereas the majority of perpetrators were between the ages of 25 and 39. Married males have been the victims of culpable killings or attempted murders in 41.88 % of the time. Similarly, the vast majority of the criminals were also married males (50.63 %). Surprisingly, around 81.56% of the criminals were employed and 79.06% of the perpetrators did not show any intoxication or illness at the time of incidence.

Despite the fact that crimes in this research area are highly unexpected, the probabilistic neural network model that was finally obtained and which captured the underline interconnections could be used by criminal investigators as a support tool for wise decision-making.

Authors: H. L. A. Weerakoon, N. V. Chandrasekara
Department of Statistics & Computer Science, University of Kelaniya, Sri Lanka
(This article was written based on the first author's final year research project)

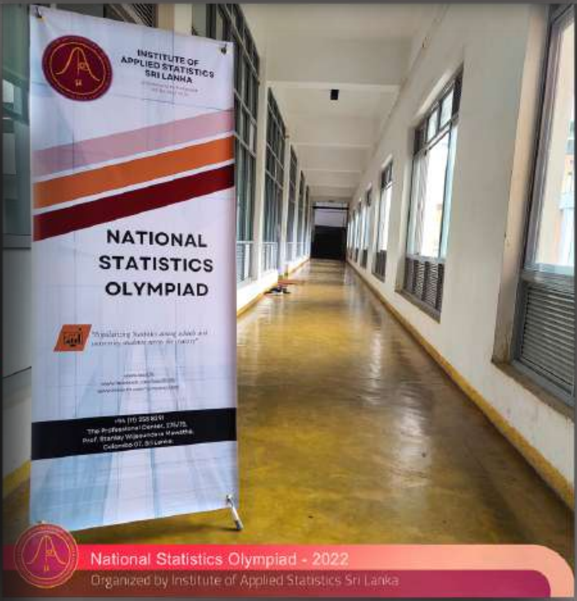


NATIONAL STATISTICS OLYMPIAD 2022

Institute of Applied Statistics Sri Lanka (Incorporated by Parliament Act No. 38 of 2011) organized the 9th National Statistics Olympiad on the 18th of December, 2022 at the University of Sri Jayewardenepura, with the aim of popularizing Statistics among schools and university students across the country. The contest revealed a ferocious battle between outstanding brains. Each participant was unique and talented. The winners (top 5 scorers and the merit passes) will be registered for the 13th Statistics Olympiad - 2023 organized by C. R. Rao Advanced Institute of Mathematics, Statistics, and Computer Science (AIMSCS) which will be held on 29th of January, 2023.

National Statistics Olympiad - 2022
Organized by Institute of Applied Statistics Sri Lanka

NATIONAL STATISTICS OLYMPIAD
National Statistics Olympiad - 2022
Organized by Institute of Applied Statistics Sri Lanka


INSTITUTE OF APPLIED STATISTICS SRI LANKA
NATIONAL STATISTICS OLYMPIAD
National Statistics Olympiad - 2022
Organized by Institute of Applied Statistics Sri Lanka

National Statistics Olympiad - 2022
Organized by Institute of Applied Statistics Sri Lanka

Best Research Awards-2022– Calling for Applications (Undergraduate/Postgraduate/Open Category)

The Institute of Applied Statistics, Sri Lanka (IASSL) is calling applications for the Best Research Awards – 2022 for those who have successfully completed research with the development or application of Statistics/Applied Statistics.

Applications are called under three categories

- Undergraduate
- Postgraduate
- Open category.

- Abstract & Extended Abstract should be submitted in two separate files. The extended abstract (Maximum number of text pages in the Extended Abstract should be four (04) pages - English only). Abstract (Maximum number of words in the text 350) should be submitted along with the application form
- For the Undergraduate category, it should be the extended abstract of the research project Report / Thesis (The effective date of the relevant degree program should be between 01st January 2022 to 31st December 2022)
- For the postgraduate category, it should be the extended abstract of the Thesis/ Directed Study/ Independent Study/ Project report. (The effective date of the relevant degree program should be between 01st January 2022 to 31st December 2022)
- For the Open category, it should be the extended abstract of the article. (The date of the relevant article published should be between 01st January 2022 to 31st December 2022)

Please submit your application form along with the abstract, extended abstract and a verification for the effective date of your Degree (Degree Certificate or Letter from your Head of the Department or your Supervisor) on or before the 10th February 2023 through any of the following email addresses:

iasslresaward2021@gmail.com / kapilar@appsc.sab.ac.lk

A webinar on Business Statistics

We are pleased to announce that the Statistics Popularization Committee, IASSL recently held a successful webinar on November 22nd , 2022 for Advanced Level students who are studying Business Statistics. This webinar was designed to raise awareness and provide motivation on the various degree options available to students after completing their A/L exams.

The webinar featured a panel of experts in the field of Statistics, who shared their insights and experiences on the various career paths available to those with a strong foundation in statistical analysis. The panelists were Mr. P. Dias, a senior lecturer of the Department of Statistics, University of Sri Jayewardenepura, and Dr. Chathurani Silva of the Department of Decision Sciences, University of Sri Jayewardenepura. They also discussed the importance of continuing education in the field of Statistics, and highlighted the many exciting opportunities available to those who pursue advanced degrees in the field of statistics .

Overall, the webinar was a great success, with many attendees expressing their appreciation for the valuable information and advice provided by the panelists. We hope that this event will inspire more students and teachers to consider the many possibilities available to them in the field of statistics, and to pursue higher education in this exciting and rewarding field.

Thank you to all who participated in this event, and we look forward to continuing to support the professional development of our members in the future. IASSL would like to sincerely thank Mr. P. Dias and Dr. Chathurani Silva for their great contribution as the panelists on this webinar. Further, we would also like to thank the Statistics Popularization Committee, IASSL led by Dr. Rajitha M. Silva for organizing this useful and informative session.

Degrees Suitable for Business Statistics Students
Thank you!
Statistics Popularization Committee
IASSL

Achievements

It is with great pleasure we announce the PhD completion of our life member Dr. Manjula Perera. Below we give a brief overview of her PhD study.

Estimation of global scale carbon fluxes using Maximum Likelihood Ensemble Filter

Dr. K. M. P. Perera

ABSTRACT

More advanced data assimilation methods based on statistical and mathematical knowledge are needed to cater to the increased amount of CO2 measurements collected in various platforms. In addition to existing flasks, continuous and aircraft data, CO2 measurements obtained by passenger aircrafts and satellites increase the observation network and provide more constraints on surface carbon flux estimation. This thesis mainly focuses on estimating the surface carbon sources and sinks using CONTRAIL aircraft observations, in addition to the existing in-situ measurements using the ensemble based data assimilation method called Maximum Likelihood Ensemble Filter (MLEF) coupled with Parameterized Chemistry Transport model (PCTM). A pseudodata experiment was carried out by adding CONTRAIL measurements to the observation vector using MLEF coupled with PCTM model, which was driven by GEOS-4 (Goddard Earth Observation System, version 4) weather data for the model validation. Next, MLEF code was developed to conduct the real data experiment to identify the capability on estimating surface carbon fluxes for the period 2009-2011. PCTM model was driven by Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA2) weather data. Solving separate multiplicative biases added for photosynthesis, respiration, and air-sea gas exchange fluxes the estimated fluxes were obtained. Hourly land fluxes, Gross Primary Production (GPP) and respiration obtained from Simple Biosphere-version 3 (SiB3) model, Takahashi ocean fluxes and Brenkert fossil fuel emissions were used.

Pseudodata experiment results showed a considerable uncertainty reduction for Asian region and more than 50% reduction for North American and European regions. According to the results of the real data experiment, North America showed about 60- 80% uncertainty reduction while the Asian and European regions showed moderate results with 50-60% uncertainty reduction. Most other land and oceanic regions showed less than 30% uncertainty reduction. The results were mainly compared with the results of well-known Carbon Tracker (CT2017) which is a CO2 measurement and modeling system developed by NOAA (National Oceanic and Atmospheric Administration). The spatial distribution of estimated mean annual fluxes over North America, Australia and Tropical Asia showed good agreement with the Carbon Tracker results when aggregated into large regions. The biases were poorly constrained in the regions where the CO2 observations are not sufficiently dense such as South America and Africa. Long-term averaged fluxes were compared with several other inversion studies and showed similar results for the Boreal North America, Temperate North America and Australia. Theresults reveal the capability of MLEF method to assimilate large CO2 observation vectors with high performance parallel computing environment with less cost and less time. The impact of satellite observations with MLEF needs to be investigated further and this study forms the basis of the future work in this area.Keywords: Data assimilation, Maximum Likelihood Ensemble Filter, CONTRAIL aircraft observations, Carbon sources and sinks, CO2 modeling.

Supervisor:
Dr. R. S. Lokupitiya, Senior Lecturer, Department of Statistics, Faculty of Applied Sciences, University of Sri Jayewardenepura, Gangodawila, Nugegoda, Sri Lanka.

Names of Co-supervisor(s):

- Prof. A. Scott Denning, Professor, Department of Atmospheric Science, Colorado State University, Fort Collins, CO 80523-1371 USA.
- Prof. E. Y. K. Lokupitiya, Professor in Environmental Science, Department of Zoology and Environment Sciences, University of Colombo, Colombo, Sri Lanka.
- Dr. Prabir Kumar Patra, Principal Scientist, Research Institute for Global Change, JAMSTEC, 3173-25 Showa-machi, Kanazawa-ku, Yokohama, 236-0001, Japan.
- Prof. R. G. N. Meegama, Professor in Computer Science, Department of Computer Science, Faculty of Applied Sciences, University of Sri Jayewardenepura, Gangodawila, Nugegoda, Sri Lanka.

Grants:

- National Research Council, Sri Lanka (NRC-036)
- Asia-Pacific Network for Global Change Research (APN); grant#ARCP2011-11NMY-Patra/Canadell.

Notices

Upcoming Courses

- **Systematic Literature Review (SLR) with Bibliometric Analysis: a way of manuscript writing with PRISMA**

If you are an undergraduate, postgraduate, Ph.D. candidate or a researcher in any area, the SLR helps you scientifically identify the extant knowledge gaps to easily justify your research problem. This course guide you on how to write a manuscript in the area where the knowledge gaps are identified through this method and publish it.

On 28th January & 3rd , 11th, 18th, 25th February 2023 (5 days)

K.G. Priyashantha
Senior Lecturer, Department of Human Resource Management,
University of Ruhuna
kgpriya80@gmail.com

- **Basic Statistics for Managers and Researches**

This Course provides the participants with broad conceptual understanding of data analysis and interpretation.

on 11th, 12th, 18th ,19th & 25th February (5 Days)

Prof. (Ms.) N.R. Abeynayake
Professor in Applied Statistics
Department of Agribusiness Management
Faculty of Agriculture & Plantation Management
Wayamba University of Sri Lanka
Makandura
Gonawila (NWP).

- **Tableau for Business Analytics**

Tableau is a Data Visualisation tool that is widely used for Business Intelligence but is not limited to it. It helps create interactive graphs and charts in the form of dashboards and worksheets to gain business insights.

on 21st , 22 nd, 23rd & 24th February 2023 (4 Days)

Ms. Samudra Bandaranayake.
Senior Data Analyst,
Wiley in Sri Lanka.

For More Information.....

☎ +94 11 2588291
✉ appstatsl@gmail.com
http://www.iassl.lk
http://www.facebook.com
/iassl2020/ https://www.linkedin.com/company/iassl/

New Members

- MISS. S. P. M. SUBODHA
- MISS. H. W. KURUPPUGE
- MR. H. A. I. MADHUSHANKA

Winners of Sudoku Puzzle Competition (Issue 2)

- **1st Place**
  Ms. Tharuni Kavishka
- **2 nd Place**
  Ms.Taniya Fernando
- **3 rd Place**
  Mr. Sathyajith Saliya

Congratulations!!

Diploma in Applied Statistics

ARE YOU AFTER A/L's?

Thinking about what to do next?

Follow Diploma in Applied Statistics at IASSL at an affordable fee. Gateway to Chartered Statistician – A Professional Qualification for a Challenging Career.

Apply Now
https://forms.gle/iafquAenonmEf2mz8
Closing date - 31st January 2023

Higher Diploma in Applied Statistics

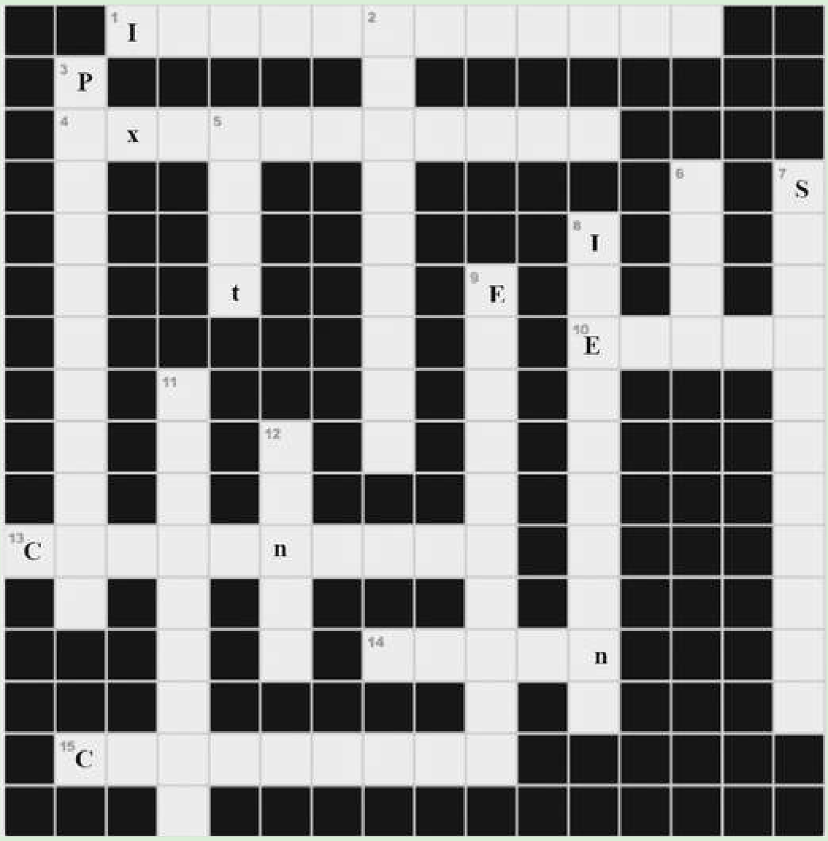
Career Advancement through hands-on experience with advanced data analysis techniques using statistical Software.

Apply Now
https://forms.gle/xCRNGKJVbMJwk4ms9
Closing date - 28th February 2023

Sudoku Puzzle Competition

Please email your submission to appstatsl@gmail.com on or before 15th of April 2023. The draw will be held on the 30th of April, 2023. Correct submissions will be short listed and the winners will be selected considering the order of submission and will be announced in the Issue 1 2023 IASSL newsletter.

We would like to sincerely thank a well-wisher for sponsoring this competition in memory of late Mr. Palitha Sarukkali the first President of IASSL.



INSTRUCTIONS

Across

1. The set of outcomes that belong to A and to B.

4. The variable is used to predict or explain differences in the response variable.

10. The collection of one or more outcomes from an experiment is called

13. A variable that assume any possible value between two points.

14. Set of outcomes that belong to either to A, to B, or both.

15. P(A∩B) = P(B) P(A|B). This result is known as:

15 P(A∩B) = P(B) P(A|B). This result known as:

DOWN

2. A numerical term that summarizes or describe a sample.

3. In mathematics, ... refers to number of ways to order a set of members.

5. For a .... skewed distribution the mean is less than the median.

6. The most frequently occurring value of a data set is called as:

7. This plot examines relationships between pairs of variables.

8. For any event A, AUA = A and A∩A = A. The law is called as.......................law.

9. $\bigcup_{i=1}^{n} B_i = \Omega$; this is called as:

11. This plot examines and compare distributions.

12. A measure of dispersion.

CONTRIBUTIONS TO THE JANUARY-APRIL (ISSUE 1) 2023 NEWSLETTER:

If you have any submissions, comments, suggestions & feedback, please send them to editor@iassl.lk.

WE SINCERELY APPRECIATE ALL WHO CONTRIBUTED TO THIS ISSUE, AND THOSE WHO PARTICIPATED IN THE PREPARATION OF IT.

EDITORIAL BOARD/IASSL

IASSL NEWSLETTER

Official Newsletter of the Institute of Applied Statistics Sri Lanka

Institute of Applied Statistics Sri Lanka
The Professional Center
275/75
Prof. Stanley Wijesundara Mawatha
Colombo 07
Sri Lanka

☎ +94 11 2588291
✉ appstatsl@gmail.com/editor@iassl.lk
🌐 http://www.iassl.lk
f http://www.facebook.com/iassl2020/
in https://www.linkedin.com/company/iassl/