ISSN 0971-0388

Aligarh Journal of Statistics
Vol. 38 (2018), 59-82

Categorizing Tropical Cyclone Tracks Using Mathematical Programming

Arti Pratap Chand, M. G. M. Khan*, Gennady Gienko and James P. Terry

[Received on May, 2017. Accepted on January, 2019]

ABSTRACT

Tropical cyclones (TCs) are one of the most destructive natural hazards in the tropical South Pacific region, so understanding characteristics of their movement is important for hazard risk assessment and adaptation. This study examines the statistical characteristics of track shape for 291 TCs from 1970 to 2008 in the South Pacific (160°E–120°W, 0–25°S). The particular focus is on TC track sinuosity properties and how these may be characterised and grouped using a robust technique such that categories so constructed within which TCs are of similar sinuosity characteristics. In this paper, we propose a mathematical programming approach to determine the optimum boundary points of the categories that seeks minimisation of the sum of weighted variance of sinuosity index values between categories. The problem is formulated as a nonlinear programming problem which is then solved using a dynamic programming technique. Applying the technique we proposed five homogeneous categories of TC tracks found to be (1) straight, (2) near straight, (3) curving, (4) sinuous, and (5) convoluted tracks. The results are compared with the track-shape categories that are obtained by the K-mean cluster analysis method and a hierarchical cluster analysis with Ward's method. The comparison shows that categories constructed by the proposed mathematical programming approach are more homogenous than the categories obtained by other methods.

.

Corresponding author:* Arti Pratap Chand, **School** of Geography, Earth Science and Environment, The University of the South Pacific, Fiji. M. G. M. Khan, School of Computing, Information and Mathematical Sciences, The University of the South Pacific, Fiji. Email: khan_mg@usp.ac.fj Gennady Gienko, Department of Geomatrics, University of Alaska Anchorage, USA. James P. Terry, College of Sustainability Sciences and Humanities, Zayed University, Dubai, United Arab Emirates.

1. Introduction

Tropical cyclones (TCs) are one of the most destructive types of natural hazard that occur in the tropical South Pacific (TSP) region on an annual basis, often causing severe problems for the socio-economic and environmental sectors of developing island nations. More than half the population of the TSP region lives at or near the coast, making such communities highly vulnerable to the damaging effects of TC events. River flooding, storm surge, landslides, strong winds, heavy rainfall and coastal erosion during intense TCs can destroy properties and claim the lives of people and livestock. For example, TC Tomas was an intense cyclone that struck Fiji in 2010, destroying many homes with strong winds and storm surges (Etienne & Terry, 2013). In 1987 TC Uma struck Vanuatu, claiming 48 lives, with costs of damage estimated to approximately USD25 million (UN, 2009). In consequence, improving scientific understanding of patterns in TC formation and migration is a continual task, as this helps in disaster plannng to minimise the human and economic losses inflicted by these powerful storms.

In this study, the focus is on categorising TC track shape. TCs tend to display various track shapes from relatively straight tracks to curving types. More complex shapes may display single or multiple loops. By analysing track shapes and trajectories, other workers have been able to assign TC tracks into well-defined groups or clusters. For typhoons in the North West Pacific, Elsner & Liu (2003) used the K-means clustering method to establish three clusters: straight-moving, recurving and north-oriented tracks, while Camargo *et al.* (2007) successfully used a probabilistic clustering technique based on a regression mixture model.

Another way to quantify TC track shape is to use the metric of track sinuosity, as recently developed and applied in various ocean basins (Terry & Feng, 2010; Terry & Gienko, 2011; Terry *et al*., 2013). Track sinuosity refers to how much a TC 'meanders' during its lifespan compared to the straight distance between its genesis and decay position. Using this system, a perfectly straight-moving TC has a sinuosity value of 1 (the minimum possible value), with sinuosity values increasing as tracks display more curvature. Using sinuosity values, those authors were able to organise TCs tracks into quartile-range sinuosity groups of equivalent size. This works well for a rapid assessment. However, one possible drawback of this approach is that quartile-range groups are unlikely to be homogenous or comparable in terms of their statistical properties. Placing TCs into homogenouous groups that is statistically more robust, which may be

desirable for better comparison and prediction of TC behavoiur across sub-regions where they occur.

In this paper, we propose a mathematical programming technique for categorising TC tracks according to their sinuosity, based on a robust method that can provide greater homogeneity within groups. The objective is to group cyclones into five categories in order to have a central category with data points below representing generally straighter-moving cyclones and those above representing more sinuous tracks. The problem is formulated as a nonlinear programming problem, which is solved by developing an algorithm using a dynamic programming technique (DPT).

2. Study Area and Data Collection

TC track analysis was carried out on data over the period 1969/70 to 2007/08 for 39 cyclone seasons in the study area between 0–25°S and 160°E–120°W, as shown in Figure 1.

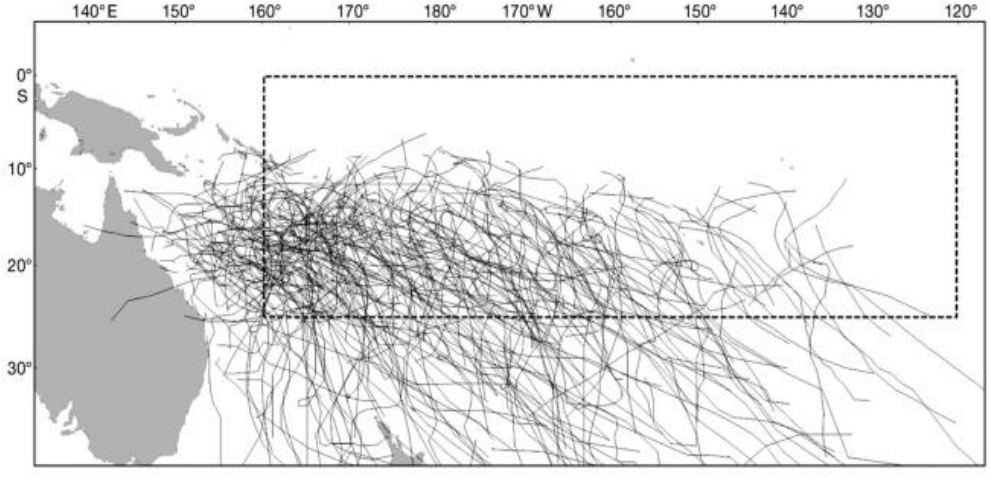

Figure 1: Map of study area showing all cyclone tracks during 1969/70 - 2007/08 cyclone seasons; n=291.

The study period represents the modern era of satellite-based TC observations. The primary data sources are the Fiji Meteorological Service (FMS) and the Tropical Cyclone Warning Centre (TCWC) in New Zealand. Analysis concentrated on the portion of tracks for which TCs were in their mature phase, i.e. with maximum sustained wind speed of 35 knots and above. Portions of

tracks during weaker stages in the early and late life of systems (i.e. depression stages) while winds remained below 35 knots were not included. The dataset records 6-hourly TC center locations and intensities, thus enabling tracks to be plotted within GIS by joining the recorded positions. Altogether 291 cyclones either formed or passed through the study area at some point of their lifespan.

TC track shape was quantified according to track sinuosity index (SI) values based on Terry & Gienko, (2011). SI values are calculated from the formula below:

$$SI = \sqrt[3]{S-1} \times 10$$

Where the sinuosity (S) of an individual TC track is calculated as:

$$S = \frac{\text{TC total travel distance}}{\text{Straight-line displacement from TC genesis to TC decay point}}$$

There are several advantages of using SI-values over S-values for quantification of TC track shape. Skew in the output dataset is reduced. SI-values have an absolute minimum of 0 (zero). This is preferable for straight-tracking storms, rather than a minimum absolute S-value of 1 (unity). TC genesis points with corresponding SI-values can also be mapped for visualization, as in Figure 2. Three extreme values (SI > 14) are identified as outliers and therefore excluded from further analysis.

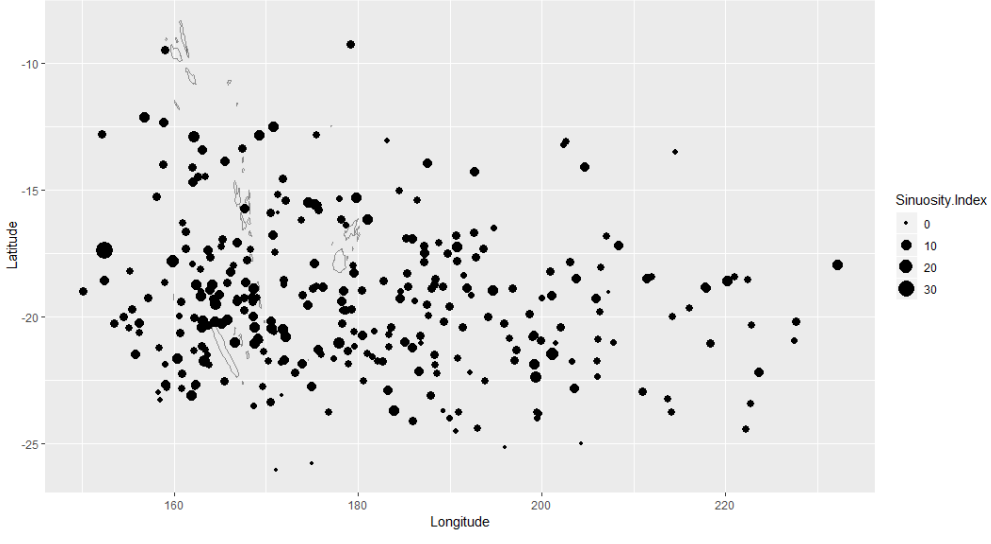

Figure 2: TC genesis points and corresponding track sinuosity index (SI) values in the TSP region; n = 291.

3. Methods for Track Categorization

3.1 K-Means Cluster Analysis:

K-means cluster is a widely-used nonhierarchical clustering method (Steinhaus, 1957; Lloyd, 1982). For *n* data points, if *K* clusters $(C_1,...,C_K)$ are to be formed with maximum cluster homogeneity, then the clusters are so formed that the data points within a cluster are similar to one another and dissimilar to data points in other clusters. In other words, if $p \in C_i$ be a data point and M_i be the mean of the cluster C_i, then the clusters are determined in such a way that the sum of squared distance between all data points in C_i and the mean M_i is maximum, that is,

$$\text{Maximize } \sum_{i=1}^{K} \sum_{p \in C_i} d(p, M_i)^2; \quad i = 1, 2, ..., K$$

Or, the clusters are so formed that the sum of the distance between means of all clusters is minimum, that is,

$$\text{Minimize } \sum_{i \neq j}^{K} d(M_i, M_j); \quad \forall\ i \text{ and } j \tag{3.1}$$

Where, $d(x, y)$ is the Euclidean distance between two points *x* and *y*; and M_i and M_j are the means of the i^{th} and j^{th} cluster, respectively.

This method tries to make the resulting *K* mutually exclusive and spherical shaped clusters as compact and as separate as possible (see Han et al., 2012, Chapter 10).

3.2 Ward's Hierarchical Cluster Analysis:

Another important cluster analysis method is the method of Ward (1963). This is a hierarchical clustering method. If *K* clusters $(C_1,...,C_K)$ are to be formed and x_{ij} $(i = 1, 2, ..., K,\ j = 1, 2, ..., t_i)$ is the value of the j^{th} data point of cluster C_i containing t_i data points, and M_i is the mean of i^{th} cluster, then this method determines the cluster in such a way that it minimizes the sum of the square of the deviation, that is,

$$\text{Minimize} \qquad E = \sum_{i=1}^{K} \sum_{j=1}^{t_i} \left(x_{ij} - M_i \right)^2 . \tag{3.2}$$

Basically, the method resembles an analysis of variance problem, instead of using distance metrics. It involves an agglomerative clustering algorithm, which uses a bottom-up strategy. It starts with individual data points as clusters, which are iteratively merged to form larger clusters. At each step of the algorithm clusters are combined in such a way as to minimize the sum of the square of the deviation (see Han et al., 2012, Chapter 10).

3.3 Proposed Mathematical Programming Technique

In Section 3.1 and 3.2, we outlined two cluster analysis techniques, namely, K-means and Ward's Hierarchical methods. Although, the K-means method is effective for small to medium size datasets, optimizing the within-cluster variation is challenging for the technique. It becomes worse when the techniques encounters a number of possible partitioning that are exponential to the number of clusters and checking the within-cluster variation values. It has been seen that the problem is NP-hard for a general number of clusters even in two-dimensional Euclidean space. The technique also has computational difficulties to achieve global optimum. Instead, it progressively improves the clustering quality and approaches a local minimum (Han et al., 2012). On the other hand, the hierarchical clustering method encounters difficulties while merging the data points. The technique may lead to low-quality clusters, if the data points are not well chosen for merging as it cannot correct the erroneous merges. Another problem in both techniques is that the clustering can be sensitive to outliers, given that outlying objects will tend to have large pairwise distances to other objects and greatly influence the progression of the hierarchical algorithm.

To overcome these problems in clustering we propose a mathematical programming technique that considers a functional form of the data points. The functional form of data consists of observations, which is intrinsically a continuous function of the responses. This can be done by converting the discretely observed data into a continuous frequency function via a smoothing method. The proposed clustering method uses a dynamic programming technique to determine the arbitrarily shaped clusters within which data points are alike as much as possible.

If a variable x is to be classified into *K* mutually-exclusive and homogeneous clusters and $f(x)$ denotes frequency function of $x\left(x_0 \le x \le x_K\right)$, the optimum cluster boundaries are obtained by determining the optimum widths of

the clusters. This is achieved by cutting the range $d = x_K - x_0$ of the distribution at optimum boundary points. The problem of determining optimum cluster widths is formulated as a Nonlinear Programming Problem (NLPP) that minimizes the variances within the clusters. The NLPP is then solved by developing a dynamic programming technique (DPT).

If the NLPP is a multistage decision problem in which the objective function and the constraint are separable functions of cluster widths, then a DPT may be used to solve the problem (Khan *et al*., 2008). DPT determines the optimum solution of a multi-variable problem by decomposing it into stages, each stage compromising a single variable subproblem. A dynamic programming model is basically a recursive equation based on Bellman's principle of optimality (Bellman, 1957). This recursive equation links the different stages of the problem in a manner which guarantees that each stage's optimal feasible solution is also optimal and feasible for the entire problem (Taha, 2007). The use of DPT in various applications, especially for clustering and grouping, can be found in Bellman (1973), Wang & Song (2011) and Nielsen & Nock (2014). The proposed method of clustering based on track SI-values is illustrated as follows:

Let *N* tropical cyclones (TCs) with SI-values (x) be classified into *K* mutually-exclusive and homogeneous categories comprising $N_h; (h = 1, 2, ..., K)$ units in the h^{th} category so that:

$$N_1 + N_2 + ... + N_K = N$$

and the variance of the sinuosity index within the category is as minimum as possible. That is, in order to make the categories internally homogenous, the categories should be constructed in such a way that the variances of the categories be as small as possible. A reasonable criterion to achieve such optimum categories is as follows.

Let x_0 and x_K be the smallest and largest values of sinuosity index (x) respectively, and $\left(x_1, x_2, ..., x_{K-1}\right)$ denotes the set of intermediate optimum boundary points of the categories. If x_{hi} are the values of sinuosity index of the i^{th} TC that falls in h^{th} category, then the problem of optimum categorization can be described as to find the intermediate category boundaries $x_1 \le x_2 \le, ..., \le x_{K-1}$ such that the sum of weighted variance due to the categorization, that is,

$$\sum_{h=1}^{K} W_h \sigma_h^2 \tag{3.3}$$

is minimum.

Where, $W_h = \frac{N_h}{N}$ = the proportion of cyclones that falls in h th category,

$\sigma_h^2 = \frac{\sum_{i=1}^{N_h} \left(x_{hi} - \mu_h \right)^2}{N_h}$ = the variance of h th category, and

$\mu_h = \frac{\sum_{i=1}^{N_h} x_{hi}}{N_h}$ = the mean of h th category.

It should be noted that N_h and x_{hi} are unknown as the categories are yet to be constructed. Further, the problem is to determine the best boundaries that make categories internally homogeneous by minimizing (3.3), which is not a function of boundary points. Therefore, a way to achieve the optimum boundary points effectively is, if (3.3) can be expressed as the function of boundary points which is possible when the distribution of sinuosity index known and then creating categories by cutting the range of the distribution at suitable points.

Let $f(x)$ denotes frequency function of the sinuosity index (x). Then the values of weights W_h and the variance σ_h^2 of h th category are obtained as the function of boundary points $\left(x_{h-1}, x_h \right)$ by

$$W_h = \int_{x_{h-1}}^{x_h} f(x) dx \tag{3.4}$$

$$\sigma_h^2 = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x^2 f(x) dx - \mu_h^2 \tag{3.5}$$

Where,

$$\mu_h = \frac{1}{W_h} \int_{x_{h-1}}^{x_h} x f(x) dx \tag{3.6}$$

Therefore, when the frequency function $f(x)$ is known and is integrable, using the expressions (3.4)-(3.6), the function $W_h\sigma_h^2$ in (3.3) can be expressed as a function of the boundary points (x_{h-1}, x_h).

Let $f_h(x_h, x_{h-1}) = W_h\sigma_h^2$

Thus, the problem of determining the optimum boundaries can be rewritten as:

"Find $x_1, x_2, ..., x_{K-1}$ which minimizes $\sum_{h=1}^{K} f_h(x_h, x_{h-1})$, subject to the constraints $x_0 \le x_1 \le x_2 \le, ..., \le x_{K-1} \le x_K$".

Define, $l_h = x_h - x_{h-1}$

That is, l_h denotes the width of the h^{th} ($h = 1, 2, ..., K$) category.

With the above definition of l_h, the range of the distribution is expressed as

$$\sum_{h=1}^{K} l_h = \sum_{h=1}^{K} (x_h - x_{h-1}) = x_K - x_0 = d$$

The k^{th} boundary point x_k; $k = 1, 2, ..., (K-1)$ is then expressed as:

$$\begin{aligned} x_k &= x_0 + l_1 + l_2 + ... + l_k \\ &= x_{k-1} + l_k \end{aligned}$$

Then, the problem of determining optimum category boundaries can be considered as the problem of determining optimum category widths and could be expressed as the following Nonlinear Programming Problem (NLPP):

$$\left.\begin{array}{ll} \text{Minimize} & \sum_{h=1}^{K} f_h(l_h, x_{h-1}) \\ \text{subject to} & \sum_{h=1}^{K} l_h = d, \\ \text{and} & l_h \ge 0;\ h = 1, 2, ..., K. \end{array}\right\} \qquad (3.7)$$

For $h=1$ the term $f_1(l_1, x_0)$ in the objective function of (3.7) is a function of l_1 alone, as x_0 is known. Similarly, for $h=2$ the term $f_2(l_2, x_1) = f_2(l_2,\ x_0 + l_1)$ will become a function of l_2 alone once l_1 is known. Thus, stating the objective function as a function of l_h alone we may rewrite the NLPP (3.7) as:

$$\left.\begin{aligned} &\text{Minimize} && \sum_{h=1}^{K} f_h(l_h) \\ &\text{subject to} && \sum_{h=1}^{K} l_h = d, \\ &\text{and} && l_h \geq 0;\ h = 1, 2, ..., K. \end{aligned}\right\} \tag{3.8}$$

4. The Solution Procedure using Dynamic Programming Technique

The problem (3.8) is a multistage decision problem in which the objective function and the constraint are separable functions of l_h, which allows us to use a dynamic programming technique (see Khan et al., 2008).

Consider the following subproblem of (8) for first $k\ (< K)$ category.

$$\left.\begin{aligned} &\text{Minimize} && \sum_{h=1}^{k} f_h(l_h) \\ &\text{subject to} && \sum_{h=1}^{k} l_h = d_k, \\ &\text{and} && l_h \geq 0;\ h = 1, 2, ..., k. \end{aligned}\right\} \tag{4.1}$$

where $d_k < d$ is the total width available for division into k category.
Note that $d_k = d$ for $k = K$
Let $f(k, d_k)$ denotes the minimum value of the objective function of (4.1), that is,

$$f(k, d_k) = \left[\min \sum_{h=1}^{k} f_h(l_h) \,\middle|\, \sum_{h=1}^{k} l_h = d_k, \text{ and } l_h \geq 0; h = 1, 2, ..., k \right] \tag{4.2}$$

With this definition of $f(k,d_k)$ the NLPP (3.8) is equivalent to $f(K,d)$, which can be obtained by finding $f(k,d_k)$ recursively for $k=1,2,...,K$ and for all feasible d_k, that is, $0 \le d_k \le d$.

(4.2) can be written as

$$f(k,d_k)=\left[\min\left(f_k(l_k)+\sum_{h=1}^{k-1} f_h(l_h)\right)\middle|\sum_{h=1}^{k-1} l_h = d_k - l_k, \text{ and } l_h \ge 0; h=1,2,...,k\right]$$

For a fixed integer value of l_k, $f(k,d_k)$ is given by

$$f(k,d_k)=f_k(l_k)+\left\{\min\sum_{h=1}^{k-1} f_h(l_h)\middle|\sum_{h=1}^{k-1} l_h = d_k - l_k, \text{ and } l_h \ge 0; h=1,2,...,k-1\right\} \quad (4.3)$$

By the definition (4.2), the quantity inside $\{\ \}$ in (4.3) is $f(k-1,d_k-l_k)$. Thus the required recurrence relation of the Dynamic Programming thus be given as:

$$f(k,d_k)=\min_{0\le l_k\le d_k}\left[f_k(l_k)+f(k-1,d_k-l_k)\right] \quad (4.4)$$

At the final stage of the solution i.e. at $k=K$, $f(K,d)$ is obtained by solving (4.4) recursively for all d_k. From $f(K,d)$ the optimum value l_K^* of l_K is obtained, from $f(K-1,d_{K-1})$ the optimum value l_{K-1}^* of l_{K-1} is obtained and so on until finally we obtain the optimum value l_1^* of l_1.

5. Analysis

5.1 Distribution of Sinuosity Index Values

A probability - probability (P-P) plot of sinuosity index (x) is obtained to determine whether the distribution of x matches a particular distribution. Figure 3 shows that x matches the gamma distribution as the points cluster around a straight line.

Also Kolmogrov-Smirnov test (D = 0.0375, p-value = 0.8212) and the relative frequency histogram shown in Figure 4 reveal that x is assumed to follow Gamma distribution with a probability density function given by

$$f(x)=\frac{1}{\theta^r\Gamma(r)}x^{r-1}e^{-\frac{x}{\theta}};\quad x>0; r,\theta>0. \quad (5.1)$$

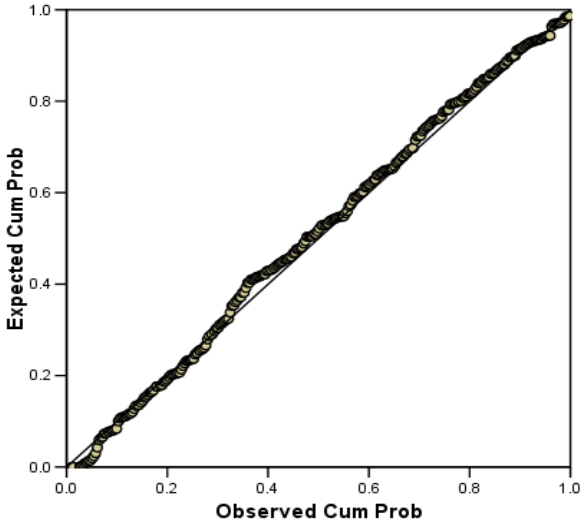

Figure 3: Gamma P-P plot for sinuosity index values (x) of TC tracks; n = 288.

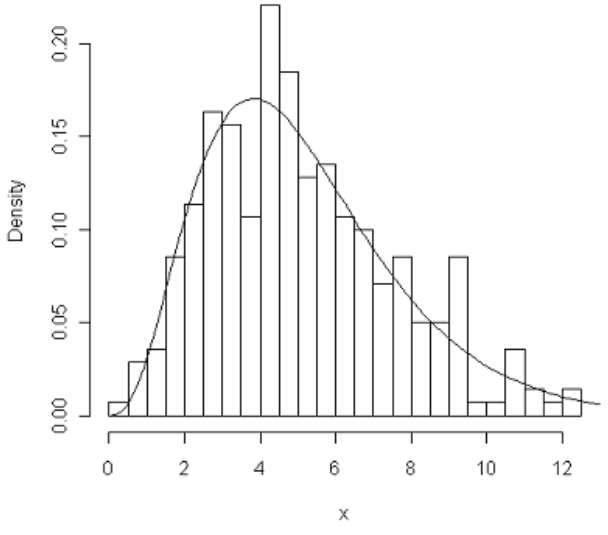

Figure 4: Frequency distribution of sinuosity index of TC tracks; n = 288.

Where *r* is the shape parameter and θ is the scale parameter of the distribution, and $\Gamma(r)$ is a Gamma function defined by:

$$\Gamma(r) = \int_0^\infty t^{r-1} e^{-t} dt, \qquad r > 0\,. \tag{5.2}$$

The function in (5.2) is also defined by an upper incomplete gamma function $\Gamma(r,x)$ and a lower incomplete gamma function $\gamma(r,x)$, respectively, as follows:

$$\Gamma(r,x)=\int_x^{\infty} t^{r-1}e^{-t}dt \tag{5.3}$$

$$\gamma(r,x)=\int_0^{x} t^{r-1}e^{-t}dt \tag{5.4}$$

There also exist regularized/normalized incomplete gamma function, which give a value restricted between 0 and 1 and can be stated as:

$$Q(r,x)=\frac{1}{\Gamma(r)}\int_x^{\infty} t^{r-1}e^{-t}dt, \qquad r,x>0;\Gamma(r)\neq 0 \tag{5.5}$$

$$P(r,x)=\frac{1}{\Gamma(r)}\int_0^{x} t^{r-1}e^{-t}dt, \qquad r,x>0;\Gamma(r)\neq 0, \tag{5.6}$$

where $Q(r,x)$ denotes the upper regularized incomplete gamma function and $P(r,x)$ denotes the lower regularized incomplete gamma function. Note that $Q(r,x)=1-P(r,x)$.

5.2 Estimate of the Distribution Parameters

Using the maximum likelihood estimate (MLE) method for the sinuosity index data, the parameters of Gamma distribution given in (5.1) are found to be:

$$\text{Shape, } \hat{r} = 3.822976 \quad \text{and} \quad \text{scale, } \hat{\theta} = 1.351949 \tag{5.7}$$

5.3 Determination of Optimum Category

Using (3.4), (3.5), (3.6), (5.3) and (5.5) we obtain W_h, μ_h and σ_h^2 as follow:

$$W_h=\left[Q\left(r,\frac{x_{h-1}}{\theta}\right)-Q\left(r,\frac{x_h}{\theta}\right)\right] \tag{5.8}$$

Note that from the definition of l_h,

$$x_h=x_{h-1}+l_h. \tag{5.9}$$

Thus,

$$W_h=\left[Q\left(r,\frac{x_{h-1}}{\theta}\right)-Q\left(r,\frac{x_{h-1}+l_h}{\theta}\right)\right] \tag{5.10}$$

Similarly, μ_h is obtained as

$$\mu_h = \frac{\theta r\left[Q\left(r+1, \frac{x_{h-1}}{\theta}\right) - Q\left(r+1, \frac{x_{h-1}+l_h}{\theta}\right)\right]}{\left[Q\left(r, \frac{x_{h-1}}{\theta}\right) - Q\left(r, \frac{x_{h-1}+l_h}{\theta}\right)\right]} \tag{5.11}$$

and σ_h^2 is reduced to

$$\sigma_h^2 = \frac{\theta^2 r(r+1)\left[Q\left(r+2, \frac{x_{h-1}}{\theta}\right) - Q\left(r+2, \frac{x_{h-1}+l_h}{\theta}\right)\right]}{\left[Q\left(r, \frac{x_{h-1}}{\theta}\right) - Q\left(r, \frac{x_{h-1}}{\theta}\right)\right]} - \frac{\theta^2 r^2\left[Q\left(r+1, \frac{x_{h-1}}{\theta}\right) - Q\left(r+1, \frac{x_{h-1}+l_h}{\theta}\right)\right]}{\left[Q\left(r, \frac{x_{h-1}}{\theta}\right) - Q\left(r, \frac{x_{h-1}+l_h}{\theta}\right)\right]^2} \tag{5.12}$$

Therefore, from (5.10) and (5.12), the expression (3.3) reduces to

$$\sum_{h=1}^{K} \theta^2 r(r+1)\left[Q\left(r+2, \frac{x_{h-1}}{\theta}\right) - Q\left(r+2, \frac{x_{h-1}+l_h}{\theta}\right)\right] - \frac{\theta^2 r^2\left[Q\left(r+1, \frac{x_{h-1}}{\theta}\right) - Q\left(r+1, \frac{x_{h-1}+l_h}{\theta}\right)\right]}{\left[Q\left(r, \frac{x_{h-1}}{\theta}\right) - Q\left(r, \frac{x_{h-1}+l_h}{\theta}\right)\right]} \tag{5.13}$$

Then, using (5.13), the NLPP (3.8) to determine the optimum widths l_h and hence the optimum boundary points (x_{h-1}, x_h) of the categories could be expressed as:

$$\left.\begin{aligned} &\text{Minimize} \quad \sum_{h=1}^{K} \theta^2 r(r+1)\left[Q\left(r+2, \frac{x_{h-1}}{\theta}\right) - Q\left(r+2, \frac{x_{h-1}+l_h}{\theta}\right)\right] - \frac{\theta^2 r^2\left[Q\left(r+1, \frac{x_{h-1}}{\theta}\right) - Q\left(r+1, \frac{x_{h-1}+l_h}{\theta}\right)\right]}{\left[Q\left(r, \frac{x_{h-1}}{\theta}\right) - Q\left(r, \frac{x_{h-1}+l_h}{\theta}\right)\right]} \\ &\text{subject to} \quad \sum_{h=1}^{K} l_h = d, \\ &\text{and} \quad l_h \geq 0;\ h = 1, 2, ..., K. \end{aligned}\right\} \tag{5.14}$$

Where, d is the known constant, that is, the range of the sinuosity indexes:

$$d = x_L - x_0 = 12.1561 - 0 = 12.1561.$$

and the values of r and θ are the parameters of the distribution of sinuosity index given in (5.7). Substituting the values of d, r and θ, the NLPP (5.14) is expressed

as:

$$\text{Minimize} \quad \sum_{h=1}^{K} \left\{ \begin{array}{l} (1.351949)^2 (3.822976)(4.822976)\left[Q\left(5.822976, \frac{x_{h-1}}{1.351949}\right) - Q\left(5.822976, \frac{x_{h-1}+l_h}{1.351949}\right)\right] \\ - \dfrac{(1.351949)^2 (3.822976)^2 \left[Q\left(4.822976, \frac{x_{h-1}}{1.351949}\right) - Q\left(4.822976, \frac{x_{h-1}+l_h}{1.351949}\right)\right]}{\left[Q\left(3.822976, \frac{x_{h-1}}{1.351949}\right) - Q\left(3.822976, \frac{x_{h-1}+l_h}{1.351949}\right)\right]} \end{array} \right\}$$

$$\text{subject to} \quad \sum_{h=1}^{K} l_h = 12.1561,$$

$$\text{and} \quad l_h \geq 0; \; h = 1, 2, ..., K. \tag{5.15}$$

Then, the recurrence relation (4.4) of the Dynamic Programming reduces to:

$$f(k, d_k) = \min_{0 \leq l_k \leq d_k} \left[\begin{array}{l} (1.351949)^2 (3.822976)(4.822976)\left[Q\left(5.822976, \frac{x_{k-1}}{1.351949}\right) - Q\left(5.822976, \frac{x_{k-1}+l_k}{1.351949}\right)\right] \\ - \dfrac{(1.351949)^2 (3.822976)^2 \left[Q\left(4.822976, \frac{x_{k-1}}{1.351949}\right) - Q\left(4.822976, \frac{x_{k-1}+l_k}{1.351949}\right)\right]}{\left[Q\left(3.822976, \frac{x_{k-1}}{1.351949}\right) - Q\left(3.822976, \frac{x_{k-1}+l_k}{1.351949}\right)\right]} \\ + f(k-1, d_k - l_k) \end{array} \right] \tag{5.16}$$

Note that the $(h-1)$th boundary point x_{k-1} is obtained by

$$\begin{aligned} x_{k-1} &= x_0 + l_1 + l_2 + \cdots + l_{k-1} \\ &= l_1 + l_2 + \cdots + l_{k-1} \quad [\text{as } x_0 = 0] \\ &= d_{k-1} \\ &= d_k - l_k \end{aligned}$$

Where d_k is the total range or width of first k strata.

Substituting this value of x_{k-1}, the recurrence relation (28) becomes

$$f(k, d_k) = \min_{0 \leq l_k \leq d_k} \left[\begin{array}{l} (1.351949)^2 (3.822976)(4.822976)\left[Q\left(5.822976, \frac{d_k - l_k}{1.351949}\right) - Q\left(5.822976, \frac{d_k}{1.351949}\right)\right] \\ - \dfrac{(1.351949)^2 (3.822976)^2 \left[Q\left(4.822976, \frac{d_k - l_k}{1.351949}\right) - Q\left(4.822976, \frac{d_k}{1.351949}\right)\right]}{\left[Q\left(3.822976, \frac{d_k - l_k}{1.351949}\right) - Q\left(3.822976, \frac{d_k}{1.351949}\right)\right]} \\ + f(k-1, d_k - l_k) \end{array} \right]$$

(5.17)

6. Results and Discussion

Solving the recurrence relation (5.17), the optimum category widths $l_h; (h = 1, 2, ..., K)$ and hence the optimum category boundaries $x_h = x_{h-1} + l_h$ are achieved by executing a C++ computer program coded for the proposed algorithm discussed above. In previous work (Terry & Feng, 2010) TCs were categorized in four categories: straight – quasi-straight – curving – sinuous. The technique works well for rapid categorization and assessment. However, where statistically homogenous groups may be required for detailed analysis, it is proposed to develop this quartile-based technique further by introducing a central category and refining descriptive names as follows: straight – near straight – curving – sinuous – convoluted. If five categories are to be formed, that is $K = 5$, then the proposed method using technique gives the optimum boundary points for each category for Problem (5.15) as shown in Table 1:

Table 1: Five categories using the proposed dynamic programming approach.

Category (h)	**Width** (l_h)	**Sinuosity Index Values** (x_{h-1}, x_h)	**No. of cyclones** (N_h)	**Weight** (W_h)	**Variance** (σ_h^2)	**Weighted Variance** ($W_h\sigma_h^2$)
1	2.9648	0 – 2.9395	67	0.2326	0.7870	0.1831
2	1.5138	2.9738 – 4.4344	68	0.2361	0.2085	0.0492
3	1.6310	4.4814 – 6.0605	65	0.2257	0.1934	0.0437
4	2.1214	6.1287 – 8.1613	53	0.1840	0.3380	0.0622
5	3.9251	8.2947 – 12.1561	35	0.1215	1.2161	0.1478
			N = 288*			$\sum W_h\sigma_h^2$ = 0.4860

* Table 1 includes 288 cyclones after elimination of three cyclones as outliers.

In order to investigate the effectiveness of the proposed categorization method, we compare the results obtained by the following three methods:

1. A proposed method using a dynamic programming technique.
2. K-Means cluster analysis method (see Steinhaus, 1957; Lloyd, 1982).
3. Hierarchical cluster analysis method with Ward's method (see Ward, 1963).

Using SPSS, five homogeneous categories of cyclones were determined based on sinuosity index values for the K-Mean and Ward's methods and the results are

shown in Table 2 and 3. The variance of each category and the sum of the weighted variance are also presented.

Table 2: Five categories using K-Mean cluster analysis.

Category (h)	**Sinuosity Index** (x)	**No. of cyclones** (N_h)	**Weight** (W_h)	**Variance** (σ_h^2)	**Weighted Variance** ($W_h\sigma_h^2$)
1	0 – 2.9738	68	0.2361	0.7914	0.1869
2	0.0037-5.0093	95	0.3299	0.3381	0.1115
3	5.0723-7.2023	71	0.2465	0.4136	0.1020
4	7.2748-9.5119	43	0.1493	0.5091	0.0760
5	10.3605 – 12.1561	11	0.0382	0.3692	0.0141
		N = 288			$\sum W_h\sigma_h^2$ = 0.4905

Table 3: Five categories using Ward's hierarchical cluster analysis.

Category (h)	**Sinuosity Index** (x)	**No. of cyclones** (N_h)	**Weight** (W_h)	**Variance** (σ_h^2)	**Weighted Variance** ($W_h\sigma_h^2$)
1	0-1.5605	18	0.0625	0.3138	0.0196
2	1.5605-3.9192	85	0.2951	0.3489	0.1030
3	3.9812-5.0093	60	0.2083	0.0936	0.0195
4	5.0723-7.2748	72	0.2500	0.4279	0.1070
5	7.3846-12.1561	53	0.1840	1.6523	0.3041
		N = 288			$\sum W_h\sigma_h^2$ = 0.5532

From the Tables 1-3, the results show that all the categories except Category 5 in the proposed method produce smaller variance as compared to K-Mean method. Whereas, Categories 2, 4 and 5 in proposed method have smaller variances as compared to Ward's method. Moreover, the sum of weighted variance (0.4860) is also smaller for the proposed method as compared to K-Mean (0.4905) and Ward's (0.5532) methods.

Table 4 summarizes the sum of weighted variance $\left(\sum W_h\sigma_h^2\right)$ of each method and presents the percentage of gain in relative efficiency of the proposed method over

others. It reveals that the percentage of gain in relative efficiency of the proposed method over the K-Mean and Ward methods is 0.93% and 13.83%, respectively.

Table 4: Percentage of Gain in Relative Efficiencies (R.E) of the proposed method

Method	Weighted sum of variance ($\sum W_h \sigma_h^2$)	% gain in R.E. of proposed method
K-Mean	0.4905	0.93%
Ward's	0.5532	13.83%
Proposed	0.4860	-

In interpreting the results from a geographical point of view, Ward's approach places boundaries in the way that the last category significantly outweighs any of the first four. Thus, one would also expect that the majority of TCs will be in the first three categories (straight, near straight, and curving), but this is not the case with Ward's method. K-means seems to perform better in both aspects of categorization: placing boundaries of classes with a well-balanced distribution of the number of TCs across the range of five categories. While the fifth category is wider in the K-means' method, the proposed dynamic programming method probably gives the most homogeneous, "natural" distribution of the number of TCs in each category.

On the basis of these comparisons, it can be concluded that categorization using the proposed mathematical programming technique is a more efficient approach. Table 5 suggests suitable names for the five categories along with the outlier category. The number and the percentage of TCs that fall into each category are also presented.

Table 5: Descriptions of the categories based on sinuosity index values.

Categories	TCs with Sinuosity Index	No. of TCs	Percentage of TCs	Description
1	0-2.9648	67	23.3	Straight Tracks
2	2.9648-4.4786	68	23.6	Near Straight Tracks
3	4.4786-6.1096	65	22.6	Curving Tracks
4	6.1096-8.2310	53	18.4	Sinuous Tracks
5	8.2310-12.1561	35	12.1	Convoluted Tracks
6	14.9714, 15.1886 and 37.2637	3		Extreme Sinuous Tracks (Outlier)

Based on our designated categories (Table 5), maps can be plotted in order to visualize simultaneously both TC genesis and track type as shown in Figures 4 (a)-(e). The maps reveal that a clear geographical contrast exists between Categories 1&2 and Categories 4&5. TCs in first two categories (straight and near straight tracks) are distributed fairly evenly between 160°E to 130°W. TCs in Categories 4&5 (sinuous and convoluted tracks) are generally limited to the far west of the study area, mainly between 160°E to 180°. Such information is important for regional agencies tasked with hazard observation and disaster preparedness planning. It reveals in particular that island groups located in the geographical west of the tropical South Pacific are more likely to experience TCs that follow tracks with a larger amount of meandering and complexity. Such TCs pose a greater degree of difficulty for forecasters in terms of predicting movement and trajectories in real time than TCs which move along straighter paths, and therefore may require greater scrutiny as they evolve and mature.

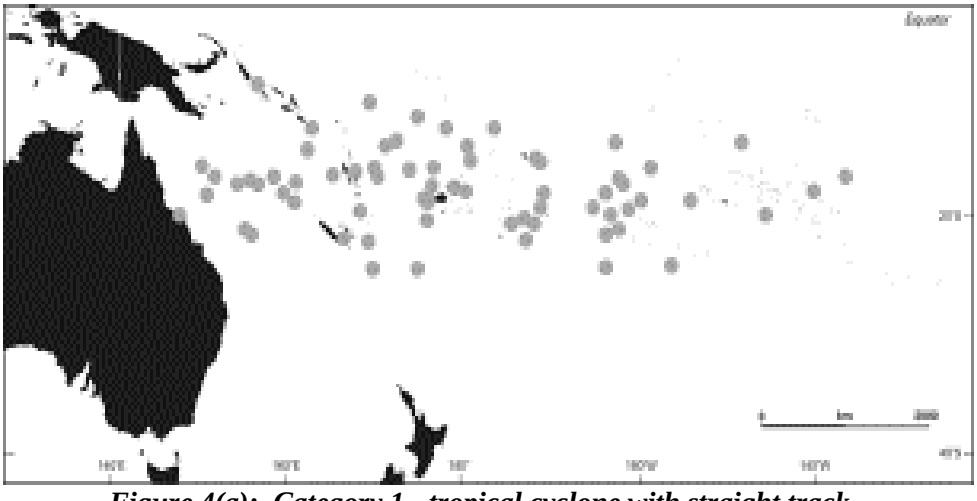

Figure 4(a): Category 1 - tropical cyclone with straight track.

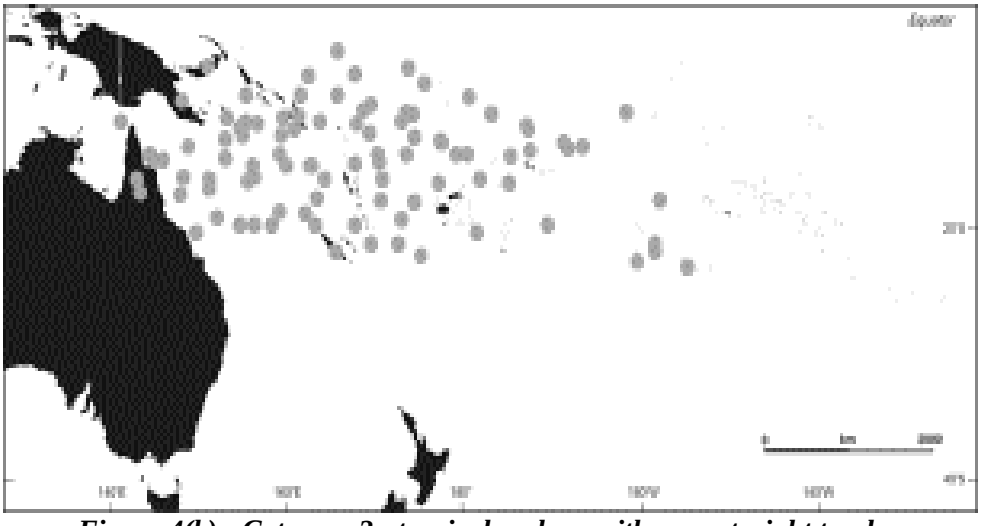

Figure 4(b): Category 2 - tropical cyclone with near straight track.

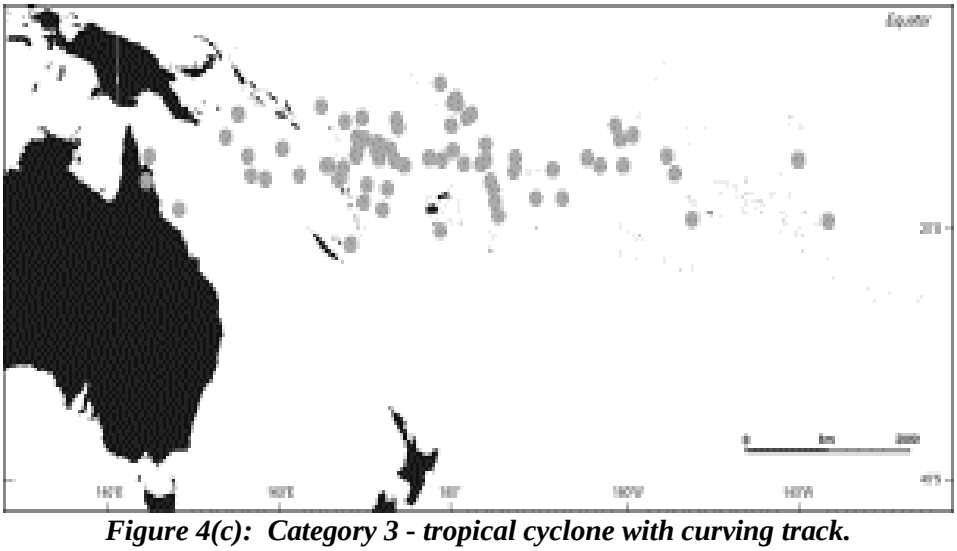

Figure 4(c): Category 3 - tropical cyclone with curving track.

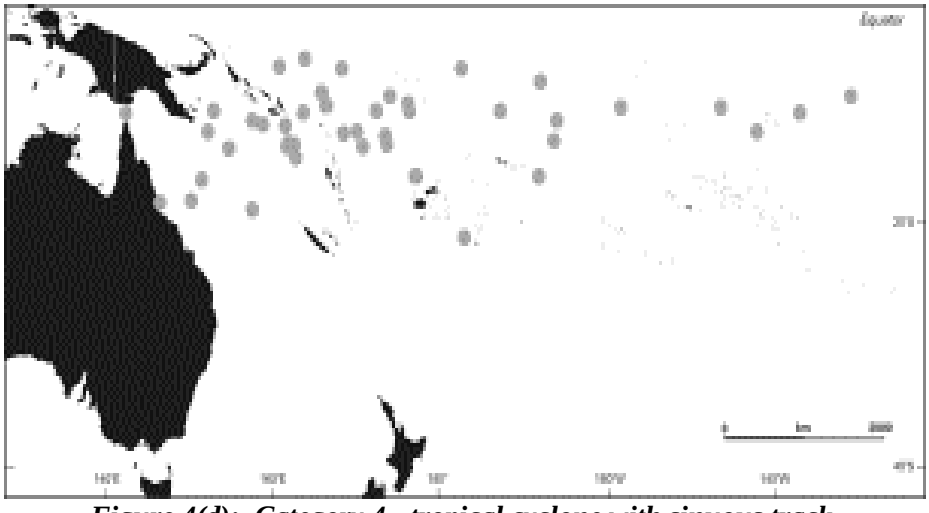

Figure 4(d): Category 4 - tropical cyclone with sinuous track.

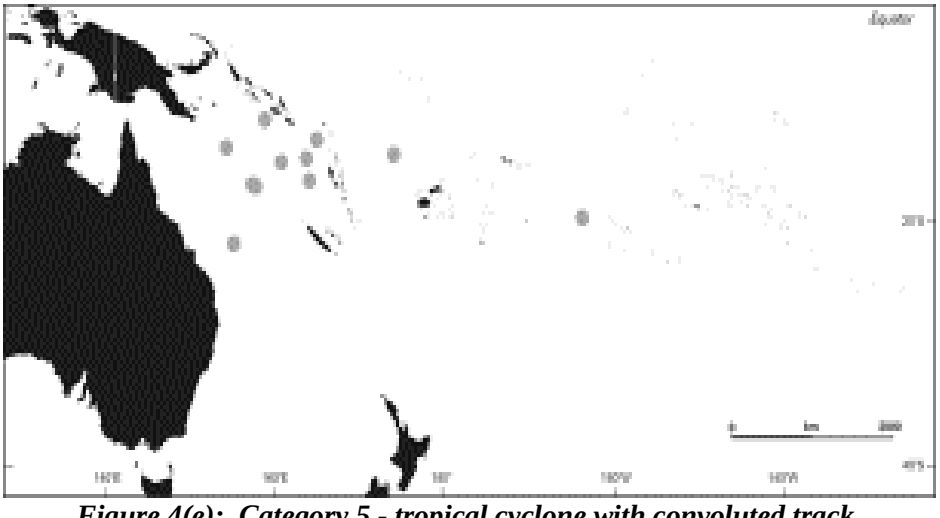

Figure 4(e): Category 5 - tropical cyclone with convoluted track.

7. Conclusion

According to the IPCC (2012), continued use of climate models to make projections of tropical cyclone behavior (in terms of frequency, location, intensity, rainfall and movement) remains a high priority for Pacific region. Improved understanding in predicting tropical cyclone behavior is important for better preparedness before, during and after TC events.

In this paper, a total of 291 tropical cyclones occurred in the Southwest Pacific over 39 seasons from 1969/70 to 2007/08 was analyzed. The linear shape of the tracks of these TCs was quantified using a dimensionless metric based on an existing method for measuring track sinuosity (S), which determines how much an individual track deviates from a straight line between cyclone genesis and decay points. Sinuosity index (SI) values calculated from measured S-values reduces skew in the dataset and fixes the absolute minimum value to zero, which is preferable for straight tracks. A Gamma distribution was found to provide the best fit to the SI data.

The major aim of the paper was to develop a new robust method to categorize TCs into groups based on their track sinuosity properties. Five homogeneous categories were formed using a mathematical programming approach to obtain optimum boundary points for each category. These five categories contain 288 of the total 291 TC tracks measured (excluding three outliers). When comparing the efficiency of our proposed categorization method against two well-known alternative clustering methods, our method was found to be more efficient. Thus, when a robust method is required that forms groups with a high degree of internal homogeneity (e.g. for comparison across sub-regions where TCs occur), the proposed mathematical programming method appears to perform better than other clustering techniques.

Reference

Bellman, R. (1957): Dynamic Programming, Princetown University Press, New Jersey.

Bellman, R. (1973): A note on cluster analysis and dynamic programming. Mathematical Biosciences, **18**(3-4), 311-312.

Camargo, S.J., Robertson, A.W., Gaffney, S.J., Smyth, P., and Ghil, M. (2007): Cluster analysis of typhoon tracks. Part I: general properties, Journal of Climate, **20**, 3635-3653.

Elsner, J. B., and Liu, K. B. (2003): Examining the ENSO-typhoon hypothesis, Climate Research **25**, 43-54.

Etienne, S., and Terry, J.P. (2012): Coral boulders, gravel tongues and sand sheets: features of coastal accretion and sediment nourishment by Cyclone Tomas (March 2010) on Taveuni Island, Fiji, Geomorphology, 175-176, 54-65.

Han, J., Kember, M., and Pei, J. (2012): Data Mining Concepts and Techniques (3rd eds.). Elsevier Inc.

Harr, P.A., and Elsberry, R.L. (1991): Tropical cyclone track characteristics as a function of large – scale circulation anomalies, Monthly Weather Review, **119**, 1448-1468.

IPCC. (2014): Summary for Policymakers. Impacts, Adaptation and Vulnerability. Working Group II of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK, and New York, NY, USA. Accessed 25 April 2014 from http://ipcc-wg2.gov/AR5/images/uploads/IPCC_WG2AR5_SPM_Approved.pdf

IPCC. (2012): Summary for Policymakers. In Managing the Risks of Extreme Events and Disasters to Advance Climate Change Adaptation [Field, C. B., V. Barros, T. F.Stocker, D. Qin, D. J. Dokken, K. L. Ebi, M.D. Mastrandrea, K. J. Mach, G.-K. Plattner, S. K. Allen, M. Tignor, and P. M. Midgley (eds.)]. A Special Report of Working Groups I and II of the Intergovernmental Panel on Climate Change. Cambridge University Press, Cambridge, UK, and New York, NY, USA.1-19.

Khan, M.G.M., Nand, N., and Ahmad, N. (2008): Determining the optimum strata boundary points using dynamic programming, Survey Methodology, **34**(2), 205-214.

Lloyd, S. P. (1982): Least squares quantization in PCM, IEEE Transactions on Information Theory, **28** (2), 129-137.

Nielsen, F., and Nock, R. (2014): Optimal interval clustering: application to Bregman clustering and statistical mixture learning, signal processing letters, IEEE Signal Processing Letters. **21**(10), 1289-1292.

Republic of Vanuatu. (2009): A special submission to the UN committee for development policy on Vanuatu's LDC status. Republic of Vanuatu. Accessed 28 September 2010 from http://www.un.org/esa/policy/devplan/profile/plen4d_cdp2009.pdf.

Steinhaus, H. (1957): Sur la division des corps matériels en parties, Bull. Acad. Polon. Sci. (in French), **4**(12), 801-804.

Taha, H. A. (2007): Operations Research: An Introduction (8th edition), Pearson Education, Inc.

Terry, J.P., and Feng, C-C. (2010): On quantifying the sinuosity of typhoon tracks in the western North Pacific basin, Applied Geography, **30**, 678-686.

Terry, J.P., and Gienko, G. (2011): Developing a new sinuosity index for cyclone tracks in the tropical South Pacific, Natural Hazards, 59, 1161-1174.
Terry, J.P., Kim, I.H., and Jolivet, S. (2013): Sinuosity of tropical cyclone tracks in the South West Indian Ocean: spatio-temporal patterns and relationships with fundamental storm attributes, Applied Geography, **45**, 29-40.
Wang, H., and Song, M. (2011): Ckmeans.1d.dp: optimal k-means clustering in one dimension by dynamic programming, The R Journal, **3**(2), 29-33.
Ward, J. H., Jr. (1963): Hierarchical grouping to optimize an objective function, Journal of the American Statistical Association, **58**, 236–244.