交通部中央氣象局委託研究計畫成果報告

發展（6月至10月）行經台灣附近海域內之颱風個數預測模式

計畫類別：□國內　■國外

計畫編號：MOTC-CWB-92-3M-05

執行期間：92年2月1日至92年12月31日

計畫主持人：朱寶信

執行單位：Climate Systems Enterprise

中華民國九十二年十二月

Climate Prediction of Tropical Cyclone Counts in the Vicinity of Taiwan Using the Least Absolute Deviations Method

Pao-Shin Chu, Xin Zhao

Climate System Enterprise
Honolulu, Hawaii 96816
U. S. A.

December 1, 2003

1. Linear Regression Model

In this project, we have developed a statistical model for the purpose of predicting the annual number of tropical cyclones (TCs) in the vicinity of the Taiwan area. A linear regression model which assumes the following form is used:

$$y(t) = \sum_{j=1}^{K} c_j x_j(t) + N(t) \tag{1}$$

where $y(t)$ is the desired predictive variable or say predictand, $x_i(t), i = 1,...,K$ represent the predictors and $c_i, i = 1,...,K$ are the relative regression parameters. Note that $N(t)$ is a random variable and represents the regression (or model) deviation.

The Least-Square Error (LSE) is probably the best known method for fitting linear models and by far the most widely used. However, LSE is not necessarily the optimum regression method if the deviation $N(t)$ is not of the Normal distribution. This is certainly true for typhoon series where the distribution is non-Gaussian. To overcome this problem and the potential outliers in the records, the least absolute deviations (LAD) method has been suggested (Bloomfield and Steiger, 1983). Gray et al. (1992, 1993) have used the LAD method to predict Atlantic seasonal hurricane activity for many years with success. Because this method has been tested for many years and is quite mature, it appears reasonable to extend the LAD method to the Western Pacific for seasonal typhoon prediction. In the following, we will describe the LAD method, the procedure used in the prediction, and the results of prediction.

2. Methods and Procedures for Making Seasonal TC Prediction

2.1 Least Absolute Deviations Regression

The basic idea of Least Absolute Deviation (LAD) regression problem is stated as below. Given n points $\{\underline{\mathbf{x}}_i, y_i\}, \underline{\mathbf{x}}_i \in R^k, i = 1,...,n$, the LAD fitting problem is to find a minimizer, $\hat{\underline{\mathbf{c}}} \in R^K$, of the distance function (Absolute Deviation):

$$f(\underline{\mathbf{c}}) = \sum_{i=1}^{n} \left| y_i - \sum_{j=1}^{K} c_j x_{ij} \right| = \sum_{i=1}^{n} \left| y_i - \langle \underline{\mathbf{c}}, \underline{\mathbf{x}}_i \rangle \right| = \left\| \underline{\mathbf{y}} - \mathbf{x}\underline{\mathbf{c}} \right\|_1 \tag{2}$$

where $\underline{\mathbf{y}} = [y_1, y_2,..., y_n]'$, $\mathbf{x} = [\underline{\mathbf{x}}_1, \underline{\mathbf{x}}_2,..., \underline{\mathbf{x}}_n]'$. $\underline{\mathbf{c}} = [c_1, c_2,..., c_K]'$
such that, $f(\hat{\underline{\mathbf{c}}}) = \min(f(\underline{\mathbf{c}}))$.

This problem is solvable because $f(c)$ is continuous and convex.

Due to the non-linearity of absolute deviations, solving LAD problem is no longer a linear problem. A variety of computer algorithms were developed based on the well-studied Linear Programming (LP) problem, since LAD and LP deal with the same kind of problem in the very basic nature. The LP problem in standard form is to find $\hat{\underline{x}}$ which maximize $f(\underline{x}) = \langle \underline{c}, \underline{x} \rangle$ subject to constraints $A\underline{x} \le \underline{b}$ and $\underline{x} \ge 0$ with a given vector $\underline{c}$, $\underline{b}$ and matrix A. It can be shown that any LAD curve-fitting can be expressed as an equivalent bounded feasible LP problem and vice versa.

We chose Bloomfield-Steiger (1983) algorithm to find the minimizer. The basic idea of this algorithm is to find the normalized steepest direction in each iteration of the algorithm. Let the current fit be $\underline{c}$, and $\underline{\delta_1}, \underline{\delta_2}, ..., \underline{\delta_K}$ be a set of directions along with the next iteration could move, the optimum descent direction is $\underline{\delta_p}$ along which

$$\min(f(\underline{c} + t\underline{\delta_p}), t \in R) = \min[\min(f(\underline{c} + t\underline{\delta_i})), i \le K] \tag{3}$$

the inner minimization over t in R. To find it, K weighted median calculations would need to be performed (one for each i in the right hand side of the equation (3)). The pseudo code for the Bloomfield-Steiger algorithm is listed in the Appendix A at the end of this report.

2.2 Detrending the TC series

The TC series from 1970 to 2002 are shown in Fig. 1a. The annual number of TCs (including both tropical storms and typhoons) in the vicinity of Taiwan refers to a region between 21° N-26°N and 119°E-125°E (courtesy of G.-H. Chen of the Central Weather Bureau). Clearly there is an upward trend in this series. In order to eliminate the non-stationarity of the observed TC series, we determined the linear long-term trend by using a least-square fit to a linear function of time t. Thus, the detrended data will have zero mean and take the following form:

$$y_{Detrend}(t) = y(t) - (\hat{\alpha} * t + \hat{\beta}) \tag{4}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are regression parameters and are chosen such that

$$\sum_{t=1}^{n} (y(t) - (\hat{\alpha} * t + \hat{\beta}))^2 = \min_{\alpha,\beta} (\sum_{t=1}^{n} (y(t) - (\alpha * t + \beta))^2)$$

The linear trend and detrended TC series in the vicinity of Taiwan are plotted in Fig. 1a and Fig. 1b, respectively.

2.3 Variance Analysis

The variance analysis method is a statistical technique to test the existence of hidden periods in a time series. The basic idea of this method is as follows:
In order to testify that if there is a hidden period p existing in a given series $x_i, i = 1,..., N$, where $2 \le p \le \left\lfloor \frac{N}{2} \right\rfloor$, one can divide the series into p groups $y_j, j = 1,..., p$, where the k-th element of the sub-group y_j is defined as

$$y_{jk} = x_{(k-1)*p+j}, j = 1,..., p, k = \begin{cases} \left\lfloor \frac{N}{p} \right\rfloor, if\ j > \mathrm{mod}(N, p) \\ \left\lceil \frac{N}{p} \right\rceil, otherwise \end{cases} \quad (5)$$

Then, one calculates the within-group variance and among-group variance given by

$$S_{within} = \sum (y_{jk} - \bar{y}_j)^2 / (N - p) \quad (6a)$$

$$S_{among} = \sum_{j=1}^{p} n_j (\bar{y}_j - \bar{x})^2 / (p - 1) \quad (6b)$$

where n_j represents the number of elements in j-th sub-group

Note these variances should be normalized by their respective degrees of freedom.

At last, we calculated the ratio S_{among} / S_{within} and compare it to the critical value determined by given confidence α (for example 95%), that is $F(\alpha, p-1, N-p)$ where F represents the F-distribution. If this ratio is larger than the critical value, it implies that the original series has a significant period of p. The Variance Analysis results for the raw TC data with 95% confidence interval are shown in Fig. 2; obviously 16 years and 4 years are two significant periods for this data set. The pseudo code of variance analysis to testify the existence of a period p of a given series x is shown in the Appendix B.

2.4 CLIPER

Once the hidden periods for the series are determined, we can use CLIPER method to find the periodical oscillation. The basic idea for CLIPER prediction is very simple: for the first significant period, say p, we can re-group the data as defined in formula (5). Then we can calculate each group's mean, which will be the prediction for this group.

In the case of having more than one significant period, say, for the second period, we just do the same procedure as we did for the first one, except that the new processing data is the "predictive residual", which is the original data subtracted by the CLIPER prediction from the first period. Same procedure can be applied to the third significant period and so on and so forth. The final prediction will be the sum of the predictions based on each period.

2.5 Finding relevant predictors

We first calculated the monthly correlations between the number of TCs in the vicinity of the Taiwan area and the index of HC45, Nino3.4, and SOI from January to May respectively. The HC45 index series is the westernmost location (in longitudes) of the 500-hPa subtropical ridge over the Western North Pacific. The Nino3.4 index is the standard El Nino index and the SOI is the standard Southern Oscillation Index. Results suggest that all of these correlations are not significant with values lower than 0.1. We then calculated the correlations between the number of TCs and the monthly SST (sea surface temperatures) and SLP (sea level pressures) over the Pacific Ocean, also from January to May. It turns out that the SST or SLP in some areas near Taiwan in May are highly correlated with the number of TCs. The correlation map between the number of TCs in the vicinity of Taiwan and the May SST over the Pacific Ocean is shown in Fig. 3a. Note that the grid resolution for SST is 10° by 10°. An area of strong correlation is found to the southeast of Taiwan with a maximum around 0.49 near 0°, 140°E in the warm pool of the western Pacific. The correlation map between the number of TCs and the May SLP over the Pacific Ocean is shown in Fig. 3b. For the SLP, the horizontal grid resolution is 2.5° by 2.5°. The highest negative correlation (-0.42) is noted at 15°N, 130°E, also near Taiwan. Interestingly, the juxtaposition of the maximum correlation found in Figs. 3a and 3b suggests a Rossby-type response of atmosphere to equatorial heating as demonstrated in Gill's model.

Based on the results of correlation analyses, we chose the point, where the SST or SLP are most highly (in absolute value) correlated with the number of TCs as our predictor variables. Time series of SST and SLP in May in the chosen location as well TC series near Taiwan are listed in Table 1.

3. Summary

The overall approach for our prediction method is:

1. Find the linear long-term trend using a Least-Square fit to a linear function of time t for time series of TCs and use the detrended data in the following analysis.
2. With detrended data, find the short-term oscillation periods using the Variance Analysis method.
3. With the found significant periods, we apply CLIPER method to get the prediction for each year (in cross-validation sense).
4. We then try to find the relevant predictor regions from the SST and SLP fields over the Pacific. This is determined by examining the month-to-month correlations between the TC counts and the SST as well as the SLP fields.
5. Our prediction of the number of TCs for each year is based on the LAD technique with SST and SLP records in May and the CLIPER prediction. This predicted number is then added to the long-term linear trend to arrive at the final value.

As discussed in section 2.5, we found the highly correlated regions in terms of SST and SLP and used them as the predictor data. We then applied the cross-validation technique to test the predictability of TC from our method. In cross-validation, we

repeatedly omitted one observation from the entire data (1970-2002), reconstructed the model, and then made forecast for the omitted case (Chu and He, 1994; Yu et al. 1997). The cross-validation provides an unbiased estimate of forecast skill. The cross-validation results are shown in Fig. 4. The linear correlation between the cross-validated predictions and the raw data is 0.58, significant at the 1% level. Another measure of forecast success is the reduction in variance of forecast error. The variance of prediction error through cross-validation (1.79) is dramatically decreased compared to the variance of the observations (2.56).

Appendix:

A. Pseudo Code for the Bloomfield and Steiger Algorithm:

[1] Accept $\{\underline{\mathbf{x}}_i, y_i\}, \underline{\mathbf{x}}_i \in R^k, i = 1,...,n$

[2] $A \leftarrow \begin{pmatrix} \mathbf{X} & \underline{\mathbf{y}} \\ \mathbf{I}_k & \underline{\mathbf{0}} \end{pmatrix}$, where $\mathbf{X} = [\underline{\mathbf{x}}_1,...,\underline{\mathbf{x}}_n]'$, $\underline{\mathbf{y}} = [y_1,...,y_n]'$, $\mathbf{I}_k$ is an $k \times k$ identity matrix and $\underline{\mathbf{0}} \in R^k$

[3] $j \leftarrow 0$

[4] $\mathbf{Z} \leftarrow \{i, 1 \le i \le n : A_{i,k+1} = 0\}$

[5] for m = 1 to k

$$g_m = -\sum_{\mathbf{Z}} |A_{i,m}|$$

$$h_m = \sum_{\mathbf{Z}'} A_{i,m} sign(A_{i,k+1})$$

$$f_m = \max(g_m - h_m, g_m + h_m) / \sum_{i=1}^{n} |A_{i,m}|$$

end

[6] $p \leftarrow \arg\min_i \{f_i, i = 1,...,n\}$

if $f_p \le 0$ go to [10]

[7] $t \leftarrow A_{q,k+1} / A_{qp}$, the weighted median of $\{A_{i,k+1} / A_{i,p}, \text{i} = 1,...\text{n}\}$ with weights $|A_{i,p}|$

[8] for matrix $\mathbf{A}$:

column $p \leftarrow$ column p / A_{qp}

column $i \leftarrow$ column $i - A_{qi}$ * column p

[9] $j \leftarrow j + 1$ go to [4]

[10] $\underline{\mathbf{c}}_i \leftarrow -A_{n+i,k+1}, \text{i} = 1,...,\text{k}$

B. Pseudo Code of Variance Analysis

[1] regroup the original serials x with length N into p sub-groups y_j as defined in formula (2)

[2] calculate the mean of each group y_j, $\bar{y}_j$

[3] calculate the grand mean of the serials x, $\bar{x}$

[4] calculate the within-group variance $S_{within} = \sum (y_{jk} - \bar{y}_j)^2 / (N - p)$

[5] calculate the among-group variance $S_{among} = \sum_{j=1}^{p} n_j (\bar{y}_j - \bar{x})^2 / (p - 1)$, where n_j represents the number of elements in j-th sub-group

[6] compare the F-ratio S_{among} / S_{within} to the critical value $F(\alpha, p-1, N-p)$, where α is the confidence level for this critical value. If the F-ratio bigger than the critical value, it means that this hidden period exists

References:

Bloomfield, P., and W. L. Steiger, 1983: *Least Absolute Deviations: Theory, Applications, and Algorithm*, Birkhauser, Boston, 349 pp.

Chu, P.-S., and Y. He, 1994: Long-range prediction of Hawaiian winter rainfall using canonical correlation analysis. *Int. J. Climatol.*, 14, 659-669.

Gray, W.M., C.W. Landsea, P.W. Mielke, Jr. and K.J. Berry, 1992: Predicting Atlantic basin seasonal hurricane activity 6-11 months in advance. *Wea. Forecasting*, 7, 440-455.

Gray, W.M., C.W. Landsea, P.W. Mielke, Jr. and K.J. Berry, 1993: Predicting Atlantic basin seasonal tropical cyclone activity by 1 August. *Wea. Forecasting*, 8, 73-86.

Yu, Z.-P., P.-S. Chu, and T.A. Schroeder, 1997: Predictive skills of seasonal to annual rainfall variations in the U.S. affiliated Pacific islands: Canonical correlation analysis and multivariate principal component regression approaches. *J. Climate*, 10, 2586-2599.

Table and Figure:

Table 1: The data of the LAD Prediction

Figure 1a: Time series of the annual number of TCs in the vicinity of Taiwan area and its linear trend (LSE fit with respect to time)

Figure 1b: Detrened TC series by subtracting the long term trend from the original series

Figure 2: Variance Analysis of TC series in the vicinity of Taiwan area with 95% confidence.

Figure 3a: Correlation map between the number of TCs near Taiwan and the May SST of the same year over the Pacific Ocean.

Figure 3b: Correlation map between the number of TCs near Taiwan and the May SLP of the same year over the Pacific Ocean.

Figure 4: Cross-validated TC count prediction based on the LAD method with predictors of SST, SLP and CLIPER.

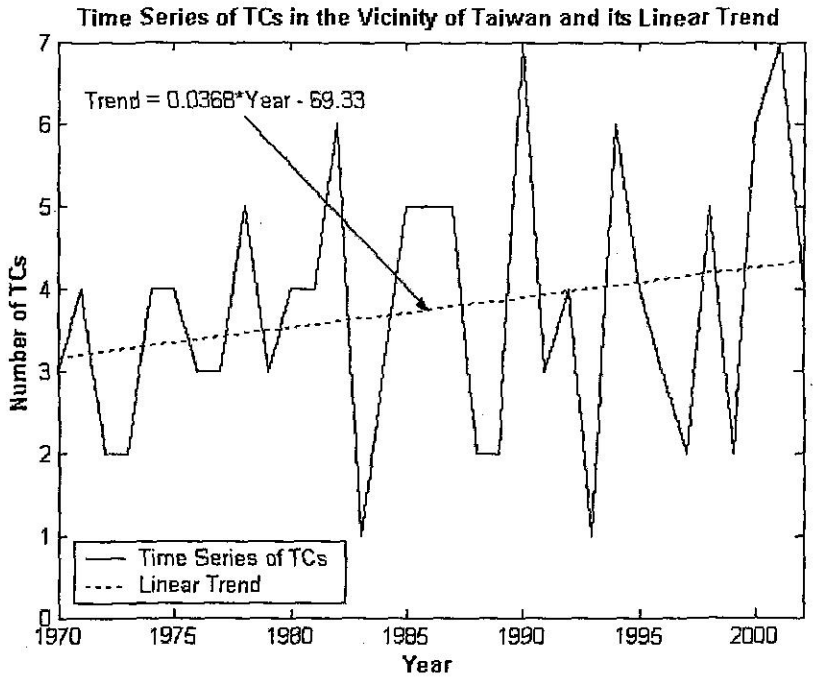
Time Series of TCs in the Vicinity of Taiwan and its Linear Trend
Trend = 0.0368*Year - 69.33
Number of TCs
Year
Time Series of TCs
Linear Trend
0
1
2
3
4
5
6
7
1970
1975
1980
1985
1990
1995
2000

Fig. 1a

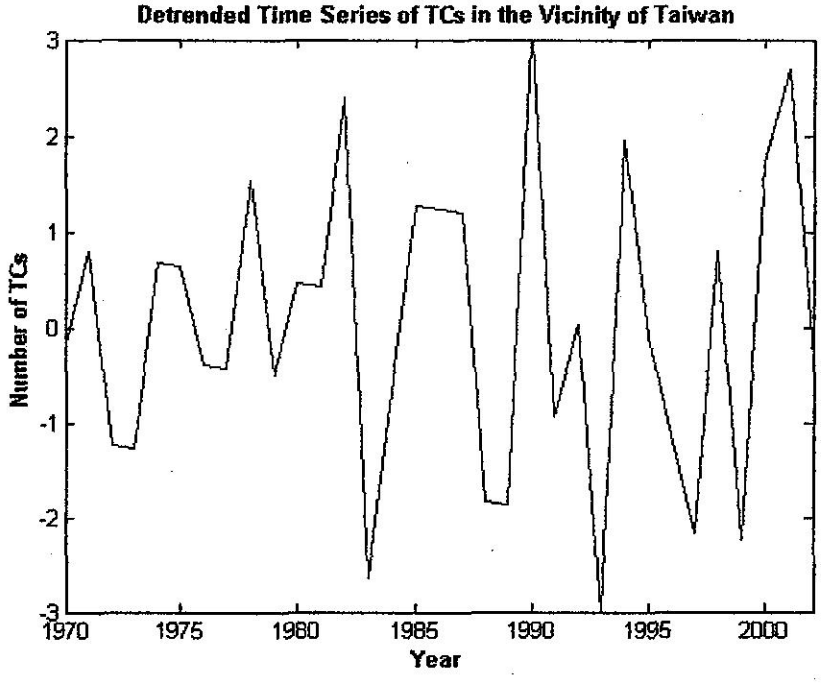
Detrended Time Series of TCs in the Vicinity of Taiwan
Number of TCs
Year
3
2
1
0
-1
-2
-3
1970
1975
1980
1985
1990
1995
2000

Fig. 1b

交通部中央氣象局委託研究計畫成果報告

利用衛星（TRMM）觀測降雨相關資料評估及改進中央氣局全球模式之物理過程
Use of TRMM Information for Evaluation and Improvement of Moist Physical Process in the Global Forecast System at CWB

計畫類別：□國內　　☑國外
計畫編號：MOTC-CWB- 92 - 3M-06
執行期間： 92 年 2 月 1 日 至 92 年 12 月 31 日
計畫主持人：李瑞麟
執行單位：Real One Connection

中華民國九十二年十二月