A Bayesian Analysis of the Change Point Problem for Directional Data using SIR

Ashis SenGupta
Indian Statistical Institute, Kolkata

and

Arnab Kumar Laha
Indian Institute of Management, Ahmedabad

ABSTRACT

In this paper we discuss a simple fully Bayesian analysis of the change point problem for the directional data in the parametric framework with circular normal distribution as the underlying distribution. We discuss the problem of detecting change in the mean direction of the circular normal distribution when the concentration parameter is unknown. Beginning with proper priors for all the unknown parameters, the sampling-importance-resampling (SIR) technique is used to obtain the posterior marginal distribution of the change point. The method is illustrated using the wind data (Weijer's et. al.(1995)). The method can be adapted to a variety of situations involving both angular and linear data and can be used with profit in the context of statistical process control in Phase I of control charting and also in Phase II in conjunction with control charts.

Introduction

The onset of an abrupt change, which usually leads to poor quality products, is a phenomenon which is common in the industrial context. Several of the techniques discussed commonly under the heading statistical process control (SPC) are for early detection of any sudden change in the process parameters. Some examples are Shewhart control charts and its variants, Cusum charts, EWMA charts etc. The primary aim of charting in the context of SPC is to detect the occurrence of such a sudden change in the value of a process parameter or quality characteristic as quickly as possible. Woodall and Montgomery (1999) consider change-point estimation as an important research area in SPC. Stoumbos et. al. (2000) advocates a greater synthesis of theoretical change point and applied SPC literatures. It is of interest to note that the formulation of the change point problem as given by Page (1955) does not consider any possible correlation between the successive time sequenced observations which makes it markedly different from the usual time series models. The problem of detecting whether at all there is a point of abrupt change in a given dataset thereafter the problem of detecting the change point has received a lot of attention. Over the last fifty years the problem has been examined extensively for the case of linear data. The two main streams of work pertain to the parametric set-up with normal distribution as the underlying distribution and the non-parametric set-up see eg, Chernoff and Zacks (1964), Hinkley (1970), Sen and Srivastava (1973, 1975a, 1975b), Chen and Gupta (2000) etc. The change-point problem assumes great practical significance in the context of many real-life encounters with directional data, e.g. in applications relating to meteorological data like

wind directions, movements of icebergs, propagation of cracks, biological and periodic phenomena (like circadian rhythm) etc. Lombard (1986) initiated work in the context of directional data in the non-parametric frame work. As in the linear case in the angular case also, the change point problem has an important role during the Phase I (retrospective analysis phase) of control charting. The construction of control charts for angular variables and definitions of some associated process capability indices are discussed in Laha (2002, 2004).

In this paper we discuss the change point problem for angular variables in the parametric framework with circular normal distribution (also called von Mises distribution), the most popular distribution for directional data, with probability density function given by

$$f(\theta;\mu,\kappa)=\frac{1}{2\pi I_0(\kappa)}e^{\kappa\cos(\theta-\mu)},\ 0\le\mu<2\pi,\ \kappa>0$$

where $I_0(\kappa)$ is the modified Bessel function of order 0. If the angular random variable Θ follows the circular normal distribution with parameters μ and κ then we write $\Theta\sim CN(\mu,\kappa)$. The parameter μ is called the mean direction and the parameter κ is called the concentration parameter. The circular normal distribution is a symmetric unimodal distribution with mode at μ. For more details on the circular normal distribution the reader may look into Jammalamadaka and SenGupta (2001).

Gibbs Sampler

One of the most popular tools for sampling based inference is the Gibbs sampler. In this section we discuss the construction of a Gibbs sampler for the change point problem for the mean direction of the circular normal distribution and point out some difficulties encountered in using this approach. Damien and Walker (1999) demonstrate the use of strategic latent variables to construct a Gibbs sampler for analysis of circular data, having the circular normal distribution as its underlying distribution, which has all full conditional distributions of known type. We use an approach similar to theirs to construct a Gibbs sampler for the change point problem with circular normal distribution as the underlying distribution. The details are given below.

Let the joint prior distribution of $(\mu_1, \mu_2, \kappa, r)$ be

$$\pi(\mu_1, \mu_2, \kappa, r) \propto p(r) I_0(\kappa)^{-c} e^{\kappa[R_0 \cos(\mu_1 - \mu_0)]} \{e^{\kappa[R_1 \cos(\mu_2 - \mu_0^*)] I(r<n)}\} \quad \ldots(1)$$

where $p(r)$ is a probability mass function on {1,2,…,n}, $I(r < n)$ is the indicator function of the event $\{r < n\}$, and R_0, R_1, μ_0, μ_0^* , c are all constants. In what follows $I(H)$ will denote the indicator function of the set H. The posterior distribution of $(\mu_1, \mu_2, \kappa, r)$ is then

$$\pi_p(\mu_1, \mu_2, \kappa, r \mid \theta_1, \ldots, \theta_n) \propto p(r) I_0(\kappa)^{-m} e^{(\kappa R_{r1} \cos(\mu_1 - \mu_{r1}))} . [e^{(\kappa R_{r2} \cos(\mu_2 - \mu_{r2}))} I(r < n) + I(\mu_1 = \mu_2) I(r = n)] \quad \ldots(2)$$

where *m=c+n*, R_{r1} and μ_{r1} are obtained by solving simultaneously the system of equations $R_{r1} \cos \mu_{r1} = R_0 \cos \mu_0 + \sum_{i=1}^{r} \cos \theta_i$ and $R_{r1} \sin \mu_{r1} = R_0 \sin \mu_0 + \sum_{i=1}^{r} \sin \theta_i$, R_{r2} and μ_{r2} are obtained by solving simultaneously the system of equations

$R_{r2}\cos\mu_{r2} = R_1\cos\mu_0^* + \sum_{i=r+1}^{n}\cos\theta_i$ and $R_{r2}\sin\mu_{r2} = R_1\sin\mu_0^* + \sum_{i=r+1}^{n}\sin\theta_i$. We now introduce latent variables t,ν and w and define the joint distribution of $(\mu_1,\mu_2,\kappa,r,t,\nu,w)$ to be

$$f(\mu_1,\mu_2,\kappa,r,t,\nu,w) \propto p(r)e^{-R_{r1}\kappa}I(t < e^{kR_{r1}[1+\cos(\mu_1-\mu_{r1})]})[e^{-R_{r2}\kappa}I(\nu < e^{kR_{r2}[1+\cos(\mu_2-\mu_{r2})]})I(r<n) + I(\nu<1)I(\mu_1=\mu_2)I(r=n)].w^{m-1}e^{-wI_0(\kappa)} \quad \ldots(3)$$

Note that the marginal distribution of (μ_1,μ_2,κ,r) is same as that given in (2) above. Now since $I_0(\kappa) = \sum_{k=0}^{\infty}\lambda_k\kappa^{2k}$ where $\lambda_k = (k!)^{-2}0.5^{2k}$. Therefore the above joint distribution can be written as

$$f(\mu_1,\mu_2,\kappa,r,t,\nu,w) \propto p(r)e^{-R_{r1}\kappa}I(t < e^{kR_{r1}[1+\cos(\mu_1-\mu_{r1})]})[e^{-R_{r2}\kappa}I(\nu < e^{kR_{r2}[1+\cos(\mu_2-\mu_{r2})]})I(r<n) + I(\nu<1)I(\mu_1=\mu_2)I(r=n)].w^{m-1}\prod_{k=0}^{\infty}e^{-w\lambda_k\kappa^{2k}} \quad \ldots(4)$$

Introducing the latent variables $\mathbf{u} = (u_1,u_2,\ldots)$ and x we define the joint distribution of $(\mu_1,\mu_2,\kappa,r,t,\nu,w,x,\mathbf{u})$ as

$$f(\mu_1,\mu_2,\kappa,r,t,\nu,w,x,\mathbf{u}) \propto p(r)e^{-R_{r1}\kappa}I(t < e^{kR_{r1}[1+\cos(\mu_1-\mu_{r1})]})[e^{-R_{r2}\kappa}I(\nu < e^{kR_{r2}[1+\cos(\mu_2-\mu_{r2})]})I(r<n) + I(\nu<1)I(\mu_1=\mu_2)I(r=n)].I(x<w^{m-1})e^{-w}\prod_{k=1}^{\infty}I(u_k < e^{-w\lambda_k\kappa^{2k}}) \quad \ldots(5)$$

It is easy to see that the marginal distribution of (μ_1,μ_2,κ,r) is same as that given in (2) above. To implement the Gibbs sampler we need the full conditional distributions of each of the variables given the rest. These are given below. Let *U(a,b)* denote the uniform distribution on the open interval (a,b).

(i) The full conditional distribution of x is $U(0,w^{m-1})$.

(ii) The full conditional distribution of t is $U(0, \kappa R_{r1}[1+\cos(\mu_1 - \mu_{r1})])$.

(iii) The full conditional distribution of μ_1 is $U(a,b)$ where

$A = (a,b) = \{\mu : \cos(\mu - \mu_{r1}) > (R_{r1}\kappa)^{-1} \ln t - 1\}$

(iv) The full conditional distribution of ν is

$U(0, \kappa R_{r2}[1+\cos(\mu_2 - \mu_{r2})])$ if $r < n$ and $U(0,1)$ if $r = n$.

(v) The full conditional distribution of μ_2 is $U(c,d)$ where

$A = (c,d) = \{\mu : \cos(\mu - \mu_{r2}) > (R_{r2}\kappa)^{-1} \ln \nu - 1\}$ if $r < n$ and $\mu_2 = \mu_1$ if $r = n$.

(vi) The full conditional distribution of $u_k, k = 1,2,...$ is $U(0, e^{-w\lambda_k \kappa^{2k}})$.

(vii) The full conditional density of w is given by

$$f_c(w) \propto e^{-w} I(x^{\frac{1}{m-1}} < w < \inf_k (-\lambda_k \kappa^{2k})^{-1} \log u_k)$$

(viii) The full conditional density of κ is given by

$f_c(\kappa) \propto e^{-(R_{r1}+R_{r2})\kappa} I(C)$ if $r < n$ and $f_c(\kappa) \propto e^{-R_{r1}\kappa} I(D)$ if $r = n$

where

$$C = \{\kappa > \frac{\log t}{R_{r1}(1+\cos(\mu_1 - \mu_{r1}))}\} \cap \{\kappa > \frac{\log \nu}{R_{r2}(1+\cos(\mu_2 - \mu_{r2}))}\}$$
$$\cap \{\kappa < \inf_k [(-w\lambda_k)^{-1} \log u_k]^{\frac{1}{2k}}\}$$

and

$$D = \{\kappa > \frac{\log t}{R_{r1}(1+\cos(\mu_1 - \mu_{r1}))}\} \cap \{\kappa < \inf_k [(-w\lambda_k)^{-1} \log u_k]^{\frac{1}{2k}}\}$$

(ix) The full conditional probability mass function of r is given by

$$p_c(r) = \frac{f(\mu_1, \mu_2, \kappa, r, t, \nu, w, x, \mathbf{u})}{\sum_{r=1}^{n} f(\mu_1, \mu_2, \kappa, r, t, \nu, w, x, \mathbf{u})}$$

The Gibbs sampler can now be implemented quite easily.

In the above we have tried to look at the Gibbs sampler along the same lines as of Damien and Walker (1999). However, we have not yet been able to establish the necessary conditions for the ergodicity of the chain. We feel that an in depth theoretical study will be required to verify the existence of a stationary distribution for this chain. In this paper, we thus divert form that theoretical route and propose another approach which can be easily implemented in practice.

SIR in Change Point Problem

Let $\Theta_1, \Theta_2, ..., \Theta_n$ be independent observations. We are interested to test the hypothesis

$$H_0 : \Theta_1, \Theta_2, ..., \Theta_n \text{ are i.i.d. } \mathrm{CN}(\mu_0, \kappa)$$

against the alternative

$$H_1 : \Theta_1, ..., \Theta_r \text{ are i.i.d. } \mathrm{CN}(\mu_0, \kappa) \text{ and } \Theta_{r+1}, ..., \Theta_n \text{ are i.i.d. } \mathrm{CN}(\mu_1, \kappa), \mu_1 \neq \mu_0 \text{ for some } r,$$
$$1 \leq r \leq n-1.$$

The parameters μ_0, μ_1, κ and r are all unknown. Laha (2001) analyses this problem from a frequentist perspective and proposes an NR-type test for this problem. In this paper we take the Bayesian route using the Sampling Importance Resampling (SIR) technique to obtain the posterior distribution of the change point. The use of SIR in the context of change point problem even in the linear set-up has not been reported previously in the literature.

In the SIR methodology a prior (joint) distribution is specified for the unknown parameters. Samples are then drawn from this prior distribution and the likelihood is calculated for each such sample. The prior is then resampled using the likelihoods as weights. The resample constitutes a sample from the posterior (joint) distribution of the parameters. The posterior of each of the unknown parameters can be obtained by finding the appropriate marginal distribution. For an elegant discussion of SIR methodology the reader may look into Smith and Gelfand (1992).

We apply SIR with the prior joint distribution of μ_0, μ_1, κ and r taken to be the product of the four individual prior distributions. The prior distribution on μ_0 and μ_1 are both taken to be circular uniform, that of κ is taken as exponential with rate 1 and that of r is taken as discrete uniform{*1,2,…,n*}.
Thus the joint prior distribution of μ_0, μ_1, κ and r is

$$p(r, \mu_0, \mu_1, \kappa) = \frac{e^{-\kappa}}{4n\pi^2}, r = 1,..., n, 0 \le \mu_0, \mu_1 < 2\pi, \kappa > 0$$

It is to be noted that $r = n$ implies that there is no change point in the data set.
The likelihood function is

$$L(\mu_0, \mu_1, \kappa, r; \theta_1, \theta_2, ...\theta_n) = \frac{1}{(2\pi I_0(\kappa))^n} \exp\left[\kappa\left\{\sum_{i=1}^{r}\cos(\theta_i - \mu_0) + \sum_{i=r+1}^{n}\cos(\theta_i - \mu_1)\right\}\right] \text{if r} \neq 0$$

and

$$L(\mu_0, \mu_1, \kappa, r; \theta_1, \theta_2, ...\theta_n) = \frac{1}{(2\pi I_0(\kappa))^n} \exp\left[\kappa\left\{\sum_{i=1}^{n} \cos(\theta_i - \mu_0)\right\}\right] \text{if r} = 0$$

For each prior sample, $(\mu_{0i}, \mu_{1i}, \kappa_i, r_i)$ we attach the weight $q_i = \frac{L((\mu_{0i}, \mu_{1i}, \kappa_i, r_i; \boldsymbol{\theta})}{\sum_{j=1}^{m} L((\mu_{0j}, \mu_{1j}, \kappa_j, r_j; \boldsymbol{\theta})}$ where $\boldsymbol{\theta} = (\theta_1, \theta_2, ...\theta_n)$ and *m* is the number of prior samples drawn. We then resample the prior sample, using the weights q_i, *k* times to obtain a sample of size *k* from the posterior distribution of μ_0, μ_1, κ and r.

Example

In this section we apply the above methodology to a dataset of wind directions given by Weijers et. al. (1995). They investigated the horizontal perturbation wind field within thermal structures encountered in the atmospheric surface layer boundary. A field experiment with four sonic anemometers on the vertices and one in the centroid of a square was performed to obtain the necessary dataset. Structures were selected on a typical ramp-like appearance in the temperature time series. Altogether a set of 47 ``ramps'' was obtained. Ensemble averages of turbulent temperature and horizontal and vertical velocities were constructed using conditional sampling and block averaging followed by a compositing technique. We are interested in the behaviour of the direction of the horizontal wind field as recorded by the anemometer at the centroid for the 32 bins after the ensemble averaging procedure. The method of construction of the bins and the ensemble

averaging procedure show that the bins have a temporal ordering. We retrieved the actual data from the graphical representation presented in the paper. Exploratory data analysis conducted on this dataset using the Changeogram and circular difference table developed in Laha(2001) indicated the presence of two change points.

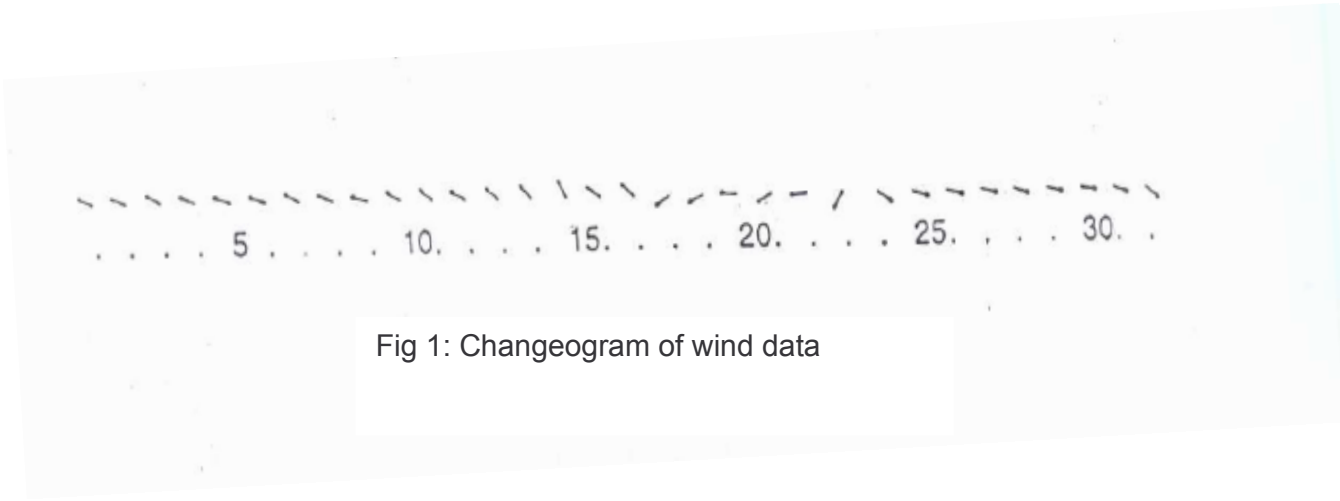

Fig 1: Changeogram of wind data

In this paper we obtain the posterior distribution of the first change point using SIR. A Changeogram displays pictorially in terms of directed arrows, each of unit length, the direction in terms of the angle as given by the corresponding observation. The circular difference table is constructed by considering the change of direction between two successive observations. For eg. if θ_t and θ_{t+1} are the two successive observations then we consider the difference to $\min(|\theta_t - \theta_{t+1}|, 2\pi - |\theta_t - \theta_{t+1}|)$. The Changeogram and the circular difference table has been incorporated in DDSTAP (SenGupta, 1996), a statistical package for the analysis of directional data. From the Changeogram (Figure 1) one notices that there are possibly two change points one around 17 and the other around 23. Since we are primarily interested in the posterior distribution of the first change point the data beyond the second change point

were omitted. Thus, we consider only the observation numbers 1 to 22 for our analysis. The marginal posterior distribution of the change point based on 69511 samples from the posterior joint distribution is given below (Fig. 2):

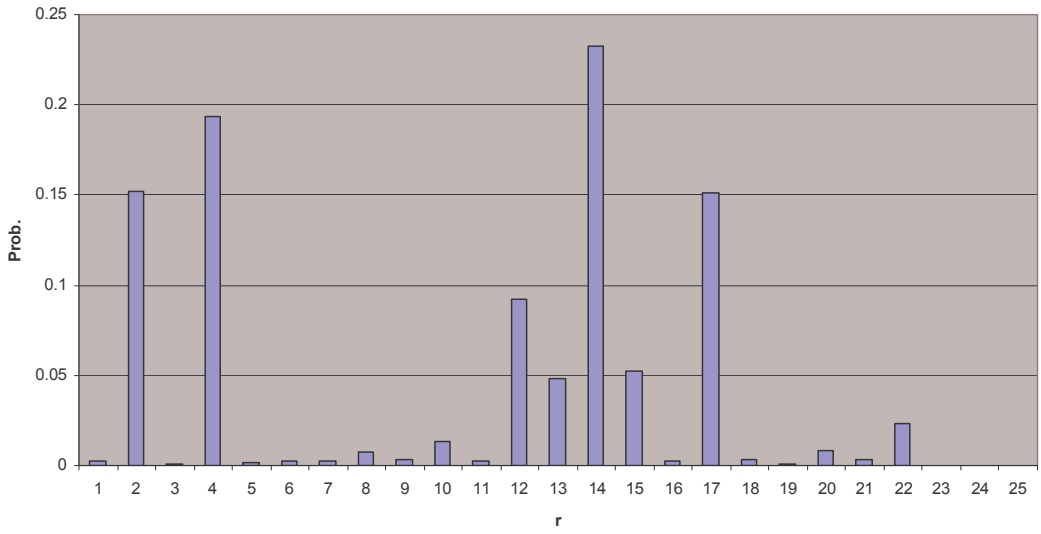

Fig 2 : Posterior distribution of change point for wind data (Obsns. 1-22)

From the above posterior marginal distribution we see that the posterior probability of no change is quite small. The posterior median is at 12, the posterior mode is at 14, and the posterior mean is 10.57. The approximately 90% posterior credible set of the change point is {2, 4, 10, 12, 13, 14, 15, 17}. The circular difference table for the wind data (Observations 1-22) is given below (Table 1).

t	Diff	t	Diff	t	Diff
1	2	8	6	15	31
2	7	9	13	16	8
3	7	10	8	17	73
4	3	11	10	18	8
5	1	12	10	19	29
6	8	13	9	20	32
7	4	14	19	21	22

Table 1 : Circular Difference Table for Wind Data (Observations 1-22).

The above dataset is analysed from a frequentist viewpoint in Laha(2001). It is reported therein that when Lombard's non-parametric test (Lombard, 1986) for single change point is applied to this dataset it indicates the presence of a change point at 5% level of significance and identifies 13 as the change point. When the NRTT, a parametric test derived under the assumption of circular normality in Laha(2001), is applied to the same dataset it also indicates the presence of change point at 5% level of significance and identifies 17 as the change point. It is interesting to note that both 13 and 17 are included in the 90% posterior credible set for change point mentioned above.

Concluding Remarks

SIR provides a very simple approach for identification of possible location of a change point through examination of the posterior marginal distribution of the change point. The method can prove to be an extremely useful tool in the context of Phase I studies for SPC. Even in Phase II when on-line control charting is in progress these methods can be used very profitably to reduce the work-load of the process engineers when a control chart indicates an out-of-control condition. The process engineers can look at the points with high posterior probability of change more intensely than spreading their effort equally over all time points since the time when the last out of control conditions were detected. The method though discussed in the context of directional data in this paper can be easily applied to other set-ups like normal or exponential. Also the method can be easily adapted for the case when the data set is suspected to have multiple change points.

References:

1. Chen, J. and Gupta, A. K. (2000): *Parametric Statistical Change Point Analysis*, Birkhauser, Boston.
2. Chernoff, H. and Zacks, S. (1964) : Estimating the current mean of a normal distribution which is subject to changes in time, *Annals of Mathematical Statistics*, 35, 999-1018.
3. Damien, P. and Walker, S. (1999) : A full Bayesian analysis of circular data using the von Mises distribution, *The Canadian Journal of Statistics*, 27, 2, 291-298.
4. Jammalamadaka, S. R and SenGupta, A (2001): *Topics in Circular Statistics*, World Scientific, Singapore
5. Laha, A. K. (2001): *Slippage and Change Point Problems with Directional Data*, Ph.D. thesis, Indian Statistical Institute, Kolkata, India.
6. Laha, A. K. (2002): Control charts for angular observations, *Statistical Methods for Quality Improvement* (Editors - A.B. Aich, A. Chatterjee and A. K. Chattopadhyay), 39-45, IAPQR, Kolkata.
7. Laha, A. K. (2003): Process capability Indices for angular data, *Statistical Methods*, 5(2), 31-40.
8. Lombard, F. (1986): The change point problem for angular data : A nonparametric approach, *Technometrics,* 28, 391-397.
9. Neyman, J. (1959): Optimal asymptotic tests for composite statistical hypothesis, *Probability and Statistics (The Harald Cramer Volume)*, Almqvist and Wiksell, Uppsala, Sweden, 213-234.

10. Page, E. S. (1955) : A test for a change in a parameter occurring at an unknown time point, *Biometrika*, 42, 523-526.
11. Sen, A. K. and Srivastava, M. S. (1973): On multivariate tests for detecting change in mean, *Sankhya*, A, 35, 173-185.
12. Sen, A. K. and Srivastava, M. S. (1975a): On tests for detecting a change of mean, *Annals of Statistics*, 3, 98-108.
13. Sen, A. K. and Srivastava, M. S. (1975b): Some one-sided tests for change in level, *Technometrics*, 17, 61-64.
14. SenGupta, A. (1996) : DDSTAP – A statistical package for directional data analysis, Indian Statistical Institute.
15. Smith, A.F.M and Gelfand, A.E. (1992): Bayesian statistics without tears: A sampling-resampling perspective, *The American Statistician*, 46, 2, 84-88.
16. Stoumbos, Z.G, Reynolds, M.R., Ryan, T.P. and Woodall, W. H. (2000): The state of statistical process control as we proceed into the 21st century, *Journal of the American Statistical Association*, 95, 451, 992-998
17. Weijers, E.P., Van Delden, A., Vugts, H.F., and Meesters, A.G.C.A. (1995): The composite horizontal wind field within convective structures of the atmospheric surface layer, *Journal of the Atmospheric Sciences*, 52, 3866-3878.
18. Woodall, W. H. and Montgomery, D. C. (1999): Research issues and ideas in statistical process control, *Journal of Quality Technology*, 31, 4, 376-386.