UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

Comunicaciones en Estadística
Diciembre 2010, Vol. 3, No. 2

Use of Additive Models to Assess the Time and Space Variability in 3D Data

Uso de modelos aditivos para evaluar la variabilidad espacial y temporal en datos 3D

Francisco Andrés Rincón Rodríguez[a]
francisco.rincon@hscb.com

Abstract

This paper is aiming to identify the presence of variability over time and space for water temperature in Santa Marta, Colombia. The modeling process consider an approach through linear models, the use of additive models as an alternative to capture nonlinear patterns, evaluation of the need for a non parametric effect for each of the covariates and finally, a diagnostic test for the residuals to assess the need to include a covariance structure over time and/or space.

***Key words*:** Additive Models, Covariance Structure Over Time and Space, Non-parametric Effect, Variability.

Resumen

Este articulo está dirigido a identificar la presencia de variabilidad sobre tiempo y espacio para la temperatura del agua en Santa Marta, Colombia. El proceso de modelamiento considera una aproximación a través de un modelo lineal, el uso de modelos aditivos como una alternativa para capturar patrones no lineales, evaluación de la necesidad de un efecto no paramétrico para cada una de las covariables y finalmente un diagnóstico sobre los residuales para valorar la necesidad de incluir una estructura de covarianza sobre tiempo y/o espacio.

***Palabras clave*:** efecto no paramétrico, estructura de covarianza sobre el tiempo y espacio, modelos aditivos, variabilidad.

1. Introduction

Nowadays the analysis of environmental information has become an important topic in the agenda for governments and academics, trying to understand changes

[a]Subgerente de riesgo. HSBC Colombia.

in patterns and how these changes affect us. This represents a challenge from a statistical point of view to provide an accurate representation and understanding of these changes, considering the presence of spatial and temporal components besides of the large number of variables involved.

These reasons make it harder to identify the presence, the magnitude and the factors involved in a change. For this reason, it is necessary to move from classical approaches to modern statistical methodologies and in some cases to use more than one methodology simultaneously.

The motivation of this article is to provide a methodology that is quite useful for environmental studies where it is necessary to analyse temporal and spatial variability in a combined way to capture both sources of variability.

This paper is aiming to assess changes in the temperature of the sea in Santa Marta, Colombia, using information collected unevenly over time and space at 17 places from August 2001 to January 2006. The goal is to identify variability over time and space using the date and location where the temperature was measured.

In section 2 a summary of previous works show different approaches that may lead to a clear understanding of the behavior of sea temperature as variable of interest or as a covariate. In section 3 a description of Santa Marta, by geographical location, temperature and economical activities is presented. Section 4 corresponds to a data description, showing the location of each of the 17 sites as well as the time series by site over the observed period. Section 5 which represent the main part of this article describes the modeling process. As a first approach a linear model is used to capture variability over time and space simultaneously, a diagnostic check for the residuals was carried out, confirming that a linear model is not a suitable approach. As an alternative, the use of additive models provides an useful approach, where the smooth functions are not restricted in shape, allowing us to capture non linear relationships. In section 6 a test to confirm the need for a non parametric effect for each variable was carried out, as well as a sensitivity analysis to assess changes in the conclusions under different degrees of freedom. Once the final model has been identified, a diagnostic check for the residuals allow us to evaluate the need to include a covariance structure over time and/or space. Finally section 7 includes a summary, explaining the findings in respect to the variability over time and space for sea temperature in Santa Marta, Colombia, as well as further steps which could lead to include more covariates and applied additive models to understand different problems in this area.

2. Previous Works

Previous papers which may be mentioned as a reference are Lewis & Ray (1997), which explains the use of time series adaptive regression splines (TSMARS) to fit a model for daily sea surface temperature, measured off the California coast, dealing with non linear effect and long range dependence, G.Bernal et al. (2006), which corresponds to the analysis of monthly sea surface time series in the colombian

caribbean coast at 8 different places and Beare & Reid (2002), which explains the use of generalized additive models to investigate the spatial-temporal change in spawning activity by Atlantic mackerel.

Looking for an alternative approach, this paper is aiming to show the analysis of spatial and temporal components simultaneously rather than in a marginal way (Rincón 2010). This approach is useful if the purpose is to capture both sources of variability and in addition, to capture tendencies properly, when the information is collected unevenly.

3. Santa Marta, Colombia

Santa Marta, is located in the North-East of the Caribbean coast (Figure 1), with an average temperature of 34^o, a maximum and minimum of 39^o and 31^o Celsius degrees respectively. The main economical activity is based on tourism, commerce, harbor activities and fishing.

4. Data Description

The data set used for this analysis, corresponds to temperature of the water collected unevenly, since August 2001 to January 2006 at 17 sites for a total number of 65 observations. The information was obtained from dives provided by the scuba diving school Naowa.

Figure 2 depicts the location of the 17 places with four points of reference Isla Aguja, Taganga, Punta Granate and Bahía Concha (left hand side) and the code assigned to each one of the 17 sites where X (direction North-South) and Y (direction East-West) corresponds to artificial coordinates in kilometers for the 17 sites (right hand side). Figure 3 shows the time series for each site, indicating a large number of observations for sites 1, 2, 7, 9 and 13. For sites 4, 10 and 14 there are 2 dives at the same day while for site 5 there is only one observation.

Further descriptions of the sites are as follow:

i For sites 1, 2, 9 and 13 the highest temperature was observed in 2005.

ii For site 7 the temperature shows an upward tendency, reaching the highest value in 2004, while in 2005 and 2006 there is a downward tendency.

iii For sites 3, 6, 8, 15 and 16 there is an upward tendency, although it is important to highlight that there is less number of observations.

iv For sites 11, 12 and 17 a downward tendency is observed, but as well as in the previous case there is less number of observations.

Figure 1: *Map of Colombia, indicating the location of Santa Marta*

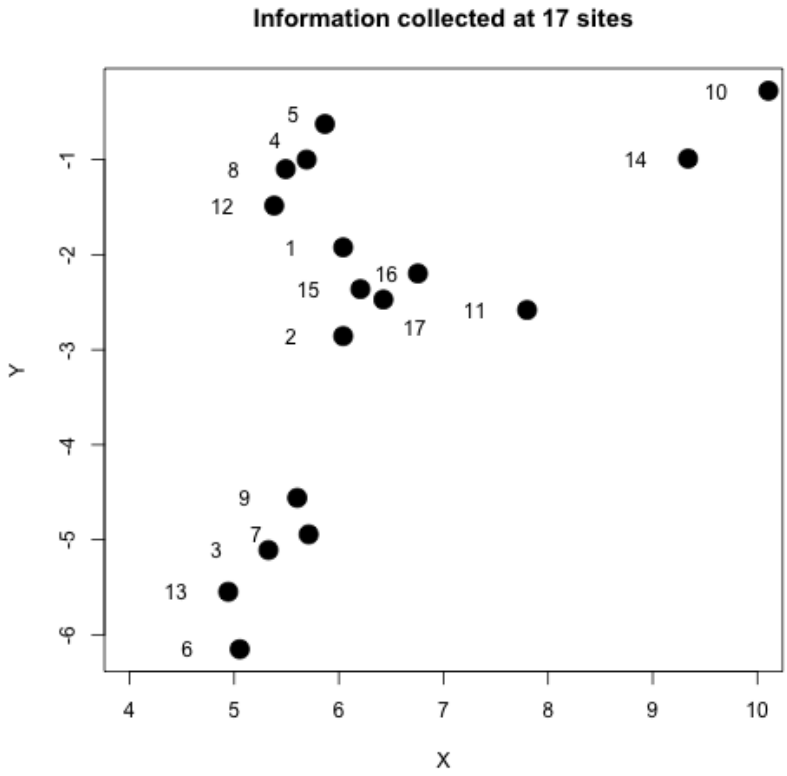

Figure 2: *17 sites where the information was collected as well as the code assigned to each point*

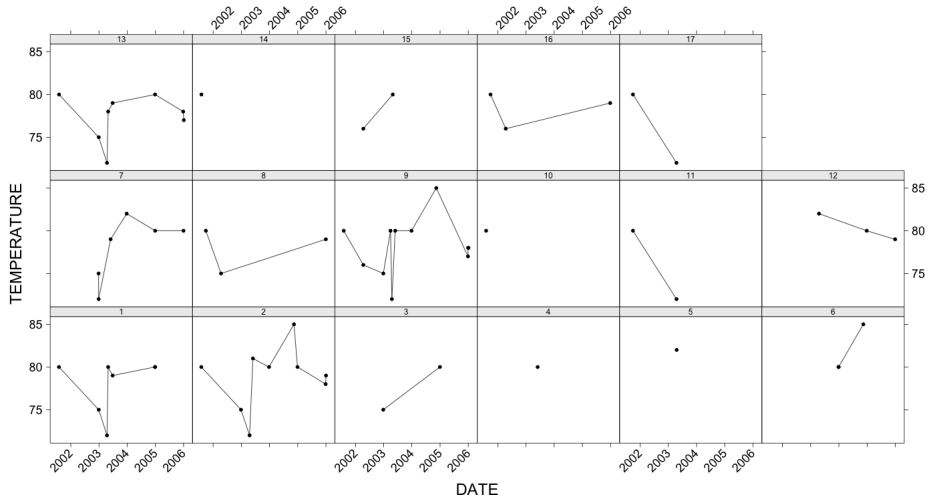

Figure 3: *Time series for water temperature by site*

5. Modeling Process

The first approach corresponds to a linear model, using as covariates the date where the information was collected in decimal form, the day (being the 1st of January 1 and the 31st of December 365 or 366) and the covariates X (direction North-South) and Y (direction East-West). These later corresponds to artificial coordinates in kilometers for the 17 sites.

To capture properly the seasonal component, the day covariate was introduced into the model as $cos\left(2\pi\left(\frac{day}{366}\right)\right)$ and $sin\left(2\pi\left(\frac{day}{366}\right)\right)$ (Esterby 1993).

These variables were included in model (1), under the the assumption that $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$. The idea is to include in the model both sources of variability to provide results for the data analyzed over time and space simultaneously, rather than in a marginal way.

$$y_i = \beta_0 + \beta_1 year_i + \beta_2 cos\left(2\pi\left(\frac{day_i}{366}\right)\right) + \beta_3 sin\left(2\pi\left(\frac{day_i}{366}\right)\right) + \beta_4 X_i + \beta_5 Y_i + \varepsilon_i \qquad i = 1, ..., n \tag{1}$$

Table 1 shows that for variables year and day, the linear model captures suitably the variability over time as well as the seasonal component, however for the coordinates X and Y the outcome is not the same. Figure 4 depicts the residuals against fitted values showing a linear pattern, indicating that a linear model may not be the more adequate approach.

Table 1: *Parameter estimated and p-values under linear model (1)*

Parameter	Estimate	Standard Error	t-value	p-value
year	1.133	0.313	3.611	<0.001
sin(day)	-3.609	0.586	-6.156	<0.001
cos(day)	-2.215	0.611	-3.623	<0.001
X	-0.617	0.451	-1.370	0.175
Y	0.291	0.233	1.247	0.217

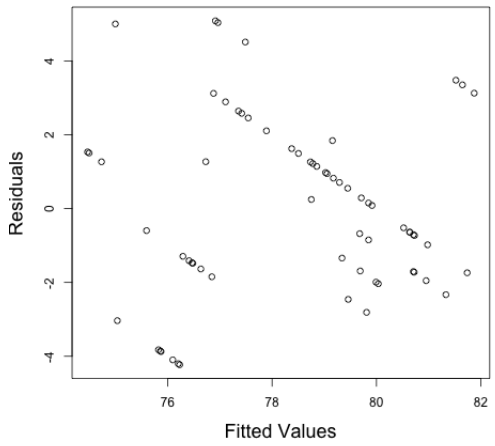

Figure 4: *Residuals against fitted values under linear model (1) for water temperature*

When the data observed is not easily described by a linear model, a suitable approach is to fit a nonparametric model of the form.

$$y_i = m(x_i) + \varepsilon_i \qquad i = 1, ..., n$$

where $m(x_i)$ corresponds to a smooth function, $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$.

There are different ways to obtain an estimate for $\hat{m}(x)$, one such approach is to use kernel estimators. Some of the most common are kernel smoothers, local regression, smoothing splines, regression splines, orthogonal series and wavelets (Green & Silverman 1994), (Wood 2006) (Fan & Gijbels 1996).

Throughout this paper a kernel smoother was chosen because the similarities with standard linear models, leading to useful statistical properties. An estimate for $\hat{m}(x)$ can be obtained by a local mean estimator (Watson 1964), (Nadaraya 1964) as

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} w(x_i - x; h) y_i}{\sum_{i=1}^{n} w(x_i - x; h)},$$

where $w(x_i - x; h)$ the weight function chosen, corresponds to a normal density centered on zero with standard deviation equal to h (Bowman & Azzalini 1997).

A two dimensional estimate for $\hat{m}(x_1, x_2)$ can be obtained from minimizing the weighted least squares

$$\sum_{i=1}^{n} y_i - \alpha - \beta_1(x_{i1} - x_1) - \beta_2(x_{2i} - x_2)^2 w(x_{1i} - x_1; h_1) w(x_{2i} - x_2; h_2),$$

over α, β_1 and β_2. It is very helpful to assess the combined effect of two variables in spatial data.

In the case of cyclical variables or seasonal effects, quite common in environmental information, an estimate for $\hat{m}(x)$ can be obtain using a local mean approach, where the weight function chosen corresponds to $w(x_i - x; h) = exp\left[\frac{r}{h} cos\left(\frac{2\pi(x_i - x)}{r}\right)\right]$, allowing us to obtain an estimate with period r.

Having chosen the way to obtain an estimate for $\hat{m}(x)$, we can introduce the additive models developed by Hastie and Tibshirani (1990).

The additive model used corresponds to model (2) under the assumption that $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$.

$$y_i = \beta_0 + m_1(year_i) + m_2(day_i) + m_3(X_i, Y_i) + \varepsilon_i \quad i = 1, ..., n \tag{2}$$

Each of the $m_j(x_j)$ $j = 1, ..., p$ smooth functions are estimated by the backfitting algorithm (Hastie & Tibshirani 1990), while β_0 corresponds to $\bar{y}$.

Figure 5 depicts each of the component for the additive model. The solid line corresponds to the smooth function fitted, the dashed line corresponds to a $\pm$ 2 standard error band and the surface corresponds to a smoothing function in two dimensions to capture the variability over space.

The degrees of freedom chosen for each covariate was 4, while for (X,Y) was 10. This provides enough flexibility, beyond of a linear shape while ensure that we capture large scale trend rather than small scale fluctuations.

The selection of degrees of freedom throughout this article is performed by a subjective method, as the main objective is to assess different models to capture tendency over time and space rather than choose a model based on automatic methods. In addition, the assessment of the partial residuals to evaluate the effect of each variable allows us to explore whether the degrees of freedom chosen is capturing well the relationship between the covariates and the dependent variable.

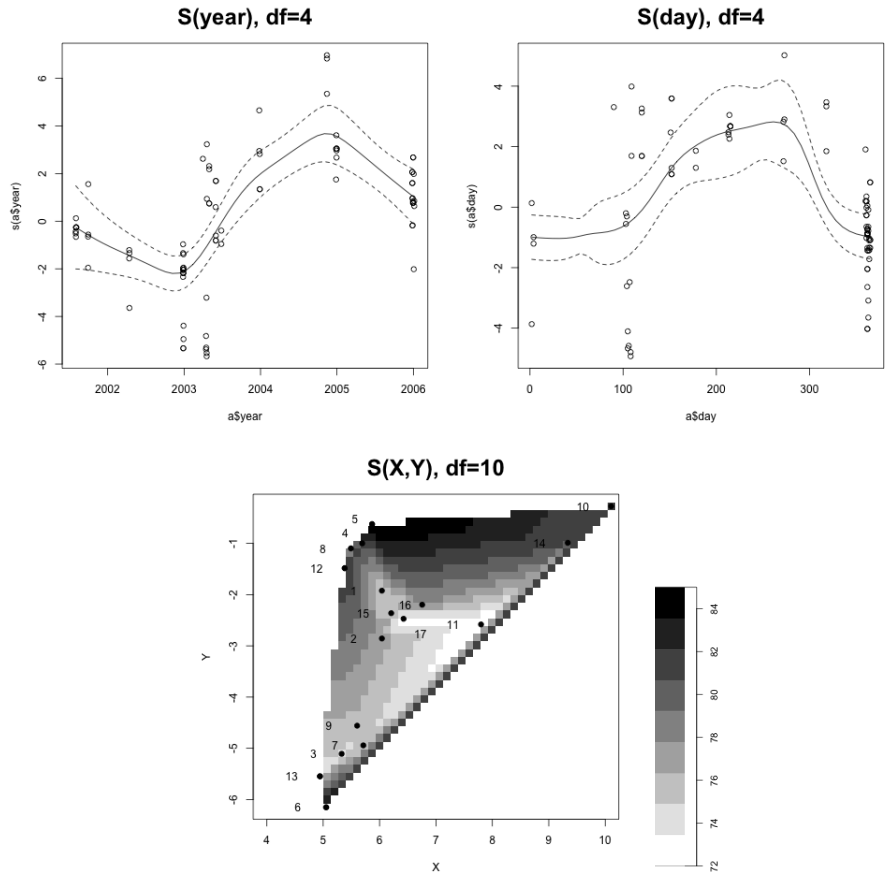

Figure 5: *Components for additive model for water temperature under model (2)*

Figure 6 depicts the residuals against fitted values under model (1) and model (2), showing that the linear pattern observed under model (1), (left hand side) is less marked under model (2) (right hand side). The odd behavior observed corresponds to temperature values 72, 75, 79 and 80 which have a frequency of 7, 7, 7 and 27 over the observed period respectively. This explains the pattern observed in the graphs where the same temperature is observed at different points over time and space.

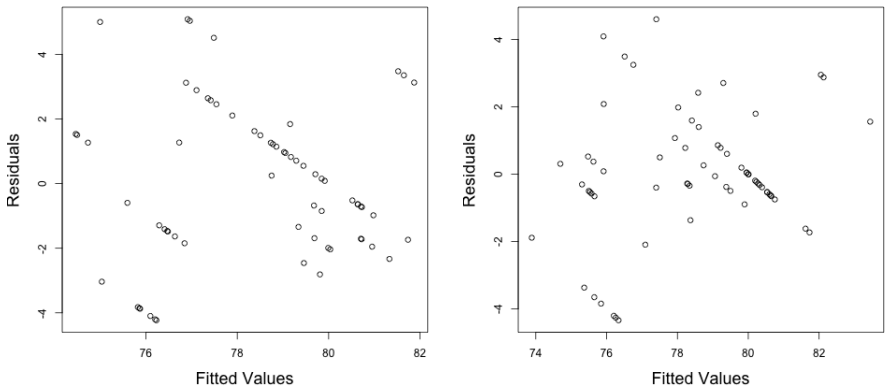

Figure 6: *Residuals against fitted values under model (1) and model (2)*

6. Testing for No Effect and Sensitivity Analysis

As part of the modeling process, it is necessary to assess the need for a non parametric effect rather than a linear effect for each variable. Following the idea of Hastie and Tibshirani (1990), the test used corresponds to an approximate F-test. This test statistic does not follow the exact F distribution, although results based on simulations (Hastie & Tibshirani 1990) provide enough evidence to support it as a guide to choose between different models. The approximate F-test is defined as

$$\frac{(RSS_1 - RSS_2)/(df_2 - df_1)}{RSS_2/(n - df_2)} \sim F_{df_2 - df_1, n - df_2},$$

where RSS_1 and RSS_2 are the residual sum of squares and df_1 and df_2 are the degrees of freedom of the models fitted.

The RSS is defined as $RSS = \sum_{i=1}^{n}(y_i - \hat{m}(x_i))^2$ or as a quadratic form as $RSS = y^t Q y$ where $Q = (I - P)^t(I - P)$. Each of the smooth functions can be expressed as a set of n x n projection matrices, providing the fitted values for an additive model as $Py = (\sum_{k=0}^{p} P_k)y$, where P_0 corresponds to a matrix with the value $1/n$ to estimate $\bar{y}$ (Bowman & Azzalini 1997).

In the same way as in a linear model, it is possible to obtain an analogous definition of approximate degrees of freedom for an additive model, where the approximate degrees of freedom for error can be defined as $df = tr[(I - P)^t(I - P)]$, with $P = \sum_{k=0}^{p} P_k$.

Table 2 shows the results under model (2) for each covariate, indicating that for the covariate day a non parametric effect is not required. This test was applied under different values of degrees of freedom to evaluate changes in the conclusion under different values.

Table 2: *Assessment of the need for a nonparametric effect rather than a linear effect and sensitivity analysis under different values of degrees of freedom*

	p-values						
parameters	df=4	df=6	df=8		df=10	df=12	df=14
year	<0.001	<0.001	<0.001	(X,Y)	<0.001	<0.001	<0.001
day	0.804	0.751	0.065				

According to this result the final model corresponds to model (3) under the assumption that $E(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$. This model corresponds to a semi-parametric model with a linear effect for day and a non parametric effect for year and (X,Y).

$$y_i = \beta_0 + \beta_1 cos\left(2\pi\left(\frac{day_i}{366}\right)\right) + \beta_2 sin\left(2\pi\left(\frac{day_i}{366}\right)\right) + m_1(year_i) + m_2(X_i, Y_i) + \varepsilon_i \quad i = 1, ..., n \tag{3}$$

Figure 7 depicts the graph for the residuals against fitted values and the graph of the fitted values against the observed values. According to the graphs the semi-parametric model fits properly the temperature of the sea. The systematic pattern as well as in model (2) corresponds to the values 72, 75, 79 and 80.

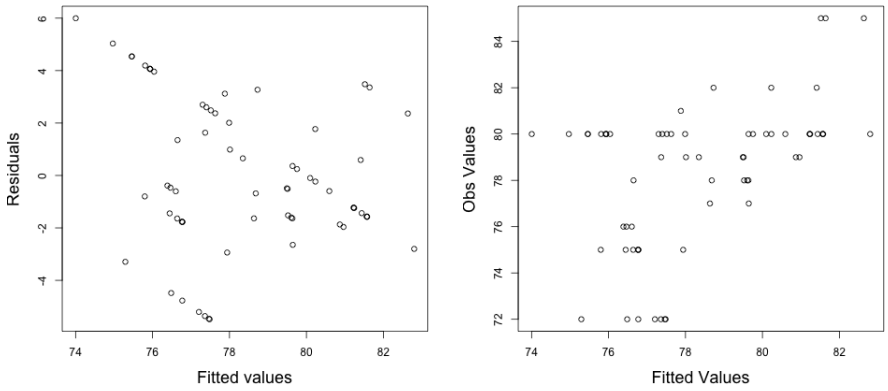

Figure 7: *Diagnostic check under semi-parametric model (3)*

Since the data was collected over time and space, it is necessary to assess the need to include a covariance structure over time and/or space, although given that the information was collected unevenly it is not possible to build the autocorrelation function. An alternative explained by Diblasi and Bowman (2001) is to build a variogram for the residuals. This test was originally developed to assess indepen-

dence over space for a single sample, although it is also useful as a diagnostic check for regression models.

Using $\hat{\gamma}(h) = \frac{1}{2}\frac{1}{|N(h)|}\sum_{N(h)}|Y(s_i)-Y(s_j)|^{\frac{1}{2}}$ as an estimator, where $N(h)$ denotes the collection of pairs of observations separated by a distance h, with h the distance between locations, independence over time or space is reflected in a constant variogram of the form $\gamma(h) = \sigma^2$, where $\gamma(h)$ is the theoretical variogram that explains the degree of dependence in two dimensions for space and one dimension for time.

Diblasi and Bowman (2001) provide a test to assess the presence of spatial correlation, allowing us to obtain a p-value under a null hypothesis that $\gamma(h) = \sigma^2$.

To assess correlation over time, the date where the information was collected was taken in Julian format. Figure 8 shows the variogram, indicating no evidence for correlation over time with a p-value of 0.145.

In the same way Figure 8 depicts the variogram to test independence over space, indicating no evidence of correlation over space with a p-value of 0.111. It is important to highlight that the reason why the test to assess independence over space can be applied, is because of the lack of evidence of autocorrelation over time.

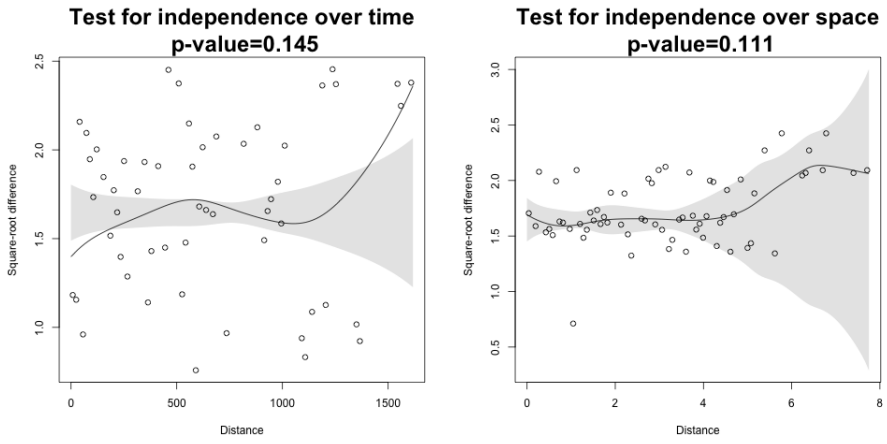

Figure 8: *Test of independence over time and space for the residuals of model (3)*

7. Discussion and Summary

Throughout this paper, different approaches have been evaluated, looking for the best way to capture variability over time and space simultaneously with information collected unevenly. The linear model did not offer a suitable description

of water temperature in Santa Marta, Colombia, mainly for the variability over space. The use of additive models provided a suitable tool based on the ability to capture non linear relationships through smooth functions unrestricted in shape.

Water temperature in Santa Marta, Colombia, collected since August 2001 to January 2006, showed a fluctuation over time with an upward tendency, reaching a peak in 2005 and a drop until 2006, although the temperature in 2006 is higher than 2004 and previous year. This is confirmed by the reports presented by the IPCC (IPCC 2007), where 2005 was identified as one of the years with the highest temperatures.

In respect to the variability over space, it is observed higher temperature in direction north-west, while the lowest temperature are observed in the south. This can be explained by the effect of the Sierra Nevada de Santa Marta, a snowy peak mountain in the central cordillera with an altitude of 5.770 meters, in the rivers Cesar, Palomino, Don Diego and Aracataca, rivers that end in the Caribbean sea.

In this particular case, the assessment of independence over time and space, suggested no evidence in both cases, however additive models can be use for correlated data, making easier the application on environmental information. In the case of correlated data, the main effect is in the calculation of standard errors and comparison models (Giannitrapani et al. 2005), where and approach through generalized least square allows this structure to be included.

As a further steps, will be interesting to include more information given the large dependency of the economical activities in this area in Colombia, looking for application on economical development, tourism and modeling fishing patterns.

Recibido: 18 de agosto de 2010
Aceptado: 26 de septiembre de 2010

References

Beare, D. & Reid, D. (2002), 'Investigating spatial-temporal change in spawning activity by atlantic mackerel between 1977 and 1998 using generalized additive models', *ICES Journal of Marince Science, 59: 711-724* .

Bowman, A. & Azzalini, A. (1997), *Applied Smoothing Techniques for Data Analyis: the kernel approach with S-Plus illustrations*, Oxford University Press.

Bowman, A., Giannitrapani, M. & Scott, E. (2009), 'Spatialtemporal smoothing and sulphur dioxide trends over europe', *Royal Statistical Society Series C 58, Part 5, pp 737-752* .

Diblasi, A. & Bowman, A. (2001), 'On the use of the variogram in checking for independence in spatial data', *Biometrics 57, 211-218* .

Eastoe, E. F., Halsall, C. J., Heffernan, J. E. & Hung, H. (2006), 'A statistical comparison of survival and replacement analyses for the use of censored data in a contaminant air database: A case of study from the canadian arctic', *Atmospheric Environment 40, 6528-6540* .

Esterby, S. (1993), 'Trend analysis methods for environmental data', *Environmetrics, 4(4), 459-481* .

Fan, J. & Gijbels, I. (1996), *Local polynomial modelling and its applications*, Chapman and Hall: London.

G.Bernal, Poveda, G., Roldan, P. & Andrade, C. (2006), 'Patrones de variabilidad de las temperaturas superficiales del mar en la costa caribe colombiana', *Rev Acad Colom Cienc 30 (115): 195-208* .

Giannitrapani, M., Bowman, A. & Scott, E. (2005), Additives models for correlated data with applications to air pollution monitoring, Technical report, University of Glasgow.

Green, P. & Silverman, B. (1994), *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*, Chapman and Hall: London.

Hastie, T. & Tibshirani, R. (1990), *Generalized Additive Models*, Monographs on Statistics and Applied Probability 43, Chapman and Hall.

IPCC (2007), *Contribution of Working Group I to the Fourth Assessment Report of the Intergovernmental Panel on Climate Change*, Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA,.

Lewis, P. A. & Ray, B. K. (1997), 'Modeling long-range dependence, nonlinearity, and periodic phenomena in sea surface temperatures using tsmars', *Journal of the American Statistical Association, Vol 92, No 439, 881-893* .

McMullan, A., Bowman, A. & Scott, E. (2007), 'Water quality in the river clyde: A case study of additive and interaction models', *Environmetrics* .

Nadaraya, E. (1964), 'Some new estimates for distribution functions', *Theory Probab. Appl. 9, 497-500* .

NAOWA(Webpage) (n.d.), Electronic Resource [Accessed 22/11/2009]. *http://www.naowa.com

Rincón, F. (2010), Statistical modelling of environmental trends over both time and space, Master's thesis, University of Glasgow.

Watson, G. (1964), 'Smooth regression analysis', *Sankhya, Ser. A, 26, 359-72* .

Wood, S. (2006), *Generalized Additive Models: An Introduction with R*, Chapman and Hall/CRC: London.