

Spatial smoothing on complex regions: a case study on the median length of deep water rose shrimps in the North-Western Ionian Sea

Smoothing spaziale su domini complessi: studio della lunghezza mediana del gambero bianco nel Mar Ionio Nord-occidentale

Simona Arcuti, Alessio Pollice, Crescenza Calculli, Nunziata Ribecco, Angelo Tursi

Abstract Conventional smoothing may produce errors across the boundary when used over complicated coastal and island regions, due to the Euclidean distances used to measure similarity between data points. In this work we want to compare the performance, in terms of prediction, of a conventional smoother based on Thin Plate Regression Splines (Wood, 2006) with the one of an alternative smoother, called Soap Film smoother (Wood, 2008). Both smoothers are considered inside a generalized additive model estimated to evaluate the changes in the median length (mm) of an aquatic population, over spatio-temporal scales and in response to anthropogenic and environmental factors.

Abstract *Gli smoother convenzionali tendono a produrre errori sul confine quando usati su regioni costiere e insulari complicate a causa della distanza Euclidea usata per misurare la similarità tra i punti. In questo lavoro si vogliono confrontare le prestazioni, in termini previsti, di uno smoother convenzionale basato su Thin Plate Regression Splines (Wood, 2006), con uno smoother alternativo, detto Soap Film smoother (Wood, 2008). Entrambi sono stati considerati all'interno di un modello additivo generalizzato stimato per valutare i cambiamenti nella lunghezza mediana di una popolazione acquatica, su scala spazio-temporale e in relazione a fattori antropici e ambientali*

Key words: Spatio-temporal models, generalized additive models, thin plate regression splines, soap film smoother

Simona Arcuti, Alessio Pollice, Crescenza Calculli, Nunziata Ribecco,
Dipartimento di Scienze economiche e metodi matematici, Università degli studi di Bari,
Largo Abbazia Santa scolastica 53, 70124 Bari, ITALY, e-mail: simona.arcuti@gmail.com,
{alessio.pollice, nunziata.ribecco}@uniba.it, calculli.enza@gmail.com

Tursi Angelo
Dipartimento di Biologia, Università degli studi di Bari, Via E. Orabona 4, 70124 BARI, e-mail:
angelo.tursi@uniba.it

1 Introduction

It is well known that most of demersal resources fluctuate over both temporal (Relini et al., 1999) and spatial scales in relation to the variability of abiotic and biotic factors in the ecosystem as well as to human activities, such as fishing (Cushing, 1982). In this paper we evaluate the relationships between the median length of a carapace species in the North-Western Ionian Sea and some environmental and anthropogenic variables from 1995 to 2006. In particular we consider the deep-water rose shrimp, *Parapenaeus longirostris*, due to its commercial importance and for its abundance in the North-Western Ionian Sea.

Statistical models used in this work are additive models, that allow to estimate the nonlinear influence of a number of predictors on the response variable, including spatial and temporal effects. The aim of this work is to compare two smooth functions types used to model the spatial effect: a conventional smoother based on *thin plate regression splines* (TPRS) and an alternative smoother known as *soap film smoother* (SOAP).

TPRS (Wood, 2006) are often the default method for modelling spatial data and are based on the Euclidean distance to measure the similarity between two points. However, when we consider a complicated coastal and island regions, the Euclidean distance does not necessarily provide a realistic representation. For example, evaluating the distance between two observations with the Euclidean distance, a straight line may cross a coastline or an exclusion zone. This can produce a low predictive power of the model and artificial over-or under-prediction in certain regions due to the distance measure.

The most recent alternative to conventional TPRS is the soap film smoother (Wood, 2008) which uses a specific basis function to model the interior domain and a cyclic penalised cubic regression spline to model each boundary. This method respects boundaries employing two tuning parameters: one global parameter for the interior domain and one for each boundary. The aim of this work is to show that SOAP has a better predictive power than TPRS.

2 Data collection

Fishery data were collected during experimental trawl surveys conducted from 1995 to 2006 in the North-Western Ionian Sea as part of the international project MED-ITS (MEDiterranean International Trawl Surveys). The study area runs from Capo d'Otranto (LE) to Capo Passero (SR) for a total surface of 16,350 km² at depths between 10 and 800 m. This Geographical region was divided into three sub-areas: Apulia; North-Calabria and South-Calabria. The sampling design adopted was random-stratified by depth: for each depth range the allocation of hauls was proportional to the extension of the corresponding stratum. In this work we focus on an important demersal resource in the North Western Ionian Sea, the deep-water rose shrimp *Parapenaeus longirostris* (Lucas, 1846). Carapace length is chosen as

the population variable influenced by environmental factors. In particular we consider: geographical factors (depth and geographic coordinates), temporal effect (year of data collection), abiotic factors for every year and each sub-areas (Fishing effort, Sea Surface Temperature and Precipitations), abiotic factors for every year (Winter North Atlantic Oscillation Index and Annual Mediterranean Oscillation Index).

In order to compare TPRS and SOAP over a closed domain, a reference region (fig. 1) was defined, bounded by the coastline and the 1000m depth contour (maximum plausible depth for *P. Longirostris*).

Fig. 1 Reference region defined by the coastline and the 1000m depth contour. Estimates and predictions are confined into this region.



3 Data modeling

Statistical models used in this work are Additive Models, in which we assume that the mean of response variable $E(Y)$ is related to the predictors by a link function. Particularly, we estimate two models as follow:

$$g(\mu_i) = \beta_0 + \sum_{j=1}^J \beta_j \cdot X_i^j + f^*(long_i, lat_i) + \sum_{k=1}^K f_k(X_i^k) = \eta_i \quad (1)$$

where: $g(\cdot)$ is a link function; β_0 is the intercept; β_j is the coefficient that resumes the j -th linear effects; X^j and X^k are the j -th and k -th linear and smooth effects, respectively; $f_k(\cdot)$ is the k -th semi-parametric smooth function modelling as TPRS; $f^*(\cdot)$ is a semi-parametric smooth function modelling as a TPRS in the first model and as a SOAP in the second model.

Both GAMs are estimated by *Maximum Penalized Likelihood* (Wood, 2006), in which a penalty, for every smooth functions, is added to the sign-switched logarithm of the likelihood function and multiplied by the degrees of freedom of the corresponding smooth function. This addendum penalizes more complex models, thus avoiding over-fitting problems. Smoothing parameters estimation is obtained by two alternative approaches: Generalized Cross Validation (GCV) and Un-Biased Risk Estimator (UBRE) (Wood, 2006).

3.1 Soap film smoother

While we refer to Wood (2006) for details about TPRS formulation, in this paper we focus the attention on the SOAP formulation when the reference region is a closed area. Authors have generalized the smooth structure to cases where the reference region includes more than one closed areas (Wood, 2008).

Consider a smooth function $f^*(x, y)$ over the x, y plane, let B the boundary and Ω the region made up of all x, y which are interior to B . Suppose that $f^*(x, y)$ is known on B , and that $z_k = f^*(x_k, y_k)$ are the values of f^* on n points within Ω . Now, f^* is a function so that:

$$\frac{\partial^2 f^*}{\partial x^2} + \frac{\partial^2 f^*}{\partial y^2} = \rho \quad \text{where} \quad \frac{\partial^2 \rho}{\partial x^2} + \frac{\partial^2 \rho}{\partial y^2} = 0 \quad (2)$$

except at the (x_k, y_k) points and $\rho = 0$ on B . Equations in (2) are the Poisson and Laplace equations, respectively.

Let $\rho_k(x, y)$ be a function which is zero on B , satisfies the Laplace equation in Ω except at the single k -th point and so that its integral on Ω is equal to 1. Then we can define the following function: $\rho(x, y) = \sum_{k=1}^n \gamma_k \rho_k(x, y)$, where γ_k are coefficients. Now, f^* can be written in terms of the γ_k :

$$f^*(x, y) = a(x, y) + \sum_{k=1}^n \gamma_k g_k(x, y) \quad (3)$$

where: $a(x, y)$ is the solution of the Poisson equation with $\rho(x, y) = 0 \forall (x, y)$ and subject to the boundary condition that $f^*(x, y)$ is known on B ; $g_k(x, y)$ is the solution of the Poisson equation $\rho(x, y) = \rho_k(x, y)$ and subject to the boundary condition that $f^*(x, y) = 0$ on B ; γ_k are parameters to be estimated.

If we do not have special knowledge about the value of f^* on the boundary, we model the first part of (3) using a *cyclic penalized regression spline* as $a(x, y) = \sum_{j=1}^J \alpha_j a_j(x, y)$, where: α_j are coefficients to be estimated and $a_j(x, y)$ are solutions of the Poisson equation with $\rho(x, y) \equiv 0$ and subject to the condition that $\alpha_j = 1$ and $\alpha_i = 0$ for all $i \neq j$.

4 Results

Table (1) summarizes the results for two generalized additive models, M.TPRS and M.SOAP, estimated considering a thin plate regression spline and a SOAP film smoother to model the spatial effect. In both models we assume that the response variable has a normal distribution. Then we consider the identity link function and predictors are chosen using a stepwise regression analysis, minimizing GVC score at each step. Finally, for each model we calculate the predictions for year 1995 on a new grid of coordinates considering the mean value of precipitations and tem-

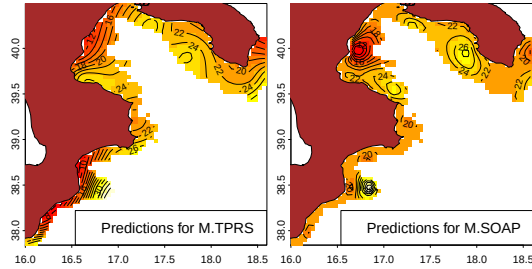
perature in this year. Looking at the table, there aren't big differences between the

Table 1 Coefficient estimates of the GAMs *P.logirostris* length. P-value in brackets.

	<i>M.TPRS</i>	<i>M.SOAP</i>
<i>Linear Terms</i>	<i>Estimate (p-values)</i>	<i>Estimate (p-values)</i>
Intercept	23.13 (< 0.001)	22.70 (< 0.001)
Precipitation	-0.02 (0.07)	
<i>Smooth Terms</i>	<i>Degree of freedom (p-values)</i>	<i>Degree of freedom (p-values)</i>
f(long, lat)	18.14 (< 0.001)	18.73 (< 0.001)
f(year)	3.60 (< 0.001)	4.00 (0.02)
f(temperature)	2.46 (0.003)	2.00 (0.04)

coefficients estimated by two the models for the linear effects, and also between the degrees of freedom of each smooth effect. The effects of time and temperature (both not reported) are similar for two models. Figure 2 shows predictions for both models. The SOAP smoother allows to put a constraint of null effect on the boundary. As highlighted in figure 2 this scheme produces a greater detail in the internal domain, and allows to highlight areas in which higher and lower values of the response variable are observed. In order to compare SOAP and TPRS performances,

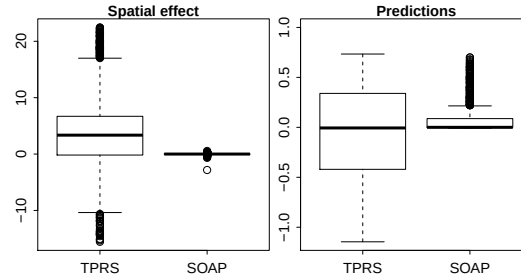
Fig. 2 Spatial smooth function components of the fitted *P. longirostris* length. The yellow coloured areas define larger values of median length. In reverse, the warm colours indicate areas characterized by lower values.



we evaluate the spatial effect estimated by both models on the boundary. Results are resumed using the boxplot, shown in figure 3 (left). As expected, M.SOAP estimates a null effect in almost all cases, whereas model M.TPRS produces a range of estimations between -10 and 20 . Despite this positive result, the SOAP constraint concerns only the spatial effect, but the full model M.SOAP is not able to estimate null predictions on the boundary, due to the influence of the intercept and of other covariates. For this reason, in order to compare the predictive power of the two strategies we estimate two new models, M2.SOAP and M2.TPRS, including only the spatial effect. We consider again a SOAP smoother and a TPRS to model the spatial effect, respectively for M2.SOAP and M2.TPRS. Finally, we repeat the predictions on the boundary. Results are resumed using a boxplot, shown in figure 3 (right). Using SOAP predictions values on the boundary are null except some

outliers, whereas using TPRS the range of predictions on the boundary increases including also negative values.

Fig. 3 Boxplots on the left resume the spatial effect estimated on the boundary by M.SOAP and M.TPRS. Boxplots on the right resume prediction values on the boundary estimated by M2.SOAP and M2.TPRS.



5 Conclusions

In order to analyze the median length distribution of the deep water rose shrimp, we suggested a GAM incorporating a soap film spline basis for space to provide more accurate spatial maps. The spatial smoother used was chosen to prevent possible over-or under-estimates of the response variable considering that the borders of domain represent the physical boundaries. This strategy allowed to restrict estimations and predictions into the areas where the species is preferentially distributed, excluding the zone where it is not captured. Furthermore, using the SOAP smoother we obtained a smoothed map with a much better representation of the length distribution of the *Parapenaeus longirostris*. Comparing the two maps of TPRS and SOAP, they show the same colors, but the SOAP allows to highlight better the areas with higher or lower lengths. In particular, we can see two areas with lower lengths, probably the nursery, near Otranto (LE) and Trebisacce (CS), and some hot-spots are highlighted along the coast of Gallipoli (LE) and off Roccella Jonica (RC). However, results have shown as the influence of the intercept and of other covariates doesn't allow to obtain null predictions on the boundary. Further developments could regard the extension of the SOAP constraint on the other covariates.

References

1. Cuscing D.H.: Climate and fisheries. In: Academic Press, London. (1982)
2. Relini, G., Bertrand, J., Zamboni, A.: Synthesis of the knowledge on bottom fishery resources in Central Mediterranean, Italy and Corsica. In: Biol. Mar. Medit. 6 (Suppl. 1), 868 pages. (1999)
3. Wood, S.: Generalized additive models. Chapman & Hall CRC, Boca Raton, FL. (2006)
4. Wood, S.: Soap film smoothing. In: J.R.Statist.Soc.B 70(5), 931-955. (2008)