SORT 36 (2) July-December 2012, 221-230

Flexible geoadditive survival analysis of non-Hodgkin lymphoma in Peru

Claudio Flores[1], Mar Rodríguez-Girondo[2,3], Carmen Cadarso-Suárez[3], Thomas Kneib[4], Guadalupe Gómez[1] and Luis Casanova[5]

Abstract

Knowledge of prognostic factors is an important task for the clinical management of Non Hodgkin Lymphoma (NHL). In this work, we study the variables affecting survival of NHL in Peru by means of geoadditive Cox-type structured hazard regression models while accounting for potential spatial correlations in the survival times. We identified eight covariates with significant effect for overall survival. Some of them are widely known such as age, performance status, clinical stage and lactic dehydrogenase, but we also identified hemoglobin, leukocytes and lymphocytes as covariates with a significant effect on the overall survival of patients with NHL. Besides, the effect of continuous covariates is clearly nonlinear and hence impossible to detect with the classical Cox method. Although the spatial component does not show a significant effect, the results show a trend of low risk in certain areas.

MSC: 62P10 (Applications to biology and medical science) 62N01 (Censored data models) 6207 (Data analysis)

Keywords: Non-Hodgkin lymphoma, structured regression models, survival analysis.

1. Introduction

Non-Hodgkin lymphomas (NHLs) are a group of lymphoproliferative malignancies of the lymphatic system defined by different morphological, immunophenotypic and genetic features. This heterogeneity determines different patterns of prognosis in the NHL patients that should be considered to optimize their treatment benefit (Friedberg *et al.*, 2008).

[1] Department of Statistics, Universitat Politècnica de Catalunya, Spain.

[2] SiDOR Research Group, University of Vigo, Spain. margirondo@uvigo.es

[3] Department of Statistics, University of Santiago de Compostela, Spain.

[4] Department of Economics, Georg August University, Göttingen, Germany.

[5] Instituto Nacional de Enfermedades Neoplásicas, Lima, Peru.

Received: November 2011
Accepted: May 2012

Traditionally, the International Prognostic Index (IPI) has been used to classify the NHL patients into four risk groups (low, intermediate low, intermediate high, high) considering five variables of prognostic significance (age, performance status, clinical stage, lactic dehydrogenase and extranodal sites) derived from a Cox regression analysis based on categorical covariates (Shipp *et al.*, 1993). However, a relatively important group of patients presents poor survival, despite being classified as good prognosis according to the IPI.

Several aspects can be related to the observed inaccuracy of the IPI. It is possible that important prognostic factors are not being included in the analysis such as new genetic and biological markers currently under investigation. Another important issue that could lead to implausible results refers to the categorization of the continuous covariates included in the IPI (age and lactic dehydrogenase).

Beyond the IPI, many studies of prognostic factors for NHL have been performed using the classical Cox's proportional hazard model. Within this framework, the effect of the continuous covariates is assumed to have a linear functional form, however it is important to note that when this assumption is not satisfied, the Cox model may lead to biased inferences, loss of statistical power and incorrect conclusions (Therneau and Grambsch, 2000).

In addition, in databases based on hospital records, referral centres, population studies or multicenter clinical trials, the results may be affected by spatial correlations. These complexities in the covariates affecting survival are not covered by the Cox model and hence a more general and flexible regression framework is required.

A variety of flexible methods have been developed in recent years. An up-to-date review of Cox-type models extensions can be found in Buchholz and Sauerbrei (2011). In this article we use geoadditive Cox-type structured hazard models to inspect the functional form of several covariates effects, including a spatial component, on the overall survival of the patients with NHL.

The rest of the paper is organized as follows. In Section 2 structured geoadditive Cox-type hazard regression models for modelling survival data are revisited. Section 3 presents the results of the analysis of the data set of NHL in Peru and finally, a discussion concludes the paper.

2. Methodology

2.1. Geoadditive survival models

In many clinical studies, the common target of analysis is to model the effect of several covariates (prognostic factors) on the survival time. A classical tool for studying the effect of a vector of covariates $\boldsymbol{v}$ on continuous survival times is the Cox proportional hazards model (Cox, 1972):

$$\lambda_i(t,\boldsymbol{v}) = \lambda_0(t)\exp(\boldsymbol{v}_i^\mathsf{T}\boldsymbol{\gamma}) \tag{1}$$

However, this specification is often not flexible enough for the correct modelling of variables affecting survival in many applications.

In our analysis, we used structured geoadditive survival models (Hennerfeind *et al.*, 2006; Kneib and Fahrmeir, 2007), a flexible spatial generalization of the Cox model. Specifically, the linear predictor of equation (1) was extended to a structured geoadditive predictor, including a spatial component for geographical effects and nonparametric terms for modelling unknown functional forms of the log-baseline hazard rate and nonlinear effects of continuous covariates. Specifically, individual hazard rates are given by:

$$\lambda_i(t) = \exp(\eta_i(t)), i = 1,\ldots,n \tag{2}$$

with geoadditive predictor

$$\eta_i(t) = g_0(t) + \boldsymbol{v}_i^\mathsf{T}\boldsymbol{\gamma} + \sum_{k=1}^{q} s_k(x_{ik}) + f_{\mathrm{spat}}(s) \tag{3}$$

where $g_0(t) = \log(\lambda_0(t))$ represents the log-baseline hazard rate, the vector $\boldsymbol{\gamma}$ contains the usual linear effects, $s_k(x_k)$ refers to the nonlinear effect of a continuous covariate x_k, and $f_{\mathrm{spat}}(s)$ is the spatial effect in region $s \in \{1,\ldots,S\}$.

In this representation, all the nonparametric effects, including the log-baseline hazard are modeled using penalized splines (P-splines, Eilers and Marx, 1996). Thus, the nonparametric problem is replaced by a parametric equivalent, in which a vector of regression coefficients is estimated under a smoothness penalty (details are given in Section 2.2.). The general idea is to approximate the functions g_0 and s_k by linear combinations of B-splines basis functions,

$$s_k(x) = \sum_{j=1}^{d_k} \beta_j B_j(x) \tag{4}$$

where vector $\boldsymbol{\beta}_k = (\beta_1,\ldots,\beta_{d_k})$ is the vector of unknown coefficients corresponding to the B-splines basis of degree a and defined over a grid of k knots lying on the domain of x, with $d_k = a + k - 1$. Specifically, we considered B-splines basis of degree 3 and a grid of 20 equidistant knots in our analyses.

At the same time, the spatial effect of each region s is split up into a structured part and an unstructured part:

$$f_{\mathrm{spat}}(s) = f_{\mathrm{str}}(s) + f_{\mathrm{unstr}}(s) \tag{5}$$

With this division of the spatial effects, we aim to distinguish between two types of geographical influential factors. On the one hand, the structured effect refers to a general smooth spatial effect along the whole studied area. On the other hand, the unstructured effect accounts for possible effects that may be present only locally. The structured spatial effects are modeled by means of Markov random fields, assuming that the effect of an area s is conditionally Gaussian, where the expectation is the mean of the effects of neighbouring areas and the variance is inversely proportional to the number of neighbours, specifically

$$f_{\text{str}}(s) = \beta_s^{\text{str}} = \frac{1}{N_s} \sum_{s' \epsilon \delta_s} \beta_{s'}^{\text{str}} + u_s, \;\; u_s \sim N\left(0, \frac{1}{\lambda_{\text{str}} N_s}\right) \tag{6}$$

where δ_s denotes the set of neighbouring areas of s and N_s the corresponding number of areas falling in δ_s. As for the unstructured spatial effects, a Gaussian region specific i.i.d. random effect is assumed.

As a result, we can express each of the predictor components as the product of an appropriate design matrix $\boldsymbol{Z}_j$ and a vector $\boldsymbol{\beta}_j$ of regression coefficients, and consequently we can represent the predictor vector $\boldsymbol{\eta}$ in a generic matrix notation as $\boldsymbol{\eta} = \boldsymbol{V}\boldsymbol{\gamma} + \boldsymbol{Z}_1 \boldsymbol{\beta}_1 + \cdots + \boldsymbol{Z}_q \boldsymbol{\beta}_q + \boldsymbol{Z}_{\text{str}} \boldsymbol{\beta}_{\text{str}} + \boldsymbol{Z}_{\text{unstr}} \boldsymbol{\beta}_{\text{unstr}}$, where $\boldsymbol{V}$ is the design matrix of parametric effects.

Interestingly, from equations (2) and (3) we can extend the concept of hazard ratio with respect to a reference value x_{ref}. In contrast to the linear hazard ratios derived from the Cox model, the structured geoadditive survival specification provides flexible hazard ratio curves. Hence, for a given smooth effect s associated to a continuous covariate X, the adjusted hazard ratio for a subject with covariate x compared to a subject with covariate x_{ref} is given by the smooth curve:

$$\text{HR}(x, x_{\text{ref}}) = \exp(s(x) - s(x_{\text{ref}})) \tag{7}$$

2.2. Estimation of the parameters

Under the usual assumptions about non-informative censoring, the log likelihood, given the vectors of all parametric effects $\boldsymbol{\gamma}$ and all nonparametric and spatial effects $\boldsymbol{\beta}$, is $\ell(\boldsymbol{\gamma}, \boldsymbol{\beta}) = \boldsymbol{\delta}^{\mathsf{T}} \boldsymbol{\eta} - \mathbf{1}^{\mathsf{T}} \boldsymbol{\Lambda}$, where $\boldsymbol{\eta}$ denotes the linear predictor defined in (3) and $\boldsymbol{\delta}$ and $\boldsymbol{\Lambda}$ are, respectively, the vector of censoring indicators and cumulative hazard rates.

However, instead of obtaining the estimates of $\boldsymbol{\beta}$ by means of the unpenalized likelihood, a penalty term is added to control the level of smoothness by penalizing wiggly functions. The most commonly used penalization term is based on the integral of the second derivative of the smooth functions, s_k:

$$\text{pen}(s_k) = \frac{1}{2}\lambda_i \int_0^\infty [s_k''(z_i)]^2 dz_i \tag{8}$$

Since equation (8) is a quadratic form of the corresponding vector of regression coefficients $\boldsymbol{\beta}_j$, it can be written as $\frac{1}{2}\lambda_j \boldsymbol{\beta}_j \boldsymbol{K}_j \boldsymbol{\beta}_j$, where the penalty matrix $\boldsymbol{K}_j$ is a positive semidefinite matrix and λ_j a smoothing parameter. Furthermore, the smooth functions for the nonlinear effects are represented in terms of B-splines and it allows to approximate the penalty term in terms of the squared differences of coefficients associated with adjacent basis functions (Eilers and Marx, 1996). As a result, the difference penalty matrix can be written as $\boldsymbol{K}_j = \boldsymbol{D}^\top \boldsymbol{D}$, with $\boldsymbol{D}$ the second order difference matrix of neighbouring coefficients.

A special remark about the spatial smoothing is required. In this case, the smoothing referees to the intuitive idea that risk in neighbouring areas should be close to each other. We define as neighbour areas those sharing a common boundary and analogously to the nonlinear effects, we penalize large deviations between neighbouring coefficients $\boldsymbol{\beta}_{\text{str}}$, where λ_{str} from equation (6) is considered as the corresponding smoothing parameter. Hence, the corresponding penalty matrix $\boldsymbol{K}_{\text{str}}$ is defined as an adjacency matrix. For the unstructured spatial effect, the penalty matrix is simply the identity matrix corresponding to independent and identically distributed random effects for the regions.

As a result, the estimation of the regression effects is based on the penalized log-likelihood to ensure a compromise between fidelity to data (in terms of the likelihood) and smoothness (in terms of the penalty terms):

$$l_{\text{pen}}(\boldsymbol{\Upsilon},\boldsymbol{\beta}) = l(\boldsymbol{\Upsilon},\boldsymbol{\beta}) - \sum_{j=1}^{q} \lambda_j \boldsymbol{\beta}_j^\top \boldsymbol{K}_j \boldsymbol{\beta}_j - \lambda_{\text{str}} \boldsymbol{\beta}_{\text{str}}^\top \boldsymbol{K}_{\text{str}} \boldsymbol{\beta}_{\text{str}} - \lambda_{\text{unstr}} \boldsymbol{\beta}_{\text{unstr}}^\top \boldsymbol{K}_{\text{unstr}} \boldsymbol{\beta}_{\text{unstr}} \tag{9}$$

Empirical Bayes inference was used to fit the model. This inferential procedure is based on a mixed model representation of equation (9) where the smoothing parameters (λ_j) are considered as variance components corresponding to the vector of regression coefficients ($\boldsymbol{\beta}_j$). It allows for the simultaneous estimation of the regression coefficients and the smoothing parameters corresponding to each unknown function g_0, s_k or f_{spat} using restricted maximum likelihood (REML) estimation. See Kneib and Fahrmeir (2007) for details.

The analysis was conducted using BayesX statistical software (Brezger *et al.*, 2005) freely available online from `www.bayesx.org`. Empirical Bayes inference was performed due to its equivalence to the penalized splines likelihood in the frequentist framework but BayesX also allows for a full Bayesian inference by means of MCMC simulation techniques (Hennerfeind *et al.*, 2006). To check the consistency of our results with regard to the inference procedure, the corresponding full Bayesian analysis was also conducted.

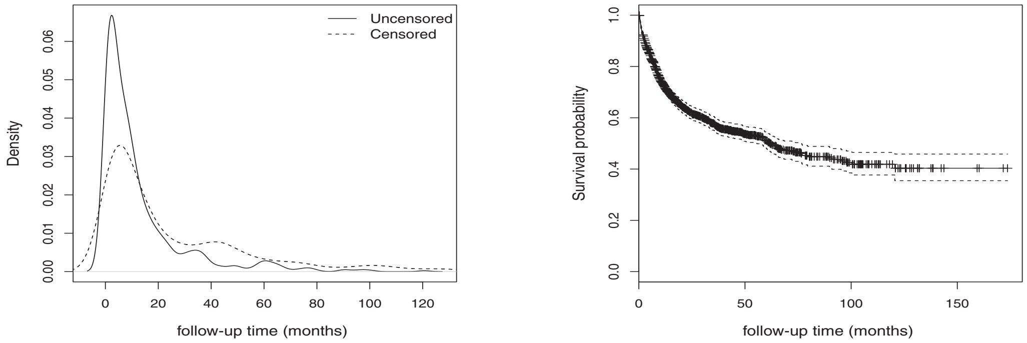

Figure 1: *Density function of survival time (left) and Kaplan-Meier estimate of the overall survival curves with corresponding 95% confidence bands(right).*

3. Application to NHL

We analyzed survival data for 2160 patients diagnosed NHL, older than 14 years and treated at the Instituto Nacional de Enfermedades Neoplásicas (INEN), Lima, Peru, between 1990 and 2002. The clinical features evaluated were age, sex, performance status (zubrod), primary disease, clinical stage (CS), B symptoms, hemoglobin (Hbg), log leukocytes (ln(WBC)), lymphocytes and log lactic dehydrogenase (ln(LDH)).

Table 1: *Fixed and random estimates of the fitted model.*

Variables	df	HR (95% CI)
Fixed effects:		
Sex: male	1	1.25 (1.07,1.46)
Zubrod: 2-4	1	1.88 (1.59,2.22)
Primary: nodal	1	0.90 (0.76,1.06)
CS: III-IV	1	1.44 (1.22,1.71)
B-symptoms	1	1.16 (0.99,1.37)
Non-parametric effecs:		
g(t)	3.52	see figure 2
Age	2.67	see figure 2
Hbg	4.16	see figure 2
ln(WBC)	2.63	see figure 2
Lymphocytes	5.79	see figure 2
ln(LDH)	4.16	see figure 2
Spatial effects:		
Random component	5.41	see figure 3
Spatial component	1.06	see figure 3
AIC	6549.95	
BIC	6750.99	

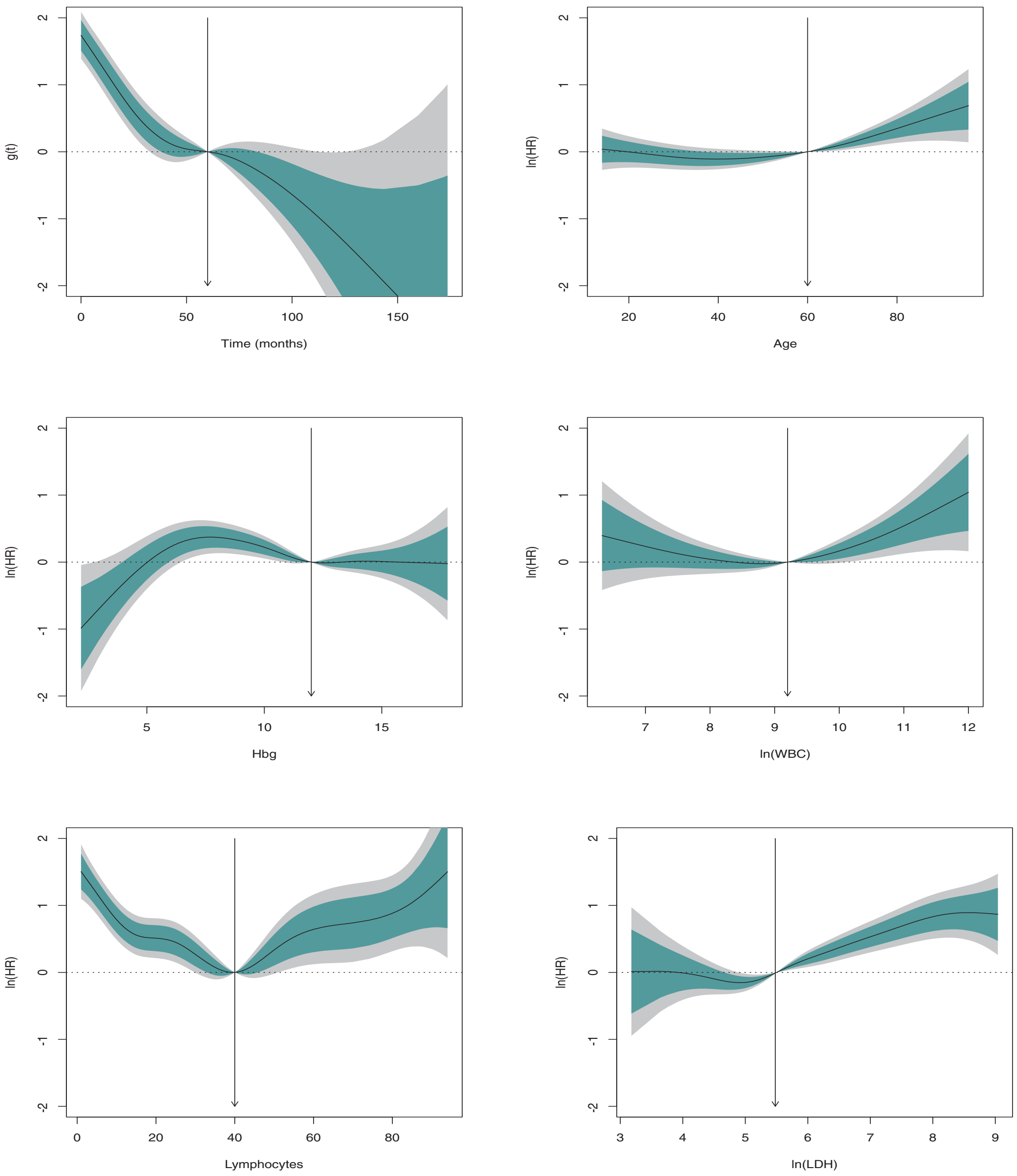

Figure 2: *Logarithm of the smooth hazard ratio curves derived from the geoadditive model fit. Clinical cut-off values were used as reference points in the analysis. 80% (green) and 95% (grey) credible intervals are shown.*

The median age was 54.0 years (range: 14-96 years). Most patients presented advanced-stage disease at diagnosis: 50.8% presented Stage I-II and 49.2% presented Stage III-IV. Thirty-eight percent of the patients had B symptoms at diagnosis. The median length of follow-up for the patients was 12.6 months. Among all the patients, 32.8% had died before the end of the follow-up period (uncensored cases) and 67.2%

remained alive (censored cases). Figure 1 shows the distribution of survival time of patients with and without censoring (left) and the Kaplan-Meier estimate of the overall survival curve (right). According to the structured geoadditive Cox-type hazard analysis, eight prognostic factors were identified associated with worse survival (Table 1). Three categorical covariates: male patients, zubrod 2-4 and clinical stage III-IV at diagnosis were associated with worse prognosis for overall survival. The location of the disease described as primary nodal or extranodal, and the symptoms B had no significant effect on the overall survival.

A significant nonlinear relationship was identified for the effects of all continuous covariates: age, Hbg, ln(WBC), lymphocytes and ln(DHL). Figure 2 shows the functional form of the covariate effects in the log hazard ratio. Usual clinical cut-off values were used as reference points: 60 years (age), 12 g/dL (Hbg), 10^3 counts/dL (WBC), 40% (lymphocytes) and 240 UI/L (DHL). Note that a strong nonlinear effect ($df = 5.79$) was found for Lymphocytes with increased hazard ratios for lowest and highest values. Risk geographical pattern is presented in Figure 3. Although Lima and Apurimac areas were identified as increased risk areas, the spatial effect was not significant according to the included variables.

As for the inference procedure, the results obtained from the full Bayesian inference (not shown) are very similar to the ones derived from REML estimation, hence we can assess that both inferential methods perform equivalently to our data.

4. Conclusions

The study of new covariates (with possible non-linear functional forms) in a flexible way and the existence of spatial correlation are examples of new challenges that the traditional tools of survival analysis do not allow to manage in an efficient way. Recent development of flexible methods for survival analysis allow for a deeper investigation of the variables affecting survival.

We used structured geoadditive survival models, a nonparametric approach that allows for the joint estimation of the baseline and covariates effects by means of a modelling through P-splines. Specifically, we considered nonlinear effects for the continuous covariates and we also account for possible geographical correlation.

In this work we identified eight covariates with significant effect for overall survival by means of the fitted geoadditive Cox-type structured hazard model. Age, zubrod, CS and DHL are prognostic factors reported in many published series, but we also identified hemoglobin, leukocytes and lymphocytes as covariates with a significant effect on the overall survival of patients with NHL.

Besides, the effect of continuous covariates is clearly non linear and hence impossible to detect with the classical Cox method. Nicely, the concept of hazard ratio is extended to obtain smooth hazard ratio curves for each of the continuous covariates.

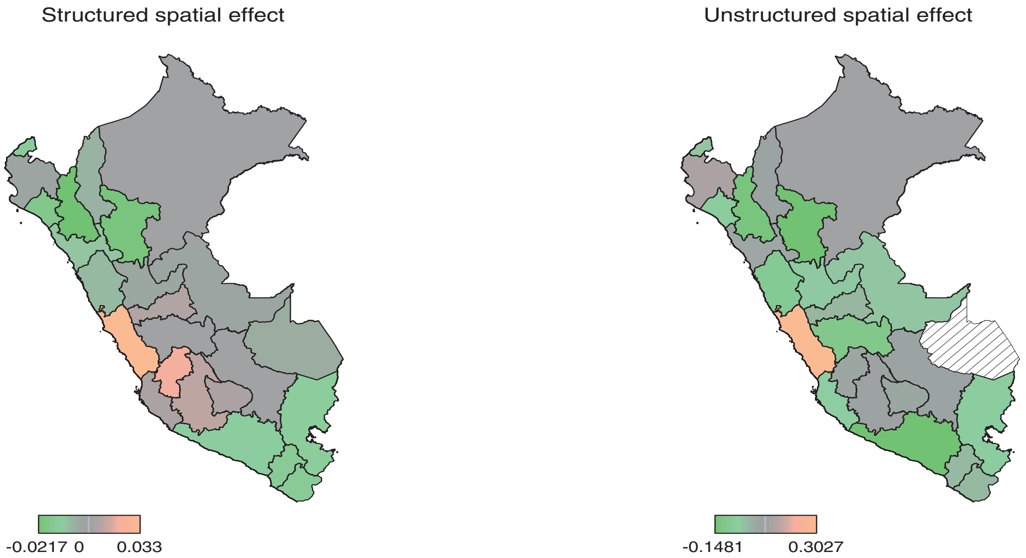

Figure 3: *Spatial effect estimates.*

Although the spatial component does not show a significant effect, the results show a trend of low risk in certain areas. This phenomenon could be associated with certain subtype of NHL more frequent in these areas. So, the spatial analysis points out that further inspection of the NHL subtypes is required.

Still, it is noteworthy that more general specifications of the predictor are possible in the structured geoadditive Cox-type hazard regression framework, such as the inclusion of time-varying effects which allows to relax the proportional hazards assumption or the inclusion of interactions between covariates. In fact, possible extensions of the present work considering time-varying prognostic factors and interactions between them are currently under investigation.

To sum-up, geoadditive Cox-type structured hazard regression is a useful tool for assessing prognostic factors for the survival in a flexible way. This methodology allows to detect variables that may affect the risk of mortality while taking the possible spatial correlation of data into account.

5. Acknowledgments

This work was partially supported by grants MTM2008-01603, MTM2008-03129, MTM2008-06747-C02-00, MTM 2011-2304 and MTM2011-28285-C02-01 from the Spanish Ministry of Science and Innovation and the INBIOMED project (DXPCTSUG, Ref. 2009/063).

References

Brezger, A., Kneib, T. and Lang, S. (2005). BayesX: Analyzing Bayesian structural additive regression models. *Journal of Statistical Software*, 14, i11.

Buchholz, A. and Sauerbrei, W. (2011). Comparison of procedures to assee non-linear and time-varying effects in multivariable models for survival data. *Biometrical Journal*, 53(2), 308–331.

Cox, D. R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society Series B*, 34, 187–220.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing using B-splines and penalties. *Statistical Science*, 11, 89–121.

Friedberg, J. W., Mauch, P. M., Rimsza, L. M. and Fisher, R. I. (2008). Non-Hodgkin's lymphomas. In: DeVita, V. T., Lawrence, T. S., Rosenberg, S. A., eds. DeVita, Hellman, and Rosenberg's Cancer: *Principles and Practice of Oncology*. 8th ed. Philadelphia, Pa: Lippincott Williams & Wilkins; 2278–2292.

Hennerfeind, A., Brezger, A. and Fahrmeir, L. (2006). Geoadditive survival models. *Journal Of the American Statistical Association*, 101, 1065–1075.

Kneib, T. and Fahrmeir, L. (2007). A mixed model approach for geoadditive hazard regression. *Scandinavian Journal of Statistics*, 34 207–228.

Shipp, M. A., Harrington, D. P. and Aderson, J. R. *et al.* (1993). A predictive model for aggressive non-Hodgkin's lymphoma. The International non-Hodgkin's lymphoma prognostic factors project. *The New England Journal of Medicine*, 329, 987–994.

Therneau, T. M. and Grambsch, P. M. (2000). *Modelling Survival Data: Extending the Cox Model*. New York: Springer.