Applying Spatial Copula Additive Regression to Breast Cancer Screening Data

Elisa Duarte[1], Bruno de Sousa[2], Carmen Cadarso-Suárez[1], Jenifer Espasandín-Domínguez[1], Oscar Lado-Baleato[1], Giampiero Marra[3], Rosalba Radice[4], and Vítor Rodrigues[5]

[1] Unit of Biostatistics, Department of Statistics, Mathematical Analysis, and Optimization, School of Medicine, University of Santiago de Compostela, Santiago de Compostela, Spain.
[2] Faculty of Psychology and Education Sciences, University of Coimbra, CINEICC, Portugal.
[3] Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, UK.
[4] Department of Economics, Mathematics and Statistics, Birkbeck, University of London, Malet Street, London WC1E 7HX, UK.
[5] Faculty of Medicine, University of Coimbra, Portugal.

Abstract. Breast cancer is associated with several risk factors. Although genetics is an important breast cancer risk factor, environmental and sociodemographic characteristics, that may differ across populations, are also factors to be taken into account when studying the disease. These factors, apart from having a role as direct agents in the risk of the disease, can also influence other variables that act as risk factors. The age at menarche and the reproductive lifespan are considered by the literature as breast cancer risk factors so that, there are several studies whose aim is to analyze the trend of age at menarche and menopause along generations. Also, it is believed that these two moments in a woman's life can be affected by environmental, social status, and lifestyles of women. Using the information of 278,000 registries of women which entered in the breast cancer screening program in Central Portugal, we developed a bivariate copula model to quantify the effect a womans year of birth in the association between age at menarche and a womans reproductive lifespan, in addition to explore any possible effect of the geographic location in these variables and their association. For this analysis we employ CGAMLSS models and the inference was carried out using the R package SemiParBIVProbit.

1 Introduction

The age at menarche and age at menopause are well known breast cancer risk factors, since these moments set a woman's reproductive lifespan, during which the woman is exposed to endogenous hormones responsible to ensure the regular functioning of her reproductive system. There are several factors that affect the beginning and the end of a woman's reproductive lifespan. A downward trend in

the age at menarche has been highlighted in recent researches [1, 2]. The results of a study conducted by [3], shows an upward trend in the number of a woman's reproductive years.

The natural menopause is defined as a complex bio-social and bio-cultural phenomenon [4]. Other studies, such as the one presented in [5], analyse the association between age at menarche and socio-economic characteristics showing that the environmental conditions may have influence in the onset of a woman's reproductive lifespan. Therefore, it should be of the utmost importance to explore how individual characteristics such as age at menarche and a woman's reproductive lifespan can be a reflection of an influence from one's social environment. In addition, rather than analyse the effect of a woman's cohort and of the environment in the age at menarche and reproductive lifespan as two separated response variables, quantify these effects in the association between them is a major topic.

For this analysis we employ Bivariate Copula Additive Models for Location, Scale and Shape. Such models extend the scope of univariate GAMLSS by binding two equations with binary, discrete or continuous responses. The equations can be flexibly specified using smoothers with single or multiples penalties, thus allowing for several types of covariate effects. The copula dependence parameter can also be specified as a function of flexible covariate effects. All the models parameters are estimated simultaneously. The inference is carried out using the R package SemiParBIVProbit [12].

2 Breast Cancer Screening Data

This study is based on data provided by the Central Regional Nucleus of the Portuguese Cancer League (LPCC-NRC), sponsored by the Breast Cancer Screening Program (BCSP) in 78 municipalities located in central Portugal's. Figure Fig. 1 shows the map of Portugal, with the blue regions representing the municipalities under study. The database consists of 278,282 women who were registered for the BCSP in central Portugal between 1990 and 2010.

Women considered in this study have a screening age between 45 and 69, with 76% (212,517) of them reaching menopause. Since we are dealing with the reproductive lifespan cycle of a woman, only the post-menopausal women were considered in the study.

The variables involved in this study are: age of menarche, a woman's reproductive lifespan cycle (calculated by subtracting the age of menarche from the age of menopause), year of birth, and the code of the municipality where a woman resides. Table 1 shows a summary description of these variables.

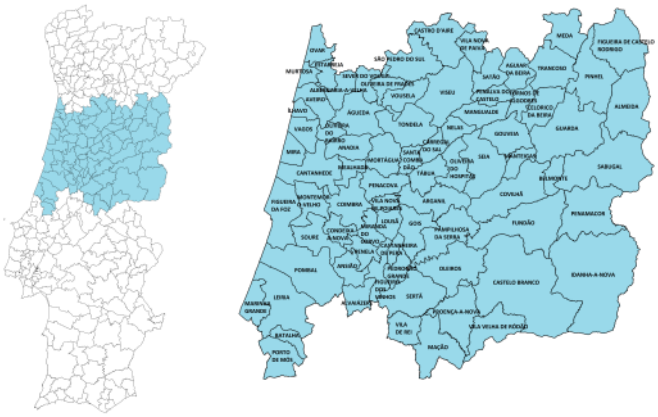

Fig. 1. The map of Portugal, with the blue regions representing the municipalities under study.

Table 1. Statistics of the variables in the study.

Variable	Mean	Standard Deviation (SD)	Min-Max
Birth year	1946	9.8	1920-1965
Age of menarche	13.3	8.0	8-18
Reproductive life span	34.9	5.5	3-50

3 Model Formulation

The main goal of this study is to apply the Bivariate Copula Additive Models for Location, Scale and Shape in order to explain the dependence structure of a bivariate response consisting of age at menarche and a woman's reproductive lifespan. In addition, the model will regress the complete distributional of the response on the year of birth and a woman's place of residence. The Bivariate Copula Additive Models for Location, Scale and Shape extends the use of GAMLSS [6] models to situations in which two responses are modeled simultaneously conditional on some covariates using copula [7]. Using additive predictors, casting several types of covariates such as nonlinear effects of continuous covariates, random effects, interactions or spatial dependence, the approach allows to model a bivariate response consisted of a copula function. Besides that, the regression is not restrict to the response expectation, being able to be extended to other distributional parameters.

One of the strengths of the copula approach is the possibility of the marginal distribution be of different families, providing different types of response dis-

tributions (continuous, discrete, and mixed discrete continuous) as opposed to the classical statistical bivariate response models, that assume each marginal response as Gaussian. A complete description of the CGAMLSS theory can be found in the recent work of Marra and Radice [7, 8].

In the present application of the CGAMLSS, it is considered two bivariate continuous responses, Y_1 and Y_2, representing, respectively, the age at menarche and the reproductive lifespan and covariate information (year of birth and a woman's place of residence) collected in the generic vector z_i. The joint cumulative distribution function (cdf) of Y_1 and Y_2 can be expressed in terms of the marginal cdfs of Y_1 and Y_2 and a copula function C that binds them together [7] as follows:

$$F(y_1, y_2|\vartheta) = C(F_1(y_1|\mu_1, \sigma_1, \nu_1), F_2(y_2|\mu_2, \sigma_2, \nu_2); \theta)$$

where $\vartheta = (\mu_1, \sigma_1, \nu_1, \mu_2, \sigma_2, \nu_2, \theta)^T$, $F_1(y_1|\mu_1, \sigma_1, \nu_1)$ and $F_2(y_2|\mu_2, \sigma_2, \nu_2)$ are the marginal cdfs of Y_1 and Y_2 taking values in $(0, 1)$, μ_m, σ_m, ν_m, for $m = 1, 2$ are the marginal distribution parameters. $C(\cdot, \cdot)$ is a uniquely defined two-place copula function which does not depend on the marginals, and θ is an association copula parameter measuring the dependence between the two random variables [9, 10].

By considering a suitable additive predictors $\eta's$ for all parameters of the bivariate response distribution defined above for an observation i, the predictor could be written as:

$$\eta_i = \beta_0 + \sum_{k=1}^{K} f_k(\mathbf{z_{ki}}), i = 1, ..., n \tag{1}$$

where β_0 is an overall intercept, and the function f_k represent the different covariate effects (as binary, categorial, continuous and spatial variables). The K functions f are chosen according the type of covariate considered (z_{ki}).

As defined in Generalize Additive Models (GAM) [14], each function f_k can be approximated as a linear combination of J_k basis functions $b_{kjk}(z_{ki})$ and regression coefficients $\beta_{kjk} \in \mathbb{R}$, i.e.

$$\sum_{j_k=1}^{J_k} \beta_{kj_k} b_{kj_k}(z_{ki}) \tag{2}$$

Equation (2) implies that the vector of evaluations $\{s_k(z_{k1}), ..., s_k(z_{kn})\}^T$ can be written as $Z_k\beta_k$ with $\beta_k = (\beta_{k1}, ..., \beta_{kJ_k})^T$ and the design matrix $Z_k[i, j_k] = b_{kj_k}(z_{ki})$. This allows the predictor in equation **(2)** to be written as:

$$\eta = \beta_0 1_n + Z_1\beta_1 + ... + Z_K\beta_K \tag{3}$$

where 1_n is an n-dimensional vector made up of ones. Equation **(3)** can also be written in a more compact way as $\eta = Z\beta$ where $Z = (1_n, Z_1, ..., Z_K)$ and $\beta = (\beta_0, \beta_1^T, ..., \beta_K^T)^T$. Each β_k has an associated quadratic penalty $\lambda_k \beta_k^T D_k \beta_k$ with the smoothing parameter λ that controls the trade off between model fit and smoothness.

To model spatial information, Marra and Radice [7] proposed the use of a Markov random field smoother, that is useful in our application where we have the spatial information split up in discrete contiguous geographic units. In this case, $f_k(z_{ki}) = ...$, where β_k represents the vector of spatial effects, R denotes the total number of regions z_{ki}. Thus, the design matrix linking an observation i with the corresponding spatial effect is defined as:

$$Z_k[i, r] = \begin{cases} 1 & \text{if the observation belongs to region r} \\ 0 & \text{otherwise} \end{cases}$$

where $r = 1, \ldots, R$. The smoothing penalty D_λ associated with the Markov random field is constructed based on the neighborhood structure of the geographic units:

$$D_k[r, q] = \begin{cases} -1 & \text{if } r \neq q \bigwedge \text{r and q are adjacent neighbors} \\ 0 & \text{if } r \neq q \bigwedge \text{r and q are not adjacent neighbors} \\ N_r & \text{if } r = q \end{cases}$$

where r and q are two regions and N_r the total number of regions.

The inference is based on penalised maximum likelihood estimation. First, it is considered the log-likelihood function for a copula model with two continuous margins [11]:

$$l(\boldsymbol{\delta}) = \sum_{i=1}^{n} log\left\{C(F_{1i}(y_{1i}|\mu_{1i}, \sigma_{1i}, \nu_{1i}), F_2(y_{2i}|\mu_{2i}, \sigma_{2i}, \nu_{2i}); \theta_i)\right\} + \sum_{i=1}^{n} \sum_{m=1}^{2} log\left\{f_m(y_{mi} \mid \mu_{mi}, \sigma_{mi}, \nu_{mi}))\right\}$$

where parameter $\boldsymbol{\delta}$ is defined as $(\beta_{\mu 1}^T, \beta_{\mu 2}^T, \beta_{\sigma 1}^T, \beta_{\sigma 2}^T, \beta_{\nu 1}^T, \beta_{\nu 2}^T, \beta_\rho^T)^T$.
The use of a classic unpenalized optimization algorithm is likely to result unduly wiggly estimates, therefore Marra and Radice (2016) [7] proposes a penalised maximum likelihood estimation of the form:

$$l_p(\boldsymbol{\delta}) = l(\delta) - \frac{1}{2}\boldsymbol{\delta}^T \boldsymbol{S_\lambda} \boldsymbol{\delta} \tag{4}$$

where $S_\lambda = diag(\lambda_{\mu 1} D_{\mu 1}, \lambda_{\mu 2} D_{\mu 2}, \lambda_{\sigma 1} D_{\sigma 1}, \lambda_{\sigma 2} D_{\sigma 2}, \lambda_{\nu 1} D_{\nu 1}, \lambda_{\nu 2} D_{\nu 2}, \lambda_\rho D_\rho)$ with each smoothing parameters related to the corresponding D component and the

overall λ is defined as $(\lambda_{\mu 1}^T, \lambda_{\mu 2}^T, \lambda_{\sigma 1}^T, \lambda_{\sigma 2}^T, \lambda_{\nu 1}^T, \lambda_{\nu 2}^T, \lambda_{\rho}^T)^T$.

To estimate the regression coefficients the CGAMLSS methodology use a two-step algorithm, in the first step it estimates the δ that maximize the log-likelihood function using a trust region algorithm which is generally more stable and faster than the line search methods such as Newton-Raphson, particularly for functions that are, for example, non-concave and/or exhibit regions that are close to flat [13]. In the second step the algorithm estimate the smoothing parameter λ, using an expression that is equivalent to the Un-Biased Risk Estimator (UBRE) given in Wood (2006, Chapter 4)[14], solved with the methodology proposed by Wood in 2004 [15].
In the CGAMLSS approach the researcher should decide about the distribution to use for the margins of the bivariate response, as well as the copula that best modelize the structure of dependence between this margins.

In our study, from the continuous distribution families available in the SemiParBIVProbit package [12], a Log-normal distribution for the age at menarche and a Gumbel to the reproductive lifespan of the woman were chosen. This choice was based on the AIC (Akaike information criterion) and on the BIC (Bayesian information criterion). Both distributions are defined by two parameters: a location parameter μ and a scale parameter σ,thus the equation model can be defined as follows:

$$\begin{cases} \eta_i^{\mu_1} = \beta_i^{\mu_1} + f_i^{\mu_1}(Year\ of\ birth) + f_i^{\mu_1}(Municipality) \\ \eta_i^{\sigma_1^2} = \beta_i^{\mu_1} + f_i^{\mu_1}(Year\ of\ birth) + f_i^{\mu_1}(Municipality) \\ \eta_i^{\mu_2} = \beta_i^{\mu_1} + f_i^{\mu_1}(Year\ of\ birth) + f_i^{\mu_1}(Municipality) \\ \eta_i^{\sigma_2^2} = \beta_i^{\mu_1} + f_i^{\mu_1}(Year\ of\ birth) + f_i^{\mu_1}(Municipality) \\ \eta_i^{\theta} = \beta_i^{\mu_1} + f_i^{\mu_1}(Year\ of\ birth) + f_i^{\mu_1}(Municipality) \end{cases} \tag{5}$$

The two first equations refer to the location and scale parameter of the age at menarche, the next two refer to the location and scale parameter of the reproductive lifespan of women and the last one refers to the association between both variables. All parameters were modeled using predictors involving a continuous (Year of birth) and spatial covariate (Municipality). The former was modeled using penalized low rank regression splines and the latter using a Markov random field smoother.

During the model building process we have tried a set of copulas which AIC, BIC and run-time information are presented in Table 2. Run-time is the time that the model required to reach the optimal estimation of the regression parameters. The inference was carried out in a Intel(R) Core(TM) i5-4570s CPU 2.90 GHz with operating system Windows 7 Professional.

For the choice of copula we start off with the gaussian, from which was observed a negative association between the marginals. Therefore, it was not

performed any fit with the Clayton copula rotated 90 degrees. In addition, due to the value of the range of the τ of kendall of the marginals (-0.223,-0.191), the Ali-Mikhail-Haq (AMH) and Farlie-Gumbel-Morgenstern (FGM) copulas were not tried, since they only modelize weak dependencies, below -0.18 and -0.22, respectively. Based on AIC and run-time values, the selected model is a Gaussian copula which quantile residuals are shown in figure Fig. 2.

Table 2. Copula used during the model building process ordered by their AIC.

Family	AIC	BIC	Run.time
Gaussian	2100208	2103537	24' 15"
Gumbel (90)	2101223	2101223	34' 28"
Frank	2101914	2105217	28' 17"
Clayton (180)	2102480	2105722	52' 05"
Gumbel (270)	2105144	2108527	54' 46"
Joe (90)	2105342	2108590	43' 04"
Clayton (270)	2108901	2112265	29' 32"
Joe (270)	2111979	2115350	33' 17"
Clayton	2123078	2125939	49' 33"
Joe	2123078	2125939	44' 13"
Joe (180)	2123078	2125939	46' 39"
Gumbel	2123078	2125939	50' 27"
Gumbel (180)	2123078	2125939	52' 00"

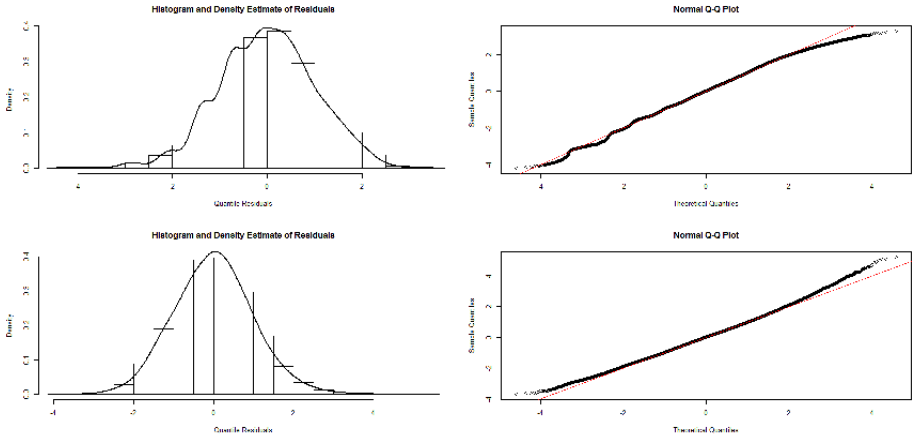

Fig. 2. Quantile residuals of the selected model

4 Results

The estimates of the smooth effect of the year of birth with the associated 95% point-wise intervals, and the spatial effects on the parameters of the marginal distributions of the variables age at menarche and a woman's reproductive lifespan, are presented in Fig. 3 to Fig. 7. The left-hand side of Fig.3 shows a clear decreasing effect of the year of birth on the expectation of the age at menarche. Regarding to the effect on the expectation of a woman's reproductive lifespan, Fig. 4 shows an increasing effect along the cohorts before 1952, followed by a sharp decrease. The spatial effects presented in right-hand side of the same figures, show the inland regions of central Portugal associated with lower ages at menarche and higher reproductive lifespans.

The effects of the year of birth on the variance of the marginal distributions of the age at menarche and reproductive lifespan, are shown in the left-hand side of the figures Fig. 5 and Fig. 6, respectively. In the first, it is observed a downward effect on the variance of age at menarche until the year 1955 followed by a slight upward effect. The second, shows a steady effect of year of birth prior to 1938, from which it starts decreasing up to the year 1949, and followed by a significantly decline. Looking to the spatial effects on the right hand-side of the figure Fig. 5, it is clear a east-west increasing effect of the variance of the age at menarche. A weak spatial effect of a woman's reproductive lifespan variance in almost all of Central Portugal's municipalities is depicted in figure Fig. 6. Nevertheless, notice that the municipalities located in the northeast part of the region have a marked negative effect on the variability of the lifespan.

The effect of the year of birth on the association between age at menarche and a woman's reproductive lifespan (Fig. 7) shows a decreasing effect until 1930, followed by a slow increase from a negative to a positive association until 1950, decreasing afterwards. Regarding the spatial effects (right-hand side of Fig. 7), only a couple of municipalities in the north part of the region show a negative effect in this association.

5 Discussion

This study was conducted in order to apply the Bivariate Copula Additive Models for Location, Scale and Shape to quantify the effect of a woman's cohort and of the environment (represented by a woman's place of residence) in the association between age at menarche and a woman's reproductive lifespan and the effects of this covariates in the location and scale parameters of the marginal distributions. This approach allows to assess easily a suitable marginal distributions for the response variables, and find the best fitted copula additive model.

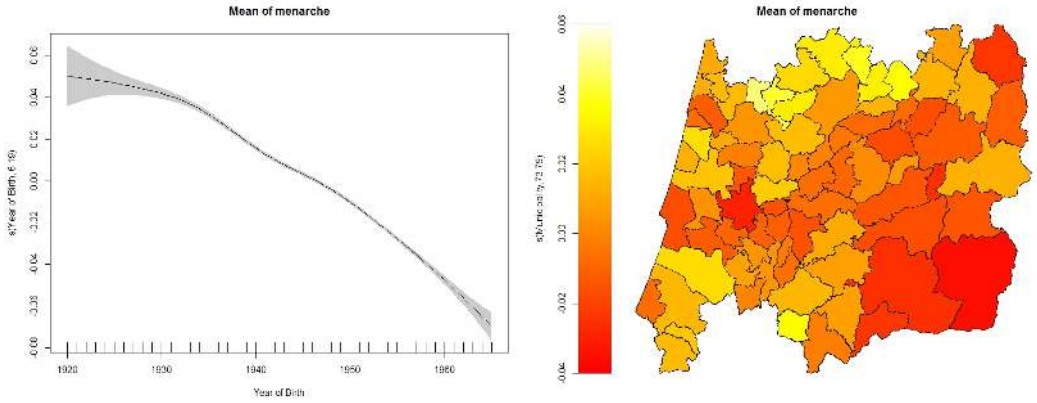

Fig. 3. Effect of the year of birth and spatial differences on the age at menarche.

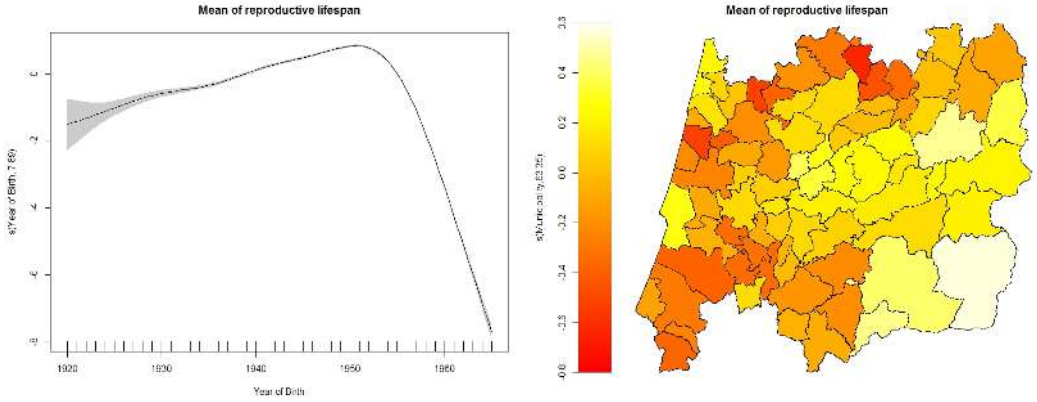

Fig. 4. Effect of the year of birth and spatial differences on the reproductive lifespan.

The package SemiParBIVProbit [12] has shown a good performance in the fit of models with big datasets such as the breast cancer screening. To our knowledge, this is the first time that the CGAMLSS is applied to a database with a considerable number of records, with the selected copula function converging in 24' 15' in a Intel(R) Core(TM) i5-4570s CPU 2.90 GHz with operating system Windows 7 Professional.

The results achieved suggested that earlier menarche is associated with younger women. An increasing of a woman's reproductive lifespan is observed, followed by a sharp decrease for women born after 1952. This drop is justified by the fact that women born after 1952 are those cases who already reported a menopause, despite of their young age. The decreasing effect of year of birth in the variabil-

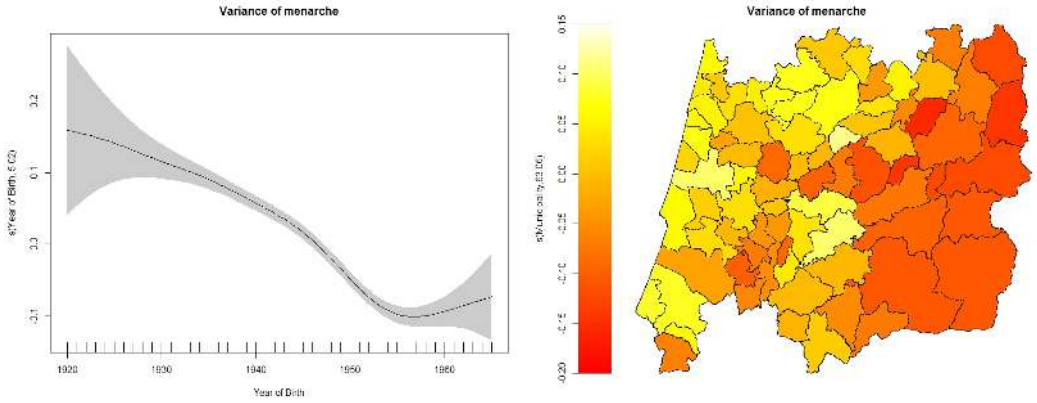

Fig. 5. Effect of the year of birth and spatial differences on the variability of the age at menarche.

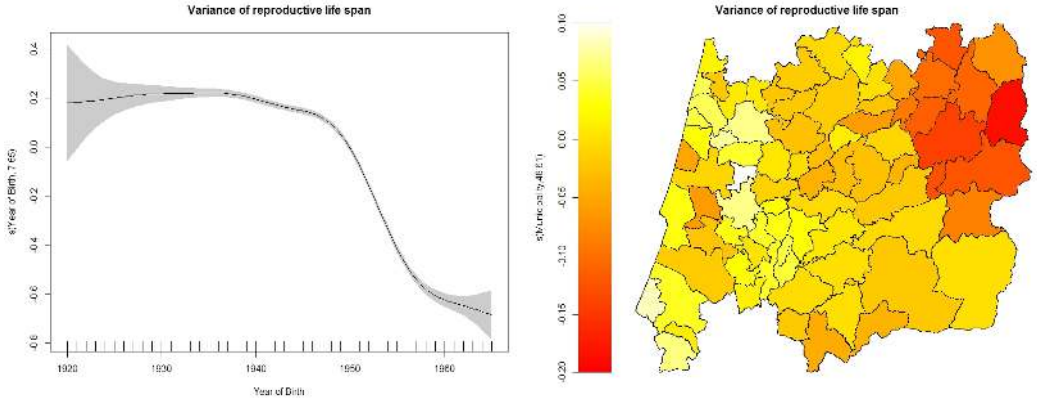

Fig. 6. Effect of the year of birth and spatial differences on the variability of the reproductive lifespan.

ity of the age at menarche may be explained by the fact that that moment is easily remembered by younger women. Since the west region in central Portugal is in general wealthier and more economically developed region in comparison to interior part, the expectation of an increasing effect on a woman's reproductive lifespan and a decreasing one on age at menarche in the east-west direction was not verified. These may be explained by factors that were not taken into account in the model, leading to the conclusion that a woman's place of residence is not the only factor that may affect women's individual characteristics, such as age at menarche and a woman's reproductive lifespan cycle.

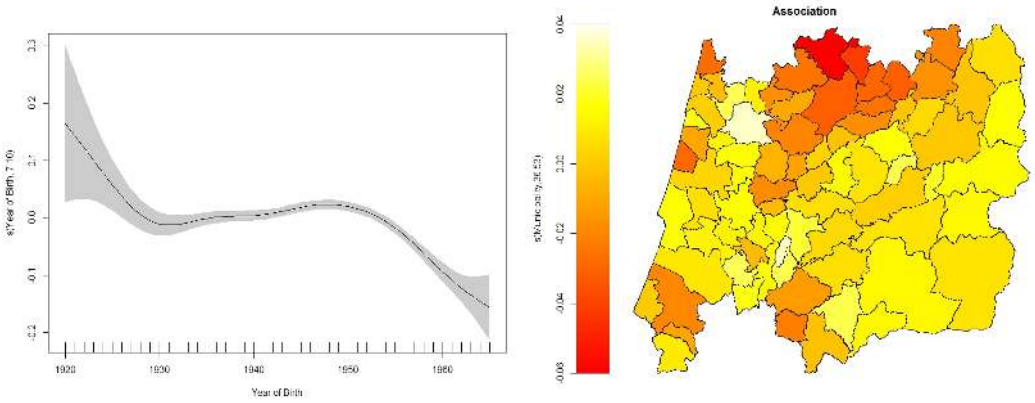

Fig. 7. Effect of the year of birth and spatial differences on the association between the menarche and reproductive lifespan.

References

1. Talma, H., Schönbeck, Y., van Dommelen, P., Bakker, B., van Buuren, S., HiraSing, R. A.: Trends in Menarcheal Age between 1955 and 2009 in the Netherlands. PLoS ONE 8(4): e60056. doi:10.1371/journal.pone.0060056. (2013)
2. Rigon, F., Bianchin, L., Bernasconi, S., Bona, G., Bozzola, M., Buzi, F.,*et al.*: Update on Age at Menarche in Italy: Toward the Leveling Off of the Secular Trend. Journal of Adolescent Health 46, 238-244.(2010)
3. Nichols, H. B., Trentham-Dietz, A., Hampton, J. M., Titus-Ernstoff, L., Egan, K. M., Willett, W. C., Newcomb, P. A.: From Menarche to Menopause: Trends among US Women Born from 1912 to 1969. American Journal of Epidemiology 164(10), 1003–1011. doi: 10.1093/aje/kwj282 (2006).
4. Kaczmarek,M.: The timing of natural menopause in Poland and associated factors. Maturitas 57 139-153 (2007).
5. Wronka, I., Pawliska-Chmara, R.: Menarcheal age and socio-economic factors in Poland. Annals of Human Biology 32 (5) 630–638 doi:10.1080/03014460500204478. (2005)
6. Rigby A. and Stasinopoulos D. M.: Generalized additive models for location, scale and shape (with discussion). Applied Statistics 54, 507 – 554 (2005).
7. Marra G., Radice R.: A Bivariate Copula Additive Model for Location, Scale and Shape. Cornell University Library. arXiv:1605.07521 [stat.ME] (2017).
8. Marra G., Radice R., Barnighausen T., Wood S.N., McGovern M.E.: A Simultaneous Equation Approach to Estimating HIV Prevalence with Non-Ignorable Missing Responses, Journal of the American Statistical Association (in press).
9. Sklar, A.: Fonctions de rpartition n dimensions et leurs marges. Publications de l'Institut de Statistique de l'Universit de Paris 8 229-231 (1959).
10. Sklar, A.: Random variables, joint distributions, and copulas. Kybernetica 9, 449 – 460 (1973).
11. Kolev, Nikolai, and Delhi Paiva. "Copula-Based Regression Models." Department of Statistics, University of So Paulo. nkolev@ ime. usp. br (2007).

12. Marra, Giampiero, Rosalba Radice, and Maintainer Giampiero Marra. "Package SemiParBIVProbit." (2016).
13. Nocedal, J. and Wright, S. J. (2006). Numerical Optimization. New York: Springer-Verlag
14. Wood, Simon. Generalized additive models: an introduction with R. CRC press, 2006.
15. Wood, Simon N. "Stable and efficient multiple smoothing parameter estimation for generalized additive models." Journal of the American Statistical Association 99.467 (2004): 673-686.