Revista Colombiana de Estadística - Applied Statistics
January 2026, volume 49, issue 1, pp. 293–305
http://doi.org/10.15446/rce.v49n1.117781

Semiparametric Modelling of Cancer Mortality Trends in Colombia

Modelamiento semiparamétrico de las tendencias de mortalidad por cáncer en Colombia

Lina Angélica Buitrago-Reyes[1,a], Juan Sosa[1,b], Cristian Andrés González-Prieto[2,c]

[1]Departamento de Estadística, Facultad de Ciencias, Universidad Nacional de Colombia, Bogotá, Colombia

[2]School of Computer Science, University of Auckland, Auckland, New Zeland

Abstract

In this paper, we compare semiparametric and parametric model adjustments for cancer mortality in breast and cervical cancer in women and gastric and lung cancer in men, according to age and period of death. Semiparametric models were adjusted for the number of deaths from the two localizations of greatest mortality by sex: breast and cervix in women; prostate and lungs in men. Adjustments in different semiparametric models were compared, which included making adjustments with different distributions and variable combinations in the parametric and non-parametric part, for localization as well as for scale. Finally, the semiparametric model with best adjustment was selected and compared to traditional model; that is, to the generalized lineal model with Poisson response and logarithmic link. Best results for the four kinds of cancer were obtained for the selected semiparametric model by comparing it to the traditional Poisson model based upon AIC, envelope correlation between estimated logarithm rate and real rate logarithm. In general, we observe that in estimation, rate increases with age; however, with respect to period, breast cancer and stomach cancer in men show a tendency to rise over time; on the other hand, for cervical cancer, it remains virtually constant, but for lung cancer in men, as of 2007, it tends to decrease.

Keywords: Cancer mortality; Log-symmetric models; Natural cubic splines; Poisson model; Semiparametric modeling.

[a]Ph.D (c). E-mail: labuitragor@unal.edu.co
[b]Ph.D. E-mail: jcsosam@unal.edu.co
[c]Ph.D. E-mail: cgon080@aucklanduni.ac.nz

Resumen

En este artículo se comparan ajustes de modelos semiparamétricos y paramétricos para la mortalidad por cáncer de mama y de cuello uterino en mujeres, y por cáncer gástrico y de pulmón en hombres, de acuerdo con la edad y el periodo de defunción. Se ajustaron modelos semiparamétricos para el número de muertes por las dos localizaciones de mayor mortalidad según sexo: mama y cuello uterino en mujeres, próstata y pulmón en hombres. Se compararon ajustes en diferentes modelos semiparamétricos, que incluyeron ajustes con distintas distribuciones y combinaciones de variables en la parte paramétrica y no paramétrica, tanto para la localización como para la escala. Finalmente, se seleccionó el modelo semiparamétrico con mejor ajuste y se comparó con el modelo tradicional, es decir, con el modelo lineal generalizado con respuesta Poisson y enlace logarítmico. Los mejores resultados para los cuatro tipos de cáncer se obtuvieron con el modelo semiparamétrico seleccionado al compararlo con el modelo Poisson tradicional, con base en el AIC y en la correlación envolvente entre el logaritmo de la tasa estimada y el logaritmo de la tasa real. En general, se observa que, en la estimación, la tasa aumenta con la edad. Sin embargo, respecto al periodo, el cáncer de mama y el cáncer de estómago en hombres muestran una tendencia creciente a lo largo del tiempo. Por otra parte, para el cáncer de cuello uterino se mantiene prácticamente constante, pero para el cáncer de pulmón en hombres, a partir de 2007, tiende a disminuir.

***Palabras clave*:** Modelación semiparamétrica; Modelo de Poisson; Modelos log-simétricos; Mortalidad por cáncer; Splines cúbicos naturales.

1. Introduction

Cancer, which can occur at any age, is a disease characterized by rapid and abnormal cell proliferation in any bodily organ (World Health Organization, 2017). Carcinogenic cells invade adjacent organs and later spread to other sites in the body through metastasis, thereby altering vital functions and, in many cases, causing death. Major cancer risk factors include tobacco use, poor diet (National Cancer Institute, 2015), and chronic infections caused by bacteria, viruses, or parasites (Ohshima & Bartsch, 1994).

Furthermore, cancer mortality is ranked among the highest worldwide. In 2015, according to the World Health Organization (WHO), 8.8 million cancer related deaths were registered worldwide (World Health Organization, 2017), among the most common were lung, hepatic, gastric, colorectal, and breast cancers. National health care institutes worldwide monitor cancer mortality as part of their follow up efforts on local cancer control policies and cancer mitigation strategies. These efforts include improving access to early detection programs and improving diagnosis and treatment, as well as identifying the most frequent cancers, thereby allowing strategies to be developed that focus on specific cancers.

Between 2000 and 2006, in Colombia, 203,907 cancer deaths were registered, and the most frequent cancers were stomach, lung, cervical, colon, and breast (Piñeros et al., 2010). Worldwide, different methods have been used to model

temporal cancer mortality trends. For example, in 2009, Cabanes et al. modeled mortality for three types of cancer, breast, ovarian, and cervical, among the female population in Spain from 1980 to 2006. They used a generalized linear model with a Poisson response, taking age into account, to evaluate changes in overall and cancer specific mortality rates over time (Cabanes et al., 2009). However, nothing has been reported on the value of this adjustment, nor on the evaluation of the underlying assumptions.

Clèries et al. (2010) used a Bayesian age drift model with a Poisson response and different priors, according to the adjusted model, to describe the trend in testicular cancer among men aged 15 to 74 years in Spain from 2005 to 2019 (Clèries et al., 2010). Additionally, they fitted an autoregressive APC model to estimate projections for years with no information and selected the best model for each case based on the deviance information criterion (DIC). For prostate cancer mortality data in Norway from 1980 to 2007, Kvåle et al. fitted a joinpoint regression model to identify linear changes in the mortality trend for this cancer type. However, no results are reported for the model quality (Kvåle et al., 2010).

Guo & Li (2012) modeled the trend in esophageal cancer mortality in China between 1987 and 2009 (Guo & Li, 2012). To do so, they modeled age, period, and birth cohort effects using a generalized additive model (GAM) with a Poisson response, where nonlinear associations between predictors are represented through nonparametric techniques. Model fit was good in this case, however, the evaluation of assumptions for this family of models was not reported.

In Colombia, cancer mortality rates in the Cancer Mortality Atlas (Piñeros et al., 2010) were modeled with generalized linear models, where the number of deaths is assumed to follow a Poisson distribution whose rate is explained by age and period, both observed from 2000 to 2006. The Colombian National Statistics Agency, DANE, provides the official registries used as sources for these mortality data. However, these models are not the most appropriate for this type of data because they may exhibit overdispersion (Berk & MacDonald, 2008). Therefore, it is necessary to consider a new methodology that overcomes these limitations and improves estimation and overall model fit.

In this article, we propose the use of semiparametric models (Vanegas & Paula, 2015) for modeling cancer mortality data, and we compare them with the Poisson models commonly used in practice. This comparison considers assumption evaluations, the Akaike information criterion, and the relation between estimated rates and observed rates for four cancer types, namely the two most frequent in men, stomach and lung, and the two most frequent in women, breast and cervical. These models allow semiparametric modeling of the median and bias.

The methods section describes the data used to compare models and provides a brief description of the semiparametric models considered. The results section presents model fits obtained with semiparametric models and the generalized linear Poisson model, and the discussion section addresses the relevance of semiparametric models and the advantages they offer relative to Poisson models.

2. Materials and Methods

Mortality data for the 1994 to 2013 period were obtained from the official registries of the *Departamento Nacional de Estadística* (DANE). Cause of death is coded according to the International Classification of Diseases (ICD), from 1994 to 1997 using ICD 9, and later using ICD 10 (World Health Organization, 2016). Regarding population, DANE projections by age and sex were used (DANE, 2017).

The two cancers in men with the highest mortality rates, prostate and lung, and in women, breast and cervical, were selected for this study. The semiparametric models described the median and bias for each cancer studied (Vanegas & Paula, 2015). Death counts were obtained by age and year of death, which corresponds to period. Therefore, the explanatory variables were age group and period midpoint. Simultaneously, using the same explanatory variables, generalized linear models with a Poisson response were fitted, and models were compared using measures of fit such as the Akaike information criterion and the correlation between the logarithm of the estimated rate and the logarithm of the observed rate.

2.1. Semiparametric Models for Modelling Median and Skewness

Let $T_1, T_2, \ldots, T_n$ be assumed to be independent random variables, for which the general model structure is given in (Vanegas & Paula, 2015),

$$T_k = \eta_k \xi_k^{\sqrt{\phi_k}}, \qquad k = 1, \ldots, n, \tag{1}$$

where η_k and $\phi_k > 0$ are the median and skewness of T_k, and $\xi_1, \ldots, \xi_n$ are random independent multiplicative errors with a log symmetric distribution:

$$\xi_k \overset{\text{i.i.d.}}{\sim} \mathcal{LS}(1, 1, g(\cdot)).$$

Taking logarithms in (1), we get

$$Y_k = \mu_k + \sqrt{\phi_k}\epsilon_k, \qquad k = 1, \ldots, n,$$

where $Y_k = \log T_k$, $\mu_k = \log \eta_k$ is the location parameter of Y_k, ϕ_k is the dispersion parameter of Y_k, and $\epsilon_k = \log \xi_k$ has a symmetric distribution around zero with dispersion 1, i.e., $\epsilon_k \overset{\text{i.i.d.}}{\sim} \mathcal{S}(0, 1, g(.))$.

It is assumed that η_k and ϕ_k follow the form

$$\eta_k = \eta(\boldsymbol{x}_k, \boldsymbol{\beta}), \qquad \log \phi_k = \boldsymbol{w}_k^\top \boldsymbol{\gamma} + f(\boldsymbol{b}_k),$$

where $\boldsymbol{x}_k$, $\boldsymbol{w}_k$, and $\boldsymbol{b}_k$ are explanatory variables, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are parameter vectors, and $f(\cdot)$ is a continuous and twice differentiable nonparametric function of $\boldsymbol{b}_k$. In this specific case, T_k is the number of deaths, the explanatory variable vectors contain age group and the death period midpoint, and the offset is the logarithm of the population.

In semiparametric log symmetric models, both the location function and the dispersion function may incorporate nonparametric components to flexibly capture nonlinear effects of the covariates. In practice, these components can be modeled using spline based smoothers, such as natural cubic splines or P splines, allowing the model to adapt to complex age and period patterns without imposing a restrictive parametric structure. This flexibility is particularly useful in mortality analyses, where nonlinear trends over age or calendar time are expected. In this study, spline functions were used to model the nonparametric parts of the location and dispersion submodels.

2.2. Generalized Lineal Model with Poisson Response

This type of model is the most commonly used for modelling the number of cases of a particular event. In this case, the fitted models for each cancer type take the form

$$\mathsf{E}(T_i) = \log \mu_i = \log p_i + \boldsymbol{x}_i^\top \boldsymbol{\beta},$$

where T_i is the number of deaths, $\boldsymbol{x}_i$ is the vector of explanatory variables, which includes the midpoint of the age group and the midpoint of the death period, and p_i is the population exposed in the corresponding age group and period. All calculations were performed using `R` software (R Core Team, 2016). Semiparametric models were fitted using the `ssym` package (Vanegas & Paula, 2016).

3. Results

3.1. Breast Cancer

During the study period, 36,012 deaths from breast cancer occurred. The semiparametric model with the best fit was the one fitted with a contaminated normal distribution, whose location parameter was modeled using natural cubic splines for age and for period, and whose scale parameter was modeled using natural cubic splines for age. Compared with the traditional model, the semiparametric model showed a better fit, since a much lower AIC was obtained. A good fit was observed in the envelopes, as well as in the median and bias, as shown in Figure 1, and the correlation between estimated rates and observed rates was higher, as shown in Table 1 and Figure 2.

When analyzing the semiparametric model results based on the median fit, it was estimated that breast cancer mortality increases rapidly between ages 20 and 50 years. It then decreases slightly until age 75 years, when it begins to increase again. Regarding period, an increasing pattern was observed during the entire study period, however, this increase is larger from 2003 onward. The bias estimate decreases until age 55 years, and then begins to increase slowly, as shown in Figure 3.

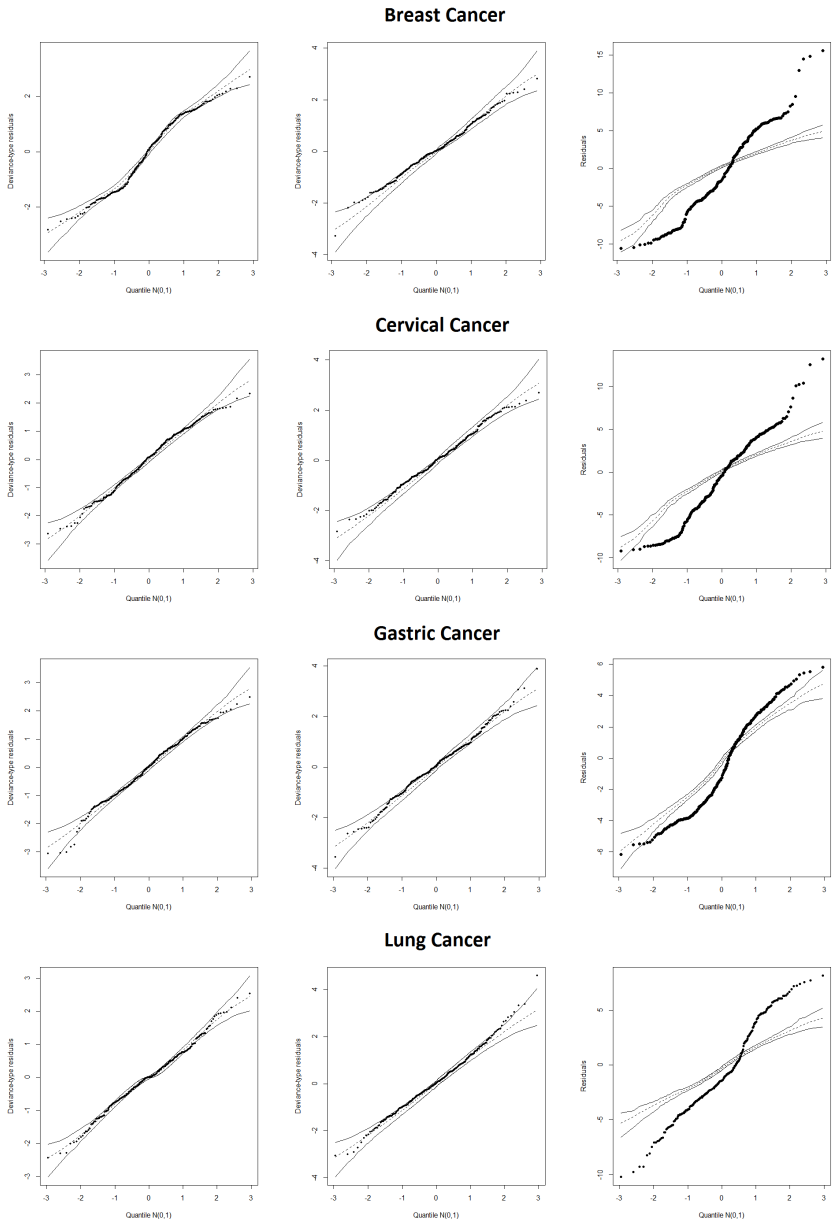

FIGURE 1: Left, envelope for the location parameter model in the semiparametric model. Center, envelope for the bias parameter in the semiparametric model. Right, envelope for the Poisson model.

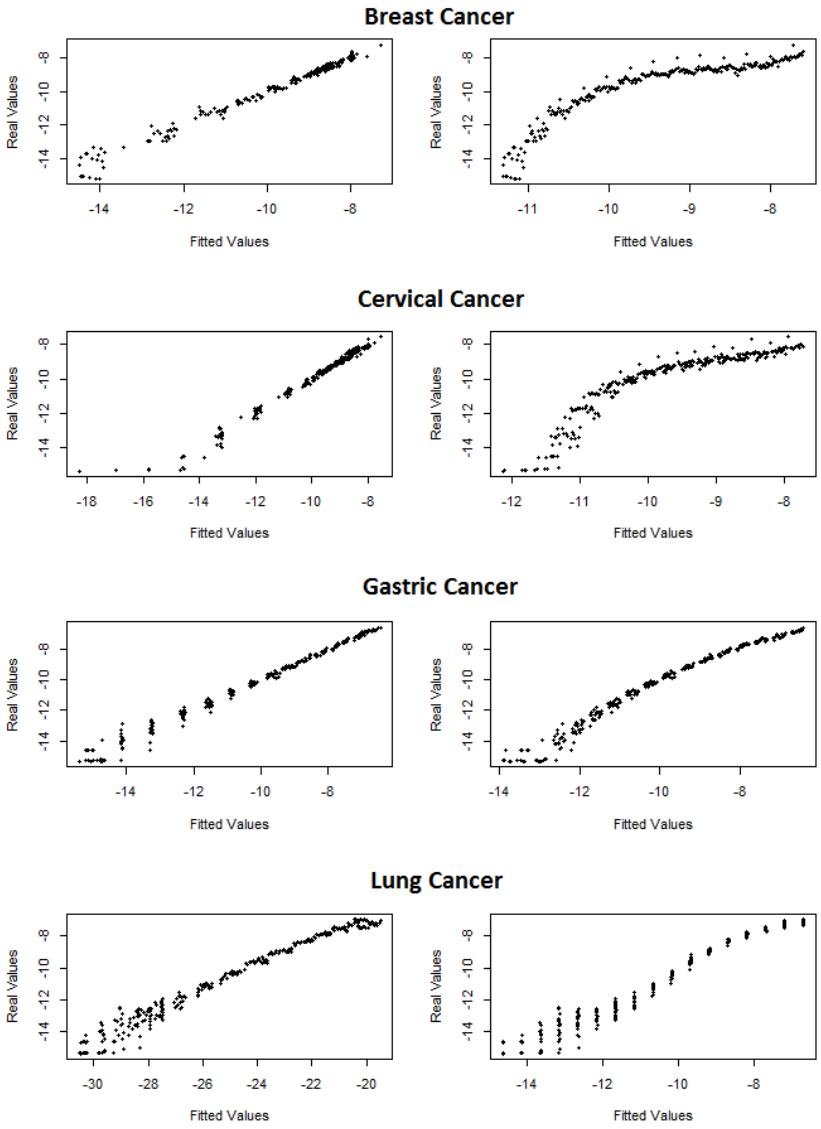

FIGURE 2: Estimated rate versus observed rate. Left, semiparametric model. Right, Poisson model.

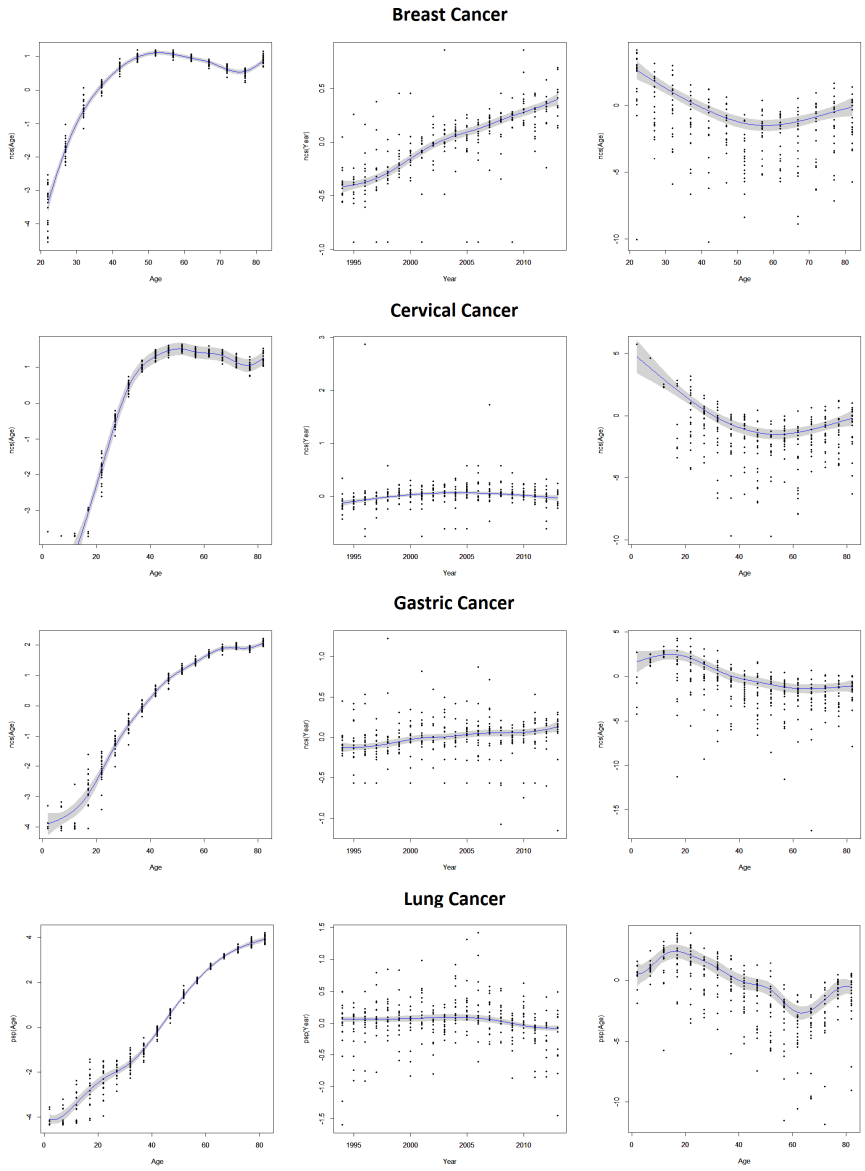

FIGURE 3: Graphs of the nonparametric model components. For the location submodel, the age and period components are shown in the left and central columns, respectively. In the right column, the age curve for the skewness submodel is shown.

3.2. Cervical Cancer

A total of 31 181 deaths from cervical cancer occurred, and the semiparametric model with the best fit was the normal one. The location component was modeled using natural cubic splines for age and for period, and the scale parameter was modeled using natural cubic splines for age. Compared with the traditional Poisson model, the semiparametric model showed a better fit, since a much lower AIC was obtained. In addition, a good fit was observed in the median and bias envelopes, as shown in Figure 2. In contrast, for the Poisson model, residuals fell completely outside the confidence bands, and the correlation between estimated rates and observed rates was higher for the semiparametric model, as shown in Table 2 and Figure 2.

When analyzing the semiparametric model results based on the median fit, the age pattern is similar to that of breast cancer mortality, since a rapid increase is observed, this time from ages 15 to 45 years. It then decreases very slightly until age 75 years, when it begins to increase again. Regarding period, the estimated mortality shows a parabolic shape and remains almost constant, reaching its peak in 2004. With respect to bias, the estimate decreases until age 55 years, and then begins to increase slowly, as shown in Figure 3.

TABLE 1: Mortality models for breast cancer. ρ, correlation between the estimated and observed values.

Semiparametric Model				**Poisson Model**			
Median							
Parametric	Estimate	Std. Err.	p value		Estimate	Std. Err.	p value
Intercept	4.311	0.016	< 0.005	Intercept	-40.790	1.845	< 0.005
				age	0.058	3.08×10^{-4}	< 0.005
				period	0.014	9.21×10^{-4}	< 0.005
Nonparam	Smooth param	d.f.	p value				
`nsc(age)`	706.700	7.846	< 0.005				
`ncs(period)`	2376.900	5.607	< 0.005				
Skewness							
Parametric	Estimate	Std. Err.	p value				
Intercept	-5.955	0.100	< 0.005				
Nonparam							
`nsc(age)`	3547.000	3.230	< 0.005				
AIC	-263.330			AIC	8832.100		
ρ	0.992			ρ	0.883		

3.3. Gastric Cancer

For gastric cancer, 51841 deaths ocurred among men. The semiparametric model with the best fit was the normal one. The location parameter was modeled using natural cubic splines for age and for period. The scale parameter was modeled using natural cubic splines for age. When the semiparametric model was compared with the traditional Poisson model, the former showed a better fit, as reflected in the AIC comparison, since the semiparametric model has a much smaller AIC. In addition, the median and bias envelopes suggest a better fit than the Poisson model, for which residuals move outside the bands, as shown in

Figure 1. The correlation between the estimated and observed values is higher for the semiparametric model than for the Poisson model, as shown in Table 3.

TABLE 2: Mortality models for cervical cancer. ρ, correlation between the estimated and observed values.

Semiparametric Model				**Poisson Model**			
Median							
Parametric	Estimate	Std. Err.	p value		Estimate	Std. Err.	p value
Intercept	3.666	0.055	< 0.005	Intercept	32.187	1.951	< 0.005
				age	0.055	3.00×10^{-4}	< 0.005
				period	-0.022	9.70×10^{-4}	< 0.005
Nonparam	Smooth param	d.f.	p value				
`nsc(age)`	2261.000	7.650	< 0.005				
`ncs(period)`	35689.000	2.937	< 0.005				
Skewness							
Parametric	Estimate	Std. Err.	p value				
Intercept	-3.595	0.097	< 0.005				
Nonparam							
`nsc(age)`	7538.000	3.252	< 0.005				
AIC	-305.784			AIC	7416.200		
ρ	0.990			ρ	0.908		

Based on the median fit provided by the semiparametric model, a pattern similar to that observed for the previous cancer types is evident, namely, an accelerated increase in mortality as age increases. Moreover, mortality increases as period advances. For the bias component, the maximum value occurs at age 20 years, after which it decreases, as shown in Figure 3.

3.4. Lung Cancer

A total of 40 461 deaths from lung cancer occurred among men. The semiparametric model with the best fit is the one based on the exponential power distribution. P splines were used for the scale parameter and for modeling the location component in age and period.

When comparing mortality modeling between the semiparametric model and the Poisson model, it is apparent that the former provides a better fit than the latter. The AIC for the semiparametric model is lower than that for the Poisson model, as shown in Table 4. When examining the envelope graphs, the Poisson model residuals move outside the bands, whereas this does not occur for the semiparametric model residuals, as shown in Figure 1. The correlation between estimated values and observed values is higher for the semiparametric model than for the Poisson model, as illustrated in Figure 2.

The median fit from the semiparametric model shows that mortality increases as age advances. However, it remains constant and tends to decrease as years progress. For bias, the maximum value occurs at 20 years of age and the minimum value occurs at 60 years of age, as shown in Figure 3.

TABLE 3: Mortality models for gastric cancer in men. ρ, correlation between the estimated and observed values.

Semiparametric Model				**Poisson Model**			
Median							
Parametric	Estimate	Std. Err.	p value		Estimate	Std. Err.	p value
Intercept	4.000	0.018	< 0.005	Intercept	17.403	1.534	< 0.005
				age	0.091	2.89×10^{-4}	< 0.005
				period	-0.016	7.66×10^{-4}	< 0.005
Nonparam	Smooth param	d.f.	p value				
`nsc(age)`	718.700	7.910	< 0.005				
`ncs(period)`	2285.600	5.422	< 0.005				
Skewness							
Parametric	Estimate	Std. Err.	p value				
Intercept	-3.728	0.090	< 0.005				
Nonparam							
`nsc(age)`	809.400	5.627	< 0.005				
AIC	-237.782			AIC	4448.300		
ρ	0.995			ρ	0.984		

TABLE 4: Mortality models for lung cancer in men. ρ, correlation between the estimated and observed values.

Semiparametric Model				**Poisson Model**			
Median							
Parametric	Estimate	Std. Err.	p value		Estimate	Std. Err.	p value
Intercept	-11.089	0.022	< 0.005	Intercept	11.990	1.747	< 0.005
				age	0.099	0.344	< 0.005
				period	-0.001	0.872	0.103
Nonparam	Smooth param	d.f.	p value				
`psp(age)`	586.851	3.121	< 0.005				
`psp(period)`	1.111	8.999	< 0.005				
Skewness							
Parametric	Estimate	Std. Err.	p value				
Intercept	71.237	2.698	0.008				
Nonparam	Smooth param	d.f.	p value				
`psp(age)`	2.72×10^{-3}	7.578	< 0.005				
`psp(period)`	0.058	8.793	< 0.005				
AIC	-67.937			AIC	6199.900		
ρ	0.988			ρ	0.984		

4. Discussion

Cancer mortality trend modeling is relevant for analyzing the impact that cancer control strategies may have. As these strategies are implemented, trends in cancer mortality become evident, and modeling can play an essential role in supporting their ongoing implementation.

Mortality rates, as expected, increase with age. However, for breast cancer and cervical cancer, rates increase until approximately age 50, and then mortality begins to decrease slightly, a pattern that may be explained by reduced estrogen production during this stage of women's lives, as shown in Figure 3 (Pike et al., 1993)(Marchant, 1982). The breast cancer trend shows no apparent decrease over

time. On the contrary, despite multiple early detection campaigns (Instituto Nacional de Cancerología et al., 2013), it tends to increase, as shown in Figure 3. For cervical cancer, the trend remains constant. For stomach cancer in men, the trend increases slightly over time. In contrast, lung cancer in men decreases slightly during the same period.

Regarding model comparisons, it was shown that the traditional model had poorer fit, since higher AIC values were obtained and residual overdispersion was observed, as shown in Figure 1 and corroborated in Figure 3. Another finding consistent with this result was the higher linearity and lower variance between predictions and observed values, as shown in Figure 2. Considering these results, it is apparent that the relation between age, period, and cancer mortality rates is not linear in this setting. Therefore, it is necessary to consider methodologies that include at least one nonparametric component to account for nonlinearity. On the other hand, fully nonparametric approaches have the disadvantage that they do not yield parameter estimates or summary measures such as annual average percentage change. Instead, the description is primarily graphical.

The semiparametric methodology used in this study allows the use of different positive bias distributions to obtain improved fits. For breast cancer, a contaminated normal distribution was used, for cervical and stomach cancers, a normal distribution was used, and for lung cancer, an exponential power distribution was used. Therefore, semiparametric models provide greater flexibility for fitting models where the response is a rate, as in cancer mortality. This is because they allow nonlinear effects and accommodate nonconstant variance, particularly when compared with the traditional generalized linear model with a Poisson response.

[Received: June 2025 — Accepted: December 2025]

References

Berk, R. & MacDonald, J. M. (2008), 'Overdispersion and Poisson regression', *Journal of Quantitative Criminology* **24**(3), 269–284.

Cabanes, A., Vidal, E., Pérez-Gómez, B., Aragonés, N., López-Abente, G. & Pollán, M. (2009), 'Age-specific breast, uterine and ovarian cancer mortality trends in Spain: Changes from 1980 to 2006', *Cancer Epidemiology* **33**(3–4), 169–175.

Clèries, R., Martínez, J. M., Escribà, J. M., Esteban, L., Pareja, L., Borrás, J. M. & Ribes, J. (2010), 'Monitoring the decreasing trend of testicular cancer mortality in Spain during 2005–2019 through a Bayesian approach', *Cancer Epidemiology* **34**(3), 244–256.

DANE (2017), 'Demografía y población', Sitio web. https://www.dane.gov.co/index.php/estadisticas-por-tema/demografia-y-poblacion

Guo, P. & Li, K. (2012), 'Trends in esophageal cancer mortality in China during 1987–2009: Age, period and birth cohort analyses', *Cancer Epidemiology* **36**(2), 99–105.

Instituto Nacional de Cancerología, Ministerio de Salud y Protección Social & Departamento Administrativo de Ciencia, Tecnología e Innovación en Salud (COLCIENCIAS) (2013), Guía de práctica clínica (GPC) para la detección temprana, tratamiento integral, seguimiento y rehabilitación del cáncer de mama, Technical report, Sistema General de Seguridad Social en Salud, Colombia. Guía completa. Guía No. GPC-2013-19.

Kvåle, R., Møller, B., Angelsen, A., Dahl, O., Fosså, S. D., Halvorsen, O. J., Hoem, L., Solberg, A., Wahlqvist, R. & Bray, F. (2010), 'Regional trends in prostate cancer incidence, treatment with curative intent and mortality in Norway 1980–2007', *Cancer Epidemiology* **34**(4), 359–367.

Marchant, D. J. (1982), 'Epidemiology of breast cancer', *Clinical Obstetrics and Gynecology* **25**(2), 387–392.

National Cancer Institute (2015), 'Risk factors for cancer', Sitio web. https://www.cancer.gov/about-cancer/causes-prevention/risk

Ohshima, H. & Bartsch, H. (1994), 'Chronic infections and inflammatory processes as cancer risk factors: Possible role of nitric oxide in carcinogenesis', *Mutation Research* **305**(2), 253–264.

Pike, M. C., Spicer, D. V., Dahmoush, L. & Press, M. F. (1993), 'Estrogens, progestogens, normal breast cell proliferation, and breast cancer risk', *Epidemiologic Reviews* **15**(1), 17–35.

Piñeros, M., Pardo, C., Gamboa, O. & Hernández, G. (2010), *Atlas de mortalidad por cáncer en Colombia*, 3 edn, Imprenta Nacional de Colombia, Bogotá. Instituto Nacional de Cancerología; IGAC.

R Core Team (2016), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Vanegas, L. H. & Paula, G. A. (2015), 'A semiparametric approach for joint modeling of median and skewness', *TEST* **24**(1), 110–135.

Vanegas, L. H. & Paula, G. A. (2016), *ssym: Fitting semi-parametric log-symmetric regression models.* R package version 1.5.7. https://CRAN.R-project.org/package=ssym

World Health Organization (2016), 'International classification of diseases', Sitio web. https://www.who.int/classifications/icd/en/

World Health Organization (2017), 'Cáncer', Sitio web. https://www.who.int/news-room/fact-sheets/detail/cancer