Reconstruction of smoking prevalence in Spain by sex and age groups in the period 1991-2020

Guerra-Tort C[1], *López-Vizcaíno E* [2], *Santiago-Pérez MI*[3], *Rey-Brandariz J*[1], *Varela L*[1], *Candal C*[1], *Ruano-Ravina A*[1], *Pérez-Ríos M*[1]

[1]carla.guerra@rai.usc.es, Departamento de Medicina Preventiva e Saúde Pública, Universidade de Santiago de Compostela

[2]esther.lopez@ige.eu, Servizo de Difusión e Información, Instituto Galego de Estatística, Xunta de Galicia

[3] Servizo de Epidemioloxía, Dirección Xeral de Saúde Pública, Xunta de Galicia

In Spain, there is incomplete available information to make a global assessment of the evolution of the smoking epidemic. In this work we propose to accurately reconstruct the annual series of smoking prevalence in our country in the period 1991-2020 applying a small-area model.

The data used are derived from public statistics, with special relevance to those derived from the different National (1993, 1995, 1997, 2001, 2003, 2006, 2011, 2017) and European (2009, 2014, 2020) health surveys. A small-area estimation method based on aggregated data was used to reconstruct smoking prevalence by sex and five-year age group (from 15-19 years to 80-84 years) for each year of the period 1991-2020. Specifically, a small-area multinomial logistic mixed model with area and time random effects was applied. The areas were the *D*=30 groups defined from crossing the 15 five-year age groups with sex, and the time periods were the *T*=30 years of the 1991-2020 series. In the model, the response variable is a vector with the number of smokers, ex-smokers and never smokers in each area and time, and the covariates are aggregated information related to tobacco consumption obtained from registries (sociodemographic, economic and morbidity data). The model is expressed as:

$$p_{dkt} = \frac{exp\ (\eta_{dkt})}{1 + exp(\eta_{d1t}) + exp\ (\eta_{d2t})},$$

$$\eta_{dkt} = \log\left(\frac{p_{dkt}}{p_{d3t}}\right) = x_{dkt}\beta_k + u_{1,dk} + u_{2,dkt},\ \ d = 1, \dots, D, k = 1,2, t = 1, \dots, T,$$

where p_{dkt} is the prevalence of each category *k* corresponding to area *d* and time *t*, $x_{dkt} = (x_{dkt1}, \dots, x_{dktr_k})'$ is the set of covariates corresponding to category *k*, area *d* and time *t*, and $\beta_k = (\beta_{k1}, \dots, \beta_{kr_k})'$ is the vector of regression parameters. The subscript *k* refers to the category of smokers (*k*=1) or ex-smokers (*k*=2). The third category of never smokers (*k*=3) is taken as the reference category. The model also considers random effects $u_{1,dk}$ and $u_{2,dkt}$ associated with area *d* and category *k*, and area *d* and category *k* and time *t*, respectively. To fit the model, we combine the penalized quasi-likelihood method, for the estimation and prediction of β_{kr_k}, $u_{1,dk}$ and $u_{2,dkt}$, with the restricted maximum likelihood method, which is used to estimate variance components.

To estimate the prevalence of tobacco consumption the following steps were carried out:1) union of the 11 survey's data and computation of the individual smoking status of the people polled from current status, age at initiation and age at cessation for each year of the period 1991-2020; 2) calculation of the prevalence of smokers, ex-smokers and never smokers by sex, age group and year applying a weighted ratio estimator; 2) preparation of covariates by sex, age group and year; 3) selection of covariates to fit the model for smokers and ex-smokers, and 4) adjustment of the small-area model.

We expect that the results of this work and its methodology can be applied to other lifestyle and health determinants such as alcohol consumption or exposure to second-hand smoke, allowing us to quantify their detailed impact on the health of populations.

Keywords: small-area analysis, smoking, health surveys.