Geographical differences in blood potassium detected using a structured additive distributional regression model

Jenifer Espasandín-Domínguez [a],*, Alfonso Javier Benítez-Estévez [b], Carmen Cadarso-Suárez [a], Thomas Kneib [c], Tegra Barreiro-Martínez [b], Balbina Casas-Méndez [d], Francisco Gude [e]

[a] Group of Biostatistics, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS). University of Santiago de Compostela, Spain.

[b] Department of Laboratory Medicine. Hospital Clínico Universitario de Santiago de Compostela, Spain.

[c] Chair of Statistics. Georg-August-University Göttingen, Germany.

[d] Department of Statistics, Mathematical Analysis, and Optimization. University of Santiago de Compostela, Spain.

[e] Clinical Epidemiology Unit. Hospital Clínico Universitario de Santiago de Compostela, Spain.

* Correspondence to: Group of Biostatistics, Center for Research in Molecular Medicine and Chronic Diseases (CiMUS), Campus Vida, 15782, Santiago de Compostela, A Coruña, Spain. E-mail address: jenifer.espasandin@usc.es (J. Espasandín-Domínguez).

Keywords: Potassium; Distributional regression; Spatial análisis; P-splines.

ABSTRACT: Recently, physicians in an area of northwestern Spain became concerned about the large number of patients whose serum potassium concentrations were above the normal range, as well as differences in the values recorded from one area to another. With the aim of identifying geographical differences in both mean and variability of potassium levels, analyses were performed using modern flexible regression techniques based on a structured additive distributional regression model. In this type of model, every parameter of a response distribution – rather than just the mean – is related to a structured additive predictor. After adjusting for variables such as age, sex, clot-contact time and spatial effects, differences in potassium concentrations were confirmed. The type of distributional regression model used permitted the mean and variance of the potassium concentrations to be modelled using additive predictors that allow for different types of covariate effects. A variety of complex distributions were contemplated. In general, higher concentrations and greater variability were recorded in areas further from the hospital laboratory responsible for the analyses, although some that were nearby also returned high values. Further actions are required to confirm whether these differences reflect reality or non-optimum pre-analytical handling (resulting in pseudohyperkalaemia) in certain districts.

1. Introduction

Recently, general practitioners working in the Santiago de Compostela Health Area (SCHA) in northwestern Spain raised concerns over the high percentage of patients whose serum potassium concentrations were above the normal range, and over differences in the values recorded from one area to another. Analytical laboratories are commonly called upon to determine serum potassium concentrations, especially for patients with diabetes, heart and kidney disease. When potassium concentrations are recorded falsely as high (pseudohyperkalaemia) owing to specimen-collection or processing errors, medical mistakes can be made with disastrous consequences for patients. Although the list of sample management factors that can modify the potassium concentration is large, problems can be prevented by good laboratory practice (Stankovic and Smith, 2004).

The SCHA covers an area of approximately 4905 km^2, and at the time of the present study had a population of 497 171 (Instituto Galego de Estatística, 2015). The Clinical and Laboratory Standards Institute (2008) recommends that procedures be established for the transport of samples to laboratories to ensure that they are protected from deterioration (Tanner et al., 2008). The 46 general practices that use the laboratory service based at the Hospital Clínico Universitario de Santiago (CHUS) are, however, up to 70 km away, and timely transport of samples to the laboratory is a challenge. The aim of the present work was to determine whether any geographical differences exist in terms of recorded serum potassium concentrations and their variability that might be attributed to preanalytical factors, such as the centre where blood was extracted, adjusting for other potential covariables that might influence the results. For this, a structured additive distributional regression model (Klein et al., 2015) was used. This type of model is closely related to generalized additive models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos, 2005). Inference in GAMLSS models is based on penalized maximum likelihood estimation, achieved via backfitting loops over the additive predictor components. The current work present a Bayesian version of a structured additive distributional regression model dependent on Markov chain Monte Carlo (MCMC) simulation algorithms. This approach has the advantage of providing credible intervals without relying on asymptotic arguments (Klein et al., 2015).

An advantage of this kind of model is the possibility of incorporating spatial effects. However, in most cases the output of spatial effects is not directly interpretable by biomedical researchers. This paper proposes a way in which spatial effects can be visualized.

Another advantage of this type of model is the possibility of contemplating a wide range of response variables. The deviance information criterion (DIC) (Spiegelhalter et al., 2002; Klein et al., 2015) is commonly used for model choice in distributional regression. Quantile residuals can be used to check the performance of a selected model (Klein et al., 2015). In practice, the residuals can be assessed graphically in terms of quantile–quantile plots. However, interpreting the resulting graphs can be difficult, and the decision on the adequacy of a model remains subjective. Sometimes, even though the model is correct, the plot may deviate substantially from a straight line (Augustin et al., 2012). We therefore here propose the use of quantile–quantile plots with reference bands. To construct these bands, the methodology of Augustin et al. (2012) was adapted to the context of distributional regression.

In Section 2 we present the description of the database used in the study. Section 3 provides an introduction to structured additive distributional regression models, including the choice of the response distribution (and thus the model selected) and the construction of quantile–quantile plots

with reference bands. Section 4 provides the results obtained following analysis of the potassium database discussed in Section 2. Finally, Section 5 provides a summary and some comments on directions of future research.

2. Data description

The database used in this work was provided by the Clinical Analysis Laboratory of the CHUS. This supplied information on all blood extractions performed between 1 June and 31 December 2015 for which serum potassium, sodium and creatinine measurements were made. The initial number of samples was 145 960, collected at 46 extraction centres within the SCHA. Those samples showing signs of haemolysis were excluded, as were those with creatinine or sodium concentrations outside the normal range (indicators of impaired kidney function). The final number of samples used in the present analysis was therefore 95 096.

2.1. Health Area of Santiago de Compostela (SCHA)

The SCHA, in Spain's northwest, covers 46 municipal districts (see Table 2 and Fig. 6 in the Appendix for a map and the distribution of the population). The CHUS reference hospital is located in the city of Santiago de Compostela, from where the SCHA's health centres and doctors' practices are coordinated (see Fig. 6 in the Appendix for their locations).

Blood samples were taken, usually daily, at designated locations in these 46 areas and transported by road to the CHUS following different routes.

The following variables were considered to be covariates: *gender*, *age* (in years), clot-contact time in minutes (*cctime* in formulae (1)), and demographic information on the district where the extraction centres were located (s).

56% of the patients who provided blood samples were female, and 44% were male. The age range of the patients was 1–103 years. The mean (respectively SD) age was 54.8 (19.0) years; the median was 61 years.

The clot-contact time was taken as the difference between the starting time of sample collection at the extraction centre and the entry time of the sample in the laboratory registry. The range was 4–458 min, the mean (SD) clot-contact time 229 (49) min, and the median time 231 min.

3. Structured additive distributional regression models

Distributional regression models (Klein et al., 2014) are innovative models that permit marginal distribution parameters to be modelled using additive predictors that allow for several types of covariate effects (such as the non-linear effects of continuous covariates, random effects, and the interactions or spatial effects). By modelling each parameter of the response at the same time – and not just the mean – they provide additional flexibility. This type of model also provides different types of (possibly non-standard) response distributions for continuous, discrete, and mixed discrete continuous distributions. However, the adequate selection of response variable in the formulation of these models is not without its difficulty. The statistical literature usually suggests one suppose the response variable to follow a normal distribution. However, in the present work, several distribution types were contemplated for the modelling of the potassium concentrations, including the log-normal distribution, the truncated normal distribution, the inverse Gaussian distribution, and the gamma distribution (see Section 3.3 and Appendix A.2).

In this work, a structured additive distributional regression model was used to study the potassium concentrations recorded. This type of distributional regression allows the effect of the covariate information on all the parameters of the response distribution to be examined.

Let us assume that observations $(y_i, \boldsymbol{v_i}, i = 1, \ldots, n)$ are made, where y_i are observations on the response variable, and $\boldsymbol{v_i}$ represents the generic covariate vector. In this scenario, the response variables y_i can be assumed independently distributed with K-parametric densities $p(y_i \mid \vartheta_{i1}, \ldots, \vartheta_{iK}) \equiv p_i$. In other words, the conditional distribution p_i of an observation y_i given $\boldsymbol{v_i}$ is expressed in terms of the K distributional parameters of the response distribution: $\vartheta_{i1}, \ldots, \vartheta_{iK}$.

In structured additive distributional regression models, each parameter ϑ_{ik}, $k = 1, \ldots, K$, of the response distribution is related to a semiparametric additive predictor $\eta_i^{\vartheta_k}$ defined in terms of the covariate vector $\boldsymbol{v_i}$. As in other types of classic regression model, such as generalized linear regression models, a suitable response function is used to map the predictor to the parameter of interest, $\vartheta_{ik} = h^{\vartheta_k}(\eta_i^{\vartheta_k})$. Following Klein et al. (2015), in this expression, the superscript ϑ_k refers to the fact that K predictors specific, for each of the distribution parameters of the response variable (and not just for the mean as in classical regression), are taken into account. Moreover, for an observation $i = 1, \ldots, n$, a suitable structured additive predictor for parameter ϑ_k can be written as:

$$\eta_i^{\vartheta_k} = \beta_0^{\vartheta_k} + f_1^{\vartheta_k}(\boldsymbol{v_i}) + \cdots + f_{J_k}^{\vartheta_k}(\boldsymbol{v_i}) + f_{spat}^{\vartheta_k}(s_i)$$

where $\beta_0^{\vartheta_k}$ represents the overall level of the predictor, and the functions $f_j^{\vartheta_k}(\boldsymbol{v_i})$, $j = 1, \ldots, J_k$ represent the different covariate effects. Note that each distribution parameter may depend on different covariates and a different number of effects, say J_k. The generic representation with the complete covariate vector can be used to simplify the notation. Finally, $f_{spat}(s)$ is the spatial effect capturing heterogeneity at the level of the districts s.

In structured additive regression, each function f_j is approximated by a linear combination of D_j appropriate basis functions:

$$f_j(\boldsymbol{v_i}) = \sum_{d_j=1}^{D_j} \beta_{j,d_j} \boldsymbol{B}_{j,d_j}(\boldsymbol{v_i}).$$

In matrix notation, we can write $\boldsymbol{f_j} = \big(f_j(\boldsymbol{v_1}), \ldots, f_j(\boldsymbol{v_n})\big)' = \boldsymbol{Z_j}\boldsymbol{\beta_j}$ where $\boldsymbol{Z_j}[i, d_j] = \boldsymbol{B}_{j,d_j}(\boldsymbol{v_i})$ is an $n \times D_j$ design matrix and $\boldsymbol{\beta_j}$ is the vector of coefficients (with dimension D_j) to be estimated. The basis function representation then lead us to the following matrix representation of the predictor 3:

$$\boldsymbol{\eta} = \beta_0 \mathbf{1} + \boldsymbol{Z_1}\boldsymbol{\beta_1} + \cdots + \boldsymbol{Z_J}\boldsymbol{\beta_J}.$$

For each of the parameter vectors $\boldsymbol{\beta_j}$, the multivariate normal prior can be assumed:

$$p(\boldsymbol{\beta_j} \mid \tau_j^2) \propto \left(\frac{1}{\tau_j^2}\right)^{\frac{rk(\boldsymbol{K_j})}{2}} \exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta_j'}\boldsymbol{K_j}\boldsymbol{\beta_j}\right)$$

in which the (potentially rank-deficient) precision matrix $\boldsymbol{K_j}$ corresponds to the penalty matrix in a frequentist formulation. Note that here, we are using a generic notation for different terms (e.g penalized splines, Markov random fields, random effects). The precision matrix, $\boldsymbol{K_j}$, for the Markov random field is rank-deficient by construction since only deviations from a constant spatial effect are penalized. This leads to a rank deficiency of one. In the case of the penalized splines, the rank deficiency comes from the fact that polynomials of degree equal to the difference order minus one are not penalized. However for other possible terms like random effects this matrix is not rank-deficient.

A prior smoothing variance of τ_j^2 is assigned as an inverse gamma hyperprior $\tau_j^2 \sim IG(a_j, b_j)$ (with $a_j = b_j = 0.001$ as a default option) in order to obtain data-driven smoothness.

Again following Klein et al. (2015) and to simplify the notation, the dependence on the distributional parameter indicated by the superscript ϑ_k, the observation index i, and the function index j, have been dropped.

Fahrmeir et al. (2013) discuss all terms included in this generic predictor. The following paragraphs outline the prior assumptions for the hierarchical predictor required for this study. See Lang and Brezger (2004) for more details.

3.1. Linear effects and continuous covariates

For the effect of the intercept, β_0, and the gender of the individuals, β_1, a flat, non-informative prior was assumed.

The non-linear effects of continuous covariates (*age* and *clot-contact time*) were modelled using Bayesian versions of penalized splines (P-splines, Lang and Brezger, 2004), introduced into a frequentist setting by Eilers and Marx (1996). To model age and clot-contact time, 20 inner knots, a cubic spline basis, and a second order random walk prior for penalized splines were contemplated. For the penalized spline specifications, we were able to rely on extensive research concerning the number and placement of knots, the order of the random walk prior, and the degree of the polynomial spline, e.g. Eilers and Marx (1996), Lang and Brezger (2004) and Brezger and Lang (2006). Their main findings can be summarized as follows: (i) The number and placement of the knots has only a very minor impact on the fit if the number of knots chosen is not too small. (ii) 20 equidistant knots yield sufficient flexibility for basically all situations of applied interest. (iii) Second order random walk priors leave a linear effect unpenalized which is in analogy to the common penalty for smoothing splines. Moreover, first order differences often yield more wiggly estimates. (iv) Finally, cubic splines yield a visually smooth function estimate which is twice continuously differentiable. This fits very well with the common visual perception of non-linear effects.

3.2. Spatial effects

The spatial effects, f_{spat}, were understood as the sum of the spatially structured correlated (smooth) effects f^{str}, and spatially uncorrelated (unsmooth) effects:

$$f_{spat}(s) = f^{str}(s) + f^{unstr}(s).$$

A spatial effect is usually a surrogate of many unobserved influential factors, some of which may obey a strong spatial structure while others may be present only locally. By estimating a structured and an unstructured component, it was hoped that distinctions could be made between these kinds of influential factor (Besag et al., 1991).

For correlated spatial effects (or structured spatial effects), we assume spatial correlations defined implicitly by assuming a Markov random field (Fahrmeir et al., 2013) as a prior distribution for the separate regression coefficients corresponding to the distinct regions. The Markovian structure is determined by the neighbourhood structure for the regions and the precise form of the prior distribution is defined by:

$$\beta_{str,s}|\beta_{str,r},\, r \neq s,\, \tau^2_{str} \sim N\left(\frac{1}{N_s}\sum_{r\in\delta_s}\beta_{str,r},\, \frac{\tau^2_{str}}{N_s}\right),$$

where $N_s = |\delta_s|$ is the number of adjacent sites or neighbours, and $r \in \delta_s$ denotes that region r is a neighbour of site s. The conditional mean of $\beta_{str,s}$, given all other coefficients, is the average of the neighbouring regions.

Additional uncorrelated random effects (or unstructured spatial effects) may be incorporated as a surrogate for unobserved local small-area, group or individual specific heterogeneity. If one ignores spatial proximity and interprets $s \in \{1, \ldots, S\}$ as a cluster variable that only represents membership to different groups (such as different individuals in longitudinal data or more generally to different clusters of observations), we can assume a standard Gaussian random effects prior, i.e. $f^{unstr}(s) \sim N(0, \tau^2_{unstr}), s \in \{1, \ldots, S\}$, where the different groups correspond to the different administrative regions in the data set.

Note that here two variances are used for the structured, τ^2_{str}, and the unstructured effect, τ^2_{unstr}. Structurally, both variances are of the same type but they refer to different prior assumptions, where one assumes spatial structure (τ^2_{str}) while the other one assumes spatial independence (τ^2_{unstr}).

3.3. Inference and choice of the response distribution

Distributional regression models can be inferred employing computationally efficient extensions of the MCMC techniques developed by Klein et al. (2014).

Many approaches have been proposed for dealing with model choice in a Bayesian framework. In the present work, the Deviance Information Criterion (DIC, Spiegelhalter et al., 2002; Klein et al., 2015)

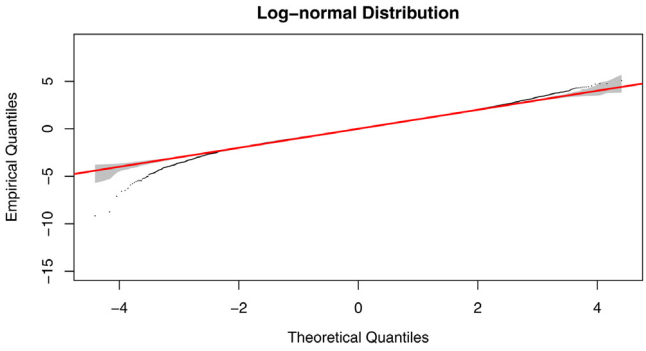


Fig. 1. Quantile–quantile residuals plot for the selected model with reference bands: the closer the residuals to the bisecting red line, the better the fit to the data. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

was used to choose the best response distribution. The DIC is a commonly used criterion for model choice in Bayesian inference. It became popular in part because of its easy implementation from the MCMC output. The performance of the DIC was valued as positive by Klein et al. (2015), who compared several mis-specified models with the true DIC model. In the present work, the DIC showed the log-normal distribution to provide the best and most parsimonious fit (see Appendix A.2 for more details).

The residuals can also be used to check the performance of a selected model (Klein et al., 2015). If the estimated model is close to the true model, the quantile residuals approximately follow a standard normal distribution, even if the model distribution itself is not a normal distribution. In practice, the residuals can be assessed graphically in terms of quantile–quantile plots. To improve the interpretability of this type of graphics, we propose plotting reference bands around the diagonal line to give a rough indication of the uncertainty implied by estimating the model from finite data. More precisely, following Augustin et al. (2012), we repeatedly simulate data from the fitted model and add pointwise minima and maxima for the resulting quantile residuals from a pre-specified number of replications. Fig. 1 shows quantile–quantile plots for the selected model with a log-normal distribution. Fig. 1 also provides reference bands for judging the relevance of departures of quantile–quantile plots from the ideal red line. The log-normal distribution turns out to be appropriate for residuals in the range between -2.5 and 2.7 but deviates from the diagonal line for extreme values. Note, however, that these extreme values correspond to only 1.37% of the total data (0.87% of the database to the left and 0.5% to the right) such that our model explains the vast majority of observations well.

Thus, the structured additive distributional regression model described at the beginning of Section 3 was used with a log-normal distribution response and with two covariate-dependent parameters (corresponding to the mean of the log-transformed potassium concentrations and the scale parameter σ^2), to describe and compare the potassium concentrations recorded in the different districts of the SCHA. This model is expressed as follows:

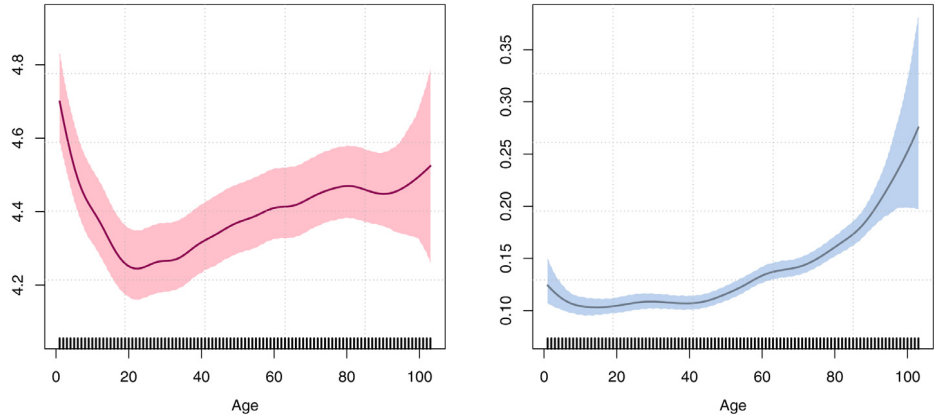
$$\begin{cases} \eta^{\mu} = \beta_0^{\mu} + gender'\beta_1^{\mu} + f_1^{\mu}(age) + f_2^{\mu}(cctime) + f_{spat}^{\mu}(s), \\ \eta^{\sigma^2} = \beta_0^{\sigma^2} + gender'\beta_1^{\sigma^2} + f_1^{\sigma^2}(age) + f_2^{\sigma^2}(cctime) + f_{spat}^{\sigma^2}(s) \end{cases} \tag{1}$$

where β_0 represents the overall level of the predictor and β_1 captures the effect of the gender. Moreover, $f_1(age)$ and $f_2(cctime)$ are non linear effects of the age and the clot-contact time, respectively. Finally, $f_{spat}(s)$ is the spatial effect of the districts s.

4. Data analysis

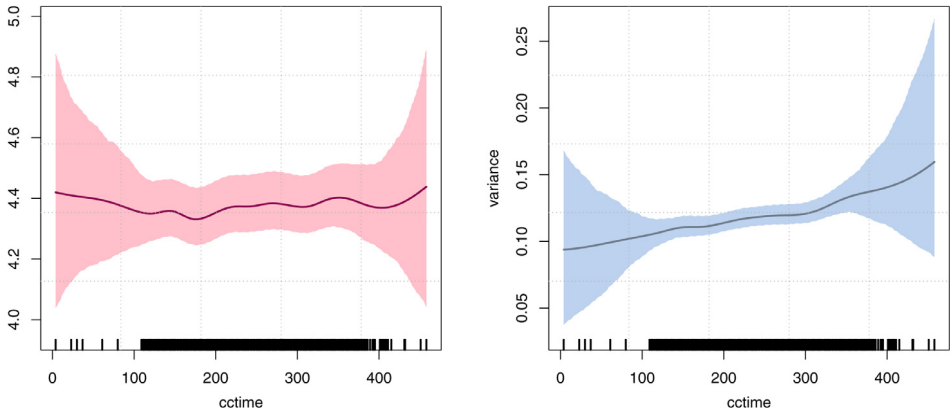
Statistical analyses were performed using open-source BayesX software (Belitz et al., 2015). The BayesX (Kneib et al., 2015) and R2Bayesx (Umlauf et al., 2015; Belitz et al., 2016) R packages were used as graphic interfaces.

Parameter	mean	2.5% quantile	median	97.5% quantile
β_0^{μ} (intercept)	1.489	1.465	1.484	1.512
$\beta_0^{\sigma^2}$ (intercept)	−0.032	−0.050	−0.031	−0.014
β_1^{μ} (gender)	−0.006	−0.008	−0.006	−0.005
$\beta_1^{\sigma^2}$ (gender)	−0.032	−0.050	−0.032	−0.014



(a) $f_1^{\mu}(age)$. (b) $f_1^{\sigma^2}(age)$.

Fig. 2. Posterior mean estimates of non linear effects of *age* on μ and σ^2.



(a) $f_2^{\mu}(cctime)$. (b) $f_2^{\sigma^2}(cctime)$.

Fig. 3. Posterior mean estimates of non linear effects of *cctime* on μ and σ^2.

All results are summarized in Table 1 and Figs. 2–5. For the continuous variables, all covariates but the one that is visualized are fixed at their average while spatial effects are set to zero. For the spatial effects, we used district-specific averages for all covariates and determined significance based on the comparison with the average of all spatial effects.

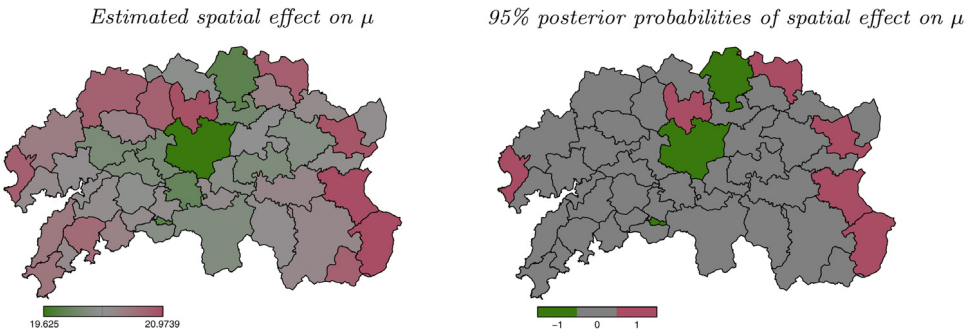


Fig. 4. Posterior mean estimates of the complete spatial effects on mean potassium levels, f^{μ}_{spat}, and 95% posterior probabilities (right). In the right panel a value of 1 corresponds to a strictly positive 95% credible interval, and a value of −1 to a strictly negative credible interval; a value of 0 indicates that 0 is contained in the corresponding credible interval.

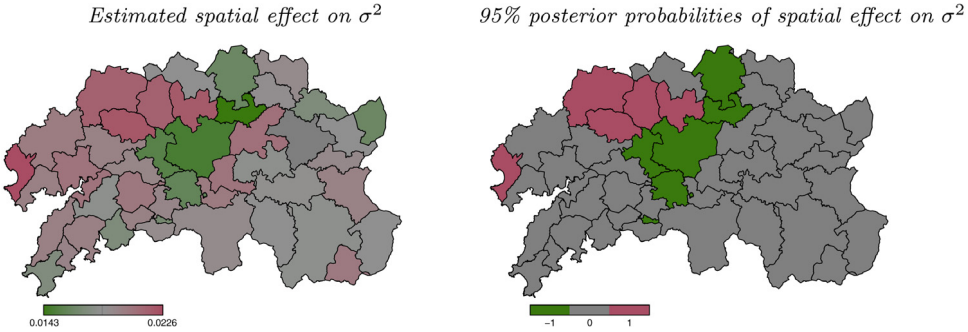


Fig. 5. Posterior mean estimates of the complete spatial effects on the variance of potassium levels, $f^{\sigma^2}_{spat}$, and 95% posterior probabilities (right). In the right panel a value of 1 corresponds to a strictly positive 95% credible interval, and a value of −1 to a strictly negative credible interval; a value of 0 indicates that 0 is contained in the corresponding credible interval.

Gender had some influence on the results (both on the mean and variance of the potassium concentrations), but this effect is clinically not relevant since the differences between both men and women were minimal (0.04 mg/dL). The same for the variance. However, age (Fig. 2), clot-contact time (Fig. 3) and the place of origin of the samples did influence the potassium concentrations recorded (Figs. 4 and 5). Children and the elderly had higher mean potassium concentrations than did patients of intermediate age; the values recorded for the elderly also showed greater variability (Fig. 2). These age-related findings might, however, be expected since venipunture is harder to perform in both children and the elderly, and both age groups show greater capillary fragility. This can lead to situations in which haemolysis occurs, releasing potassium from the red blood cells and increasing the recorded concentration for both age groups, as well as the variability of values recorded for the elderly.

The potassium concentrations recorded were clearly not uniformly distributed over the study area. In general, higher potassium concentrations were recorded in the areas farthest away from the test laboratory, although some areas close to it also returned high values. In some areas, these high concentrations were accompanied by greater variability in the results, particularly in the districts on the northern periphery of the study area. The districts to the southeast also returned high potassium concentrations but with the less variability. Although caution should be exercised when interpreting these spatial analysis results, it may be that potassium concentrations in the periphery are related to pre-analytical factors associated with the extraction centres. The affected areas are also those with the lowest population densities; they are therefore likely to have less equipment, fewer personnel,

and perhaps less well trained personnel than in the more central districts. These periphery districts may also have older inhabitants, clinical practices may be less homogeneous, and they are the worst communicated with the test laboratory (Figs. 4 and 5).

5. Discussion

Distributional regression models extends the use of generalized additive models (GAM, Hastie and Tibshirani, 1990) to situations in which the response distributions are non-standard, and in which not only the mean but multiple parameters are related to additive predictors via suitable link functions. Further, they allow additional flexibility by specifying structured additive predictors for each parameter of interest, and thus adjust for flexible non-linear effects of continuous covariates for which the smoothness is determined based on the data. They also allow the contemplation of spatial effects to capture unobserved spatial heterogeneity and spatial correlations, interaction terms such as varying coefficients or interaction surfaces, and cluster-specific random effects (Fahrmeir et al., 2013; Brezger and Lang, 2006).

The use of a new structured additive distributional regression model allowed for the flexible modelling of the distribution of potassium concentrations with a potentially non-standard response type, and permitted covariate effects to be taken into account, including the smooth estimation of the effect of continuous variables, categorical covariates, random effects and possible spatial trends. Its use clearly identified differences in serum potassium concentrations among extraction sites after adjusting for other potentially influential factors, such as age, gender and clot-contact time. The spatial analysis revealed some districts to return higher mean serum potassium concentrations, and to show greater variability in terms of these results. Two geographically-related clusters were detected: (1) districts on the periphery of the study area that returned higher potassium concentrations (and showed greater variability in the results) than those in the central area, and (2) a number of districts that returned higher potassium concentrations independent of their location.

Although classic regression analyses allow for the easy interpretation of results, they only focus on means, and may lead to erroneous conclusions when modelling complex data structures. The distributional regression models used in this manuscript provide a generic framework for performing regression analyses in which several parameters of a potentially non-standard response distribution are related to flexible regression predictors. This work shows how to visualize from a statistical viewpoint the results of distributional regression models in an analysis comprising spatial information. It is not sufficient to show the estimated spatial effects directly; rather the spatial effect has to be adjusted with respect to the covariates observed in the particular regions. We therefore plugged in covariate values obtained as spatially stratified averages for all other covariates and then compared against the overall mean of the spatial effect to determine significances.

Another benefit of this type of model is the possibility of being able to contemplate a wide family of response variables. One way of examining the goodness of fit of the selected model is via quantile–quantile plots. However, conclusions drawn from such plots can be subjective. This manuscript proposes the use of quantile–quantile plots with reference bands.

The present work examined different (non-standard) distributions that depended on the mean and variance of responses. Using the DIC, distributional regression with a log-normal response was used. This provided not only improved goodness of fit over classic distributions (e.g., a Gaussian distribution), but led to different results being obtained. Thus, although the non-linear effects of the covariables on the expectation were rather similar with both distributions, the log-normal distribution allowed differences to be identified in the variability of the spatial effects associated with the potassium concentrations that were undetectable when a Gaussian distribution was contemplated (data not shown). The majority of the central municipalities had larger populations, more health care personnel and more equipment, and followed protocols more strictly, which might explain the lower potassium concentrations they recorded (with both distributions) and their smaller variability (log-normal distribution). In medical organizations, examining clinical variation in medical practice is an important step to measuring efficiency and effectiveness in care delivery.

From a statistical viewpoint, another important feature of this type of model, is the possibility of modelling the effects of the continuous covariables and spatial effects in a flexible, unified manner

(as shown in the present work) as well as allowing for complex interactions between different types of variables, e.g., factor-curve or surface interactions. They also allow for the modelling of spatio-temporal trends. One of the hypotheses of the present work was that the holiday periods of the extraction personnel might have an effect on a number of preanalytical factors (e.g., a greater chance of heamolysis occurring when less experienced personnel perform extractions). The present work does not contemplate such effects since data were available only for a short period (6 months). Future work will include extending the observation period to 5 years, allowing these spatio-temporal effects to be taken into account.

Further work is required to determine whether the elevated potassium concentrations detected reflect a real clinical panorama or a problem of pseudohyperkalaemia. The latter scenario would appear to be more likely, however, since neither the lifestyles of those living in high value districts, nor the prevalence of disease in these areas, would seem able (at least on first inspection) to explain them. If the high values do reflect a pseudohyperkalaemia problem, a number of actions could be undertaken to help rectifying it, including: (1) the education of laboratory and non-laboratory personnel about the causes of increased potassium readings; (2) the teaching of procedures to reduce the problem; (3) improving the transport routes to the hospital to reduce clot-contact times; and (4) constant monitoring of potassium concentrations and the apparent rate of hyperkalaemia. It is worth noting that the efficient management of laboratories and other health care services has received considerable attention in the optimization literature (e.g. Green, 2006 or Mankowska et al., 2014). An investigation into the optimization techniques most appropriate for the present context might help curb the possible inefficiencies in the sample routing system.

Given the close relationship between potassium and sodium ions, it would be of great interest to study both cations at the same time in order to determine the covariables that influence them and their interactions. For this, flexible copula distributional regression models for multivariate responses can be used both in a Bayesian framework (structured additive conditional copula regression models, Klein and Kneib, 2016) or in a frequentist framework, using bivariate copula additive models for location, scale and shape (Marra and Radice, 2017).

Acknowledgements

The authors would like to acknowledge the referees for their valuable comments that improved a lot the manuscript.

The work was supported by grants from the Carlos III Health Institute, Spain (RD12/0005/0007, PI11/02219), a pre-doctoral grant (ED481A-2015/113) from the Galician Government (Plan I2C)-Xunta de Galicia and the German Research Foundation (DFG) grant KN 922/4-2. This study also was supported by the project MTM2014-52975-C2-1-R cofinanced by the Ministry of Economy and Competitiveness (SPAIN) and the European Regional Development Fund (FEDER).

Appendix

Supplementary material for "Geographical differences in blood potassium detected using a structured additive distributional regression model".

A.1. Description of the SCHA

Fig. 6 presents the locations of the 46 SCHA's municipal districts. Table 2 provides the names of the 46 districts of SCHA along with their population and population density. These values vary considerably; for example, the district of Santiago de Compostela, which has an area of 220 km^2, is home to 95 612 people, while that of Toques, with an area of just 77.9 km^2, has 1213 inhabitants.

Table 2 also shows the percentage of potassium readings that fell outside the normal range recognized by the test laboratory. The range of values across the different districts is very wide, and not easily explained by the ageing of the population or differences in the prevalence of chronic diseases.

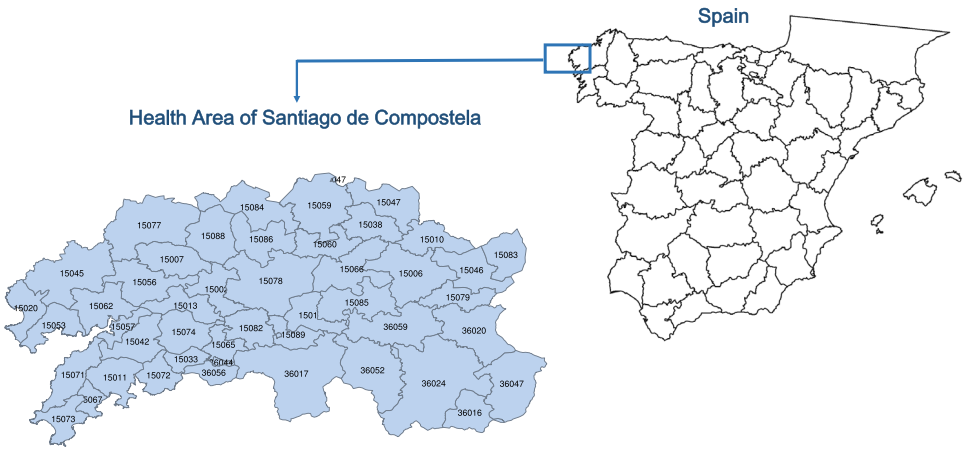


Fig. 6. Health Area of Santiago de Compostela. Codes are in Table 2.

Table 2
Districts in the Santiago de Compostela Health Area (SCHA), their codes and demographic characteristics, and the percentage of patients whose potassium results fell outside the normal range. The SCHA occupies an area of some 4095 km^2; its population was 497 171 at the time of the study.
Source: www.ige.es.

District	Population	Area (km^2)	Individuals out of range (%)	District	Population	Area (km^2)	Individuals out of range (%)
15002 - Ames	30 267	80.0	18.8	15071 - Porto do Son	9 436	94.6	24.3
15006 - Arzúa	6 219	155.5	16.1	15072 - Rianxo	11 386	58.8	21.2
15007 - A Baña	3 698	82.3	22.4	15073 - Ribeira	27 372	68.8	26.0
15010 - Boimorto	2 125	86.6	17.8	15074 - Rois	4 710	92.8	21.1
15011 - Boiro	18 950	86.6	26.9	15077 - Santa Comba	9 635	203.7	30.3
15012 - Boqueixón	4 321	73.2	20.3	15078 - Santiago de Compostela	95 612	220.0	14.9
15013 - Brión	7 564	74.9	19.0	15079 - Santiso	1 709	67.4	19.0
15020 - Carnota	4 284	70.9	28.0	15082 - Teo	18 505	79.3	12.6
15033 - Dodro	2 882	36.1	20.8	15083 - Toques	1 213	77.9	20.2
15038 - Frades	2 460	81.6	28.1	15084 - Tordoia	3 591	124.6	21.2
15042 - Lousame	3 463	93.6	20.1	15085 - Touro	3 778	115.3	15.5
15045 - Mazaricos	4 173	187.3	23.1	15086 - Trazo	3 263	101.3	32.0
15046 - Melide	7 538	101.3	29.1	15088 - Val do Dubra	4 033	108.6	29.1
15047 - Mesía	2 734	107.1	31.5	15089 - Vedra	5 059	52.8	22.7
15053 - Muros	8 960	72.9	21.5	36016 - Dozón	1 174	74.2	31.1
15056 - Negreira	6 936	115.1	17.5	36017 - A Estrada	21 025	280.8	18.6
15057 - Noia	14 472	37.2	20.7	36020 - Agolada	2 585	147.9	31.4
15059 - Ordes	12 776	157.2	15.1	36024 - Lalín	20 005	326.8	22.3
15060 - Oroso	7 413	72.6	18.6	36044 - Pontecesures	3 062	6.7	12.8
15062 - Outes	6 691	99.7	21.2	36047 - Rodeiro	700	154.9	31.3
15065 - Padrón	8 643	48.4	22.2	36052 - Silleda	8 772	168.0	19.7
15066 - O Pino	4 706	132.1	21.8	36056 - Valga	6 062	40.6	20.1
15067 - Pobra	9 623	32.5	20.4	36059 - Vila de Cruces	5 556	155.0	21.9

Table 3
Selected candidate distributions response. The response function is usually chosen to ensure appropriate restrictions on the parameter space: *exponential function* to ensure positivity and *identity function* if the parameter space is unrestricted.

Distributions	Density	Parameters	Response functions
Normal	$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$	$\mu \in \mathbb{R}, \sigma^2 > 0$	$h^{\mu}(\eta) = \eta, h^{\sigma^2}(\eta) = exp(\eta)$
Log-normal	$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2 y}} \exp\left(-\frac{(\log(y)-\mu)^2}{2\sigma^2}\right)$	$\mu \in \mathbb{R}, \sigma^2 > 0$	$h^{\mu}(\eta) = \eta, h^{\sigma^2}(\eta) = exp(\eta)$
Truncated normal	$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right) \frac{1}{\sigma\left(-\Phi(\frac{-\mu}{\sigma})\right)}$	$\mu \in \mathbb{R}, \sigma^2 > 0$	$h^{\mu}(\eta) = \eta, h^{\sigma^2}(\eta) = exp(\eta)$
Inverse Gaussian	$p(y \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2 y^3}} \exp\left(-\frac{(y-\mu)^2}{2y\mu^2\sigma^2}\right)$	$\mu, \sigma^2 > 0$	$h^{\mu}(\eta) = h^{\sigma^2}(\eta) = exp(\eta)$
Gamma	$p(y \mid \mu, \sigma) = \left(\frac{\sigma}{\mu}\right)^{\sigma} \frac{y^{\sigma-1}}{\Gamma(\sigma)} \exp\left(-\frac{\sigma}{\mu}y\right)$	$\mu, \sigma > 0$	$h^{\mu}(\eta) = h^{\sigma^2}(\eta) = exp(\eta)$

Table 4
Comparison of DIC values for the candidate distributions.

Distribution	DIC
Normal	87 816.7
Truncated normal	87 819.0
Gamma	86 880.6
Inverse Gaussian	86 872.6
Log-normal	86 821.4

A.2. Model choice

Distribution regression models can handle a variety of complex distributions of the response variable. The statistical literature normally suggests one suppose the response variable to follow a normal distribution. However, in the present work, potassium concentrations with a log-normal distribution, truncated normal distribution, inverse Gaussian, gamma and normal distribution were contemplated (Table 3).

Users need to choose a suitable response distribution to construct an adequate distributional regression model. To this end, we used DIC (Spiegelhalter et al., 2002) and quantile–quantile plots. These methods have been validated in previous simulation studies to be adequate for this task, Klein et al. (2015).

A rule of thumb says that DIC differences of 10 and more between two competing models indicate the model with the lower DIC to be superior (Klein et al., 2015). In the present work, the DIC showed the log-normal distribution to provide the best and most parsimonious fit (see Table 4).

References

Augustin, N.H., Sauleaub, E.A., Wood, S.N., 2012. On quantile–quantile plots for generalized linear models. Comput. Statist. Data Anal. 56, 2404–2409.

Belitz, C., Brezger, A., Klein, N., Kneib, T., Lang, S., Umlauf, N., 2015. BayesX: Software for Bayesian Inference in Structured Additive Regression Models. Version 3.0. Avaliable from http://www.BayesX.org/.

Belitz, C., Brezger, A., Kneib, T., Lang, S., Umlauf, N., 2016. BayesX: Software for Bayesian Inference in Structured Additive Regression Models. Version 1.1. Avaliable from http://www.BayesX.org/.

Besag, J., York, J., Mollié, A., 1991. Bayesian image restoration, with two applications in spatial statistics. Ann. Inst. Statist. Math. 43. http://dx.doi.org/10.1007/BF00116466.

Brezger, A., Lang, S., 2006. Generalized structured additive regression based on Bayesian P-splines. Comput. Statist. Data Anal. 50, 967–991.

Clinical and Laboratory Standards Institute 2008. Defining, Establishing and Verifying Reference Intervals in the Clinical Laboratory; Aproved Guideline- Third Edition. CLSI document EP28-A3c. Wayne P.A.

Eilers, P.H., Marx, B.D., 1996. Flexible smoothing with B-splines and penalties. Statist. Sci. 11, 89–121.

Fahrmeir, L., Kneib, T., Lang, S., Marx, B., 2013. Regression. Models. Methods and Applications. Springer, Heidelberg/ Berlin.

Green, L., 2006. Queueing analysis in healthcare. In: Hall, R.W. (Ed.), Patient Flow:Reducing Delay in Healthcare Delivery. Springer, New York, pp. 281–308.

Hastie, T.J., Tibshirani, R.J., 1990. Generalized Additive Models. Chapman-Hall, London.
Instituto Galego de Estatística 2015. Banco de Datos: Cifras poboacionais de Referencia. Available from http://www.ige.eu/igebdt/igeapi/datos/5230/0:2002,1:0,2:0,9915:12.
Klein, N., Kneib, T., 2016. Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. Stat. Comput. 26, 841–860.
Klein, N., Kneib, T., Klasen, S., Lang, S., 2014. Bayesian structured additive distributional regression for multivariate responses. J. Roy. Statist. Soc. Ser. C 64, 569–591.
Klein, N., Kneib, T., Lang, S., Shon, A., 2015. Bayesian structured additive distributional regression with an application to regional income inequality in Germany. Ann. Appl. Stat. 9, 1024–1052.
Kneib, T., Heinzl, F., Brezger, A., Bove, D., Klein, N., 2015. BayesX: R Utilities Accompanying the Software Package BayesX. R package version 0.2-9. Avaliable from https://CRAN.R-project.org/package=BayesX.
Lang, S., Brezger, A., 2004. Bayesian P-splines. J. Comput. Graph. Statist. 13, 183–212.
Mankowska, D.S., Meisel, F., Bierwirth, C., 2014. The home health care routing and scheduling problem with interdependent services. Health Care Manage. Sci. 17, 15–30. http://dx.doi.org/10.1007/s10729-013-9243-1.
Marra, G., Radice, R., 2017. A bivariate copula additive model for location, scale and shape. Comput. Stat. Data Anal. 112, 99–113.
Rigby, A., Stasinopoulos, D.M., 2005. Generalized additive models for location, scale and shape (with discussion). Appl. Stat. 54, 507–554.
Spiegelhalter, D.J., Best, N.J., Carlin, B.P., Van Der Linde, A., 2002. Bayesian measures of model complexity and fit. J. R. Stat. Soc. Ser. B Stat. Methodol. 64, 583–639.
Stankovic, A.K., Smith, S., 2004. Elevated serum potassium values. The Role of Preanalytic Variables. Am. J. Clin. Pathol. 121, 105–112.
Tanner, M., Kent, N., Smith, B., Fletcher, S., Lewer, M., 2008. Stability of common biochemical analytes in serum gel tubes subjected to various storage temperatures and times pre-centrifugation. Ann. Clin. Biochem. 45, 375–379.
Umlauf, N., Adler, D., Kneib, T., Lang, S., Zeileis, A., 2015. Structured additive regression models: An R interface to bayesX. J. Stat. Softw. 63, 1–46. Avaliable from http://www.jstatsoft.org/v63/i21/.