On the goodness-of-fit of generalized linear geostatistical models

Emanuele Giorgi

Lancaster Medical School, Lancaster University, Lancaster, UK

January 15, 2018

Abstract

We propose a generalization of Zhang's coefficient of determination to generalized linear geostatistical models and illustrate its application to river-blindness mapping. The generalized coefficient of determination has a more intuitive interpretation than other measures of predictive performance and allows to assess the individual contribution of each explanatory variable and the random effects to spatial prediction. The developed methodology is also more widely applicable to any generalized linear mixed model.

Keywords: coefficient of determination; generalized linear geostatistical models; goodness-of-fit.

1 Introduction

Generalized linear geostatistical models (GLGMs) are a class of mixed models where, conditional on a realisation of a Gaussian process $\mathcal{S} = \{S(x) : x \in A \subset \mathbb{R}^2\}$ in a study area A, the outcome of interest Y_i, for $i = 1, \ldots, n$, follows a classical generalized linear model (GLM) (McCullagh & Nelder, 1989). Hence, the following properties hold.

- The Y_i, conditional on $\mathcal{S}$, are a set of mutually independent variables with mean

$$E[Y_i|S(x_i)] = m_i\mu_i = m_i g^{-1}(\eta_i)$$

and variance

$$\text{Var}[Y_i|S(x_i)] = m_i V(\mu_i),$$

where: m_i is an offset (e.g. number of trials for a Binomial response); η_i is the linear predictor; $g(\cdot)$ is the link function; and $V(\cdot)$ the variance function.

- $\eta_i = d(x_i)^\top \beta + S(x_i)$ where $d(x_i)$ is a vector of explanatory variables associated with location x_i and β is a vector of regression coefficients.
- The conditional distribution of Y_i belongs to the exponential family.

In this technical note, we address the following question: how should we assess the contribution of the explanatory variables $d(x_i)$ and of the random effects $S(x_i)$ to our predictive inferences?

To answer this question, we propose a generalization of the coefficient of determination proposed by Zhang (2017) to GLGMs and show its application to a geostatistical data-set on river-blindness. For classical GLMs, Zhang's coefficient is defined as

$$R^2_{GLM} = 1 - \frac{\sum_{i=1}^n c_V(y_i, \hat{y}_i\{d(x_i)\})}{\sum_{i=1}^n c_V(y_i, \hat{y}_0)}, \tag{1}$$

where: $\hat{y}_i\{d(x_i)\}$ is the prediction for Y_i based on $d(x_i)$ by plugging-in the estimated regression coefficients via maximum likelihood; $\hat{y}_0$ is the prediction from a GLM with an intercept only; and

$$c_V(a, b) = \left\{ \int_a^b \sqrt{1 + [V'(u)]^2}\, du \right\}^2, \quad a, b \in \mathbb{R}$$

which measures the change in the variance function $V(\cdot)$ for a change in the mean from a to b. When Y_i is Gaussian and $g(\cdot)$ is the identity function, the numerator in (1) reduces to the residual sum of squares, i.e. $c_V(y_i, \hat{y}_i\{d(x_i)\}) = \sum_{i=1}^n (y_i - \hat{y}_i\{d(x_i)\})^2$. Zhang (2017) also shows that (1) does not overstate the proportion of explained variance by the explanatory variables compared to other generalizations of the coefficient of determination to GLMs that are based on the likelihood ratio (Maddala, 1983; Cox & Snell, 1989; Nagelkerke, 1991). Furthermore, unlike the generalization by Cameron & Windmeijer (1997) based on the Kullback-Leibler divergence, Zhang's coefficient of determination is also defined for quasi-models (Wedderburn, 1974) and, therefore, does not require the full specification of the likelihood function.

2 A generalization of Zhang's coefficient of determination to GLGMs

Our generalization of Zhang's coefficient of determination is based on the intuitive interpretation of random effects as accounting for the effect of unmeasured variables.

For simplicity, consider a GLM with two explanatory variables $D_1(x)$ and $D_2(x)$, hence

$$\eta_i = \beta_0 + \beta_1 D_1(x_i) + \beta_2 D_2(x_i), \quad \text{for } i = 1, \ldots, n.$$

Note that the two explanatory variables appear in the above equation in upper-case letters because we have not yet conditioned on them. Under such model, conditioning only on one of the two covariates might induce residual spatial correlation in the outcome Y_i. Hence, if, for example, we condition on $D_1(x) = d_1(x)$, a natural model for the data would be a GLGM

where $d_1(x)$ is used as an explanatory variable and $S(x)$ is used to account for the residual effect $\beta_2 D_2(x)$. This argument can also be easily extended to any number of measured and unmeasured variables.

It follows that, conditionally on a realisation of $S^\top = (S(x_1), \ldots, S(x_n))$, a natural approach to quantify the total variation in $Y^\top = (Y_1, \ldots, Y_n)$ is through

$$\sum_{i=1}^{n} c_V(y_i, \hat{y}\{d(x_i), S(x_i)\}), \tag{2}$$

where $\hat{y}\{d(x_i), S(x_i)\}$ is the prediction for Y_i based on the vector of explanatory variables $d(x_i)$ and the realisation of $S(x_i)$. Since S is not observed, we can use its predictive distribution, defined as the distribution of S conditional on $y^\top = (y_1, \ldots, y_n)$ and the covariates $d^\top = (d(x_1), \ldots, d(x_n))$ (henceforth $S|(y, d)$), to compute (2). More specifically, we average (2) over the distribution of $S|(y, d)$, which leads to

$$R^2_{GLGM} \;=\; 1 - \frac{E_{S|(y,d)}[\sum_{i=1}^{n} c_V(y_i, \hat{y}\{d(x_i), S(x_i)\})]}{\sum_{i=1}^{n} c_V(y_i, \hat{y}_0)} \tag{3}$$

In the case of a linear geostatistical model, obtained by setting $m_i = 1$, $g^{-1}(\eta_i) = \eta_i$ and $V(\mu_i) = \tau^2$ for all i, the expectation of (2) reduces to

$$(y - D\beta)^\top (y - D\beta) + \xi^\top [\xi - 2(y - D\beta)] + \mathrm{tr}(\Omega),$$

where: D is a matrix of covariates; $\xi = \Sigma(\Sigma + I\tau^2)^{-1}(y - D\beta)$, with Σ and I denoting the covariance matrix of the marginal distribution of S and the identity matrix, respectively; and, finally, $\Omega = \Sigma - \Sigma(\Sigma + I\tau^2)^{-1}\Sigma$.

For non-Gaussian responses, the expectation of (2) is generally not available in closed form. We then propose to use a Monte Carlo Markov chain (MCMC) algorithm to simulate from $S|(y, d)$ and approximate (3) with

$$R^2_{GLGM} \approx 1 - \frac{\frac{1}{B}\sum_{j=1}^{B}\sum_{i=1}^{n} c_V(y_i, \hat{y}\{d(x_i), s_{(j)}(x_i)\})}{\sum_{i=1}^{n} c_V(y_i, \hat{y}_0)} \tag{4}$$

where $s_{(j)}(x_i)$ is the j-th out of B Monte Carlo samples for the i-th component of $S|(y, d)$.

We can also define the coefficient of partial determination for the vector of explanatory variables d given S as

$$\tilde{R}^2_{GLGM} = 1 - \frac{E_{S|(y,d)}\,[\sum_{i=1}^{n} c_V(y_i, \hat{y}\{d(x_i), S(x_i)\})]}{E_{S|(y,1)}\,[\sum_{i=1}^{n} c_V(y_i, \hat{y}\{1, S(x_i)\})]}, \tag{5}$$

where $\hat{y}_i\{1, S(x_i)\}$ is the prediction for Y_i based on $S(x_i)$ but excluding the explanatory variables $d(x_i)$ from the model. We interpret (5) as the fraction of explained variation in the response Y by the explanatory variables d but unexplained by the spatial random effects S.

In the next example, we compute (3) and (5) by plugging-in the maximum likelihood estimates of the regression coefficients. These are obtained using the Monte Carlo likelihood method (Christensen, 2004) implemented in the R package PrevMap (Giorgi & Diggle, 2017). We simulate from $S|(y, d)$ using a Laplace sampling technique described in detail in Section 2.1 of Giorgi & Diggle (2017).

3 Example: River-blindness mapping in Liberia

River-blindness is an infectious disease caused by the parasite *Onchocerca volvulus* and is transmitted by a black fly of the genus *Simulium*. We analyse data from 90 communities in Liberia, where people were tested by palpation for the presence of skin nodules caused by the disease; for an Africa-wide analysis of these data, see Zouré et al. (2014).

Let x_i be the location of the i-th sampled community, where y_i out of n_i randomly selected individuals tested positive. Our model for the data is a GLGM, where the Y_i conditionally on $S(x_i)$ are mutually independent Binomial variables with number of trials n_i and probability of having skin nodules $p(x_i)$, such that

$$\log\left\{\frac{p(x_i)}{1-p(x_i)}\right\} = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + S(x_i), \tag{6}$$

where $x_{i,1}$ and $x_{i,2}$ are the abscissa and ordinate components of the geographical location x_i. The reason for using a linear trend in x_i is shown in Figure 1, where the map of the empirical nodule prevalence shows an increase in the values as we move further from the coast in the north-east direction. Finally, we model $S(x)$ as a zero-mean Gaussian process with isotropic exponential covariance function having variance σ^2 and scale parameter ϕ.

The maximum likelihood estimates of the model parameters and their 95% confidence intervals are reported in Table 1. We observe that the use of the explanatory variables leads to a remarkable reduction in the values of the estimated σ^2 and ϕ. The fitted GLGM explains about 59% of the variation in nodule prevalence compared to 27% from a classical GLM where $S(x) = 0$ for all x. However, the small value of 1% for the coefficient of partial determination, $\tilde{R}^2$, indicates that the point estimates from the GLGM with covariates, given by (6), are strongly similar to a model without covariates, where $\beta_1 = \beta_2 = 0$. Nonetheless, Figure 2 shows that the standard errors for the estimated nodule prevalence (computed using Monte Carlo samples from $S|(y,d)$ while pugging-in the Monte Carlo maximum likelihood estimates) at the observed locations from the model with covariates are smaller almost everywhere than those from the model with only the intercept. More precisely, the largest relative reduction in the standard errors is of about 10%.

Table 1: Monte Carlo maximum likelihood estimates with associated 95% confidence intervals for the regression coefficients of the model with and without covariates defined in Section 3.

	Without covariates		With covariates	
Term	Estimate	95% CI	Estimate	95% CI
β_0	-1.941	(-3.312, -0.571)	-6.327	(-9.126, -3.528)
$\beta_1 \times 10^3$			2.761	(0.223, 5.299)
$\beta_2 \times 10^3$			4.784	(2.208, 7.360)
σ^2	0.791	(0.075, 8.295)	0.145	(0.055, 0.384)
ϕ	395.050	(32.608, 4786.143)	68.526	(20.438, 229.755)
$R^2_{GLM} = 27\%; R^2_{GLGM} = 59\%; \tilde{R}^2_{GLGM} = 1\%$				

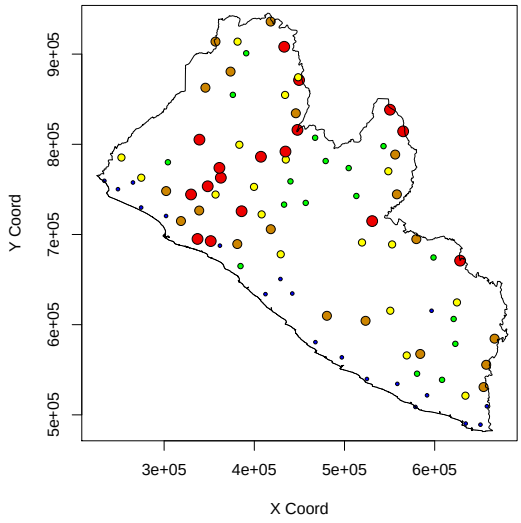

Figure 1: Map of the empirical nodule prevalence. The radius of each point is proportional the quintile class within which the associated prevalence falls, with larger radiuses corresponding to higher quintile clasees.

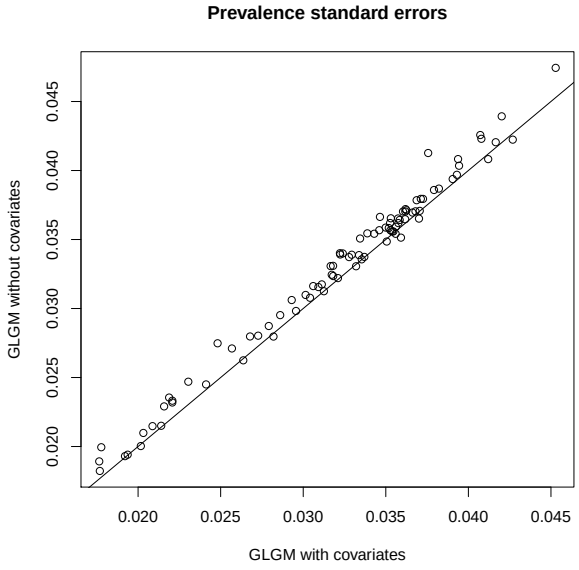

Figure 2: Standard errors for the estimated nodule prevalence from a Binomial geostatistical model without covariates against one with covariates as defined in Section 3. The solid line is the identity line.

4 Discussion

We have introduced a generalization of Zhang's coefficient of determination to quantify the proportion of explained variation in the outcome of interest by the covariates and/or the residual spatial random effects. This has a more intuitive interpretation than other measures of predictive performance, such as mean square errors, and also allows to quantify the individual contribution of each component of the linear predictor to spatial prediction. Although our focus was on geostatistical models, the developed methodology can be applied to any generalized linear mixed model.

Through an example on river-blindness mapping, we have quantified the impact of the adopted explanatory variables on the spatial estimates of prevalence. The proposed generalization of the coefficient of partial determination, $\tilde{R}^2_{GLGM}$, indicated that the impact of these on the point estimates of prevalence was negligible. We have also shown that the reduction in the standard errors, albeit small, was however more tangible than the change in the point estimates after adjusting for the north-east trend in disease prevalence. Hence, our recommendation is that $\tilde{R}^2_{GLGM}$ should not be used as a stand-alone tool but should be complemented with other measures that assess the impact on the accuracy of the spatial estimates.

Future research will aim to extend the methods of Section 2 to point process models, including log-Gaussian Cox processes.

Acknowledgements

Emanuele Giorgi holds an MRC fellowship in Biostatistics (MR/M015297/1).

References

Cameron, A. C. & Windmeijer, A. G. (1997). An r-squared measure of goodness of fit for some common nonlinear regression models. *Econometrics* **77**, 329–342.

Christensen, O. F. (2004). Monte Carlo maximum likelihood in model-based geostatistics. *Journal of Computational and Graphical Statistics* **3**, 702–718.

Cox, D. R. & Snell, E. J. (1989). *Analysis of binary data*. Chapman and Hall, London, 2nd ed.

Giorgi, E. & Diggle, P. J. (2017). Prevmap: an R package for prevalence mapping. *Journal of Statistical Software* **78**, 1–29.

Maddala, G. S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.

McCullagh, P. & Nelder, J. (1989). *Generalized Linear Models*. Chapman and Hall, London, 2nd ed.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika* **78**, 691–692.

Wedderburn, R. W. M. (1974). Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. *Biometrika* **61**, 439–447.

Zhang, D. (2017). A coefficient of determination for generalized linear models. *The American Statistician* In press. doi:10.1080/00031305.2016.1256839.

Zouré, Honorat, G. M., Noma, M., Tekle, Afework, H., Amazigo, U. V., Diggle, P. J., Giorgi, E. & Remme, J. H. F. (2014). The geographic distribution of onchocerciasis in the 20 participating countries of the african programme for onchocerciasis control: (2) pre-control endemicity levels and estimated number infected. *Parasites & Vectors* **7**.