Estimation of entropy measures for categorical variables with spatial correlation

Stima di misure di entropia per variabili categoriche con correlazione spaziale

Linda Altieri, Giulia Roli
Department of Statistical Sciences, University of Bologna, via Belle Arti, 41, 40126, Bologna, Italy.

Abstract

Entropy is a measure of heterogeneity widely used in applied sciences, often when data are collected over space. Recently, a number of approaches has been proposed to include spatial information in entropy. The aim of entropy is to synthesize the observed data in a single, interpretable number. In other studies the objective is, instead, to use data for entropy estimation; several proposals can be found in the literature, which basically are corrections of the estimator based on substituting the involved probabilities with proportions. In this case, independence is assumed and spatial correlation is not considered. We propose a path for spatial entropy estimation: instead of intervening on the global entropy estimator, we focus on improving the estimation of its components, i.e. the probabilities, in order to account for spatial effects. Once probabilities are suitably evaluated, estimating entropy is straightforward since it is a deterministic function of the distribution. Following a Bayesian approach, we derive the posterior probabilities of a binomial distribution for categorical variables, accounting for spatial correlation. A posterior distribution for entropy can be obtained, which may be synthesized as wished and displayed as an entropy surface for the area under study.

Riassunto

L'entropia è una misura di eterogeneità ampiamente utilizzata nelle scienze applicate, dove spesso i dati sono georeferenziati. Recentemente, sono stati proposti svariati approcci per includere informazione spaziale nell'entropia, il cui scopo è sintetizzare l'osservato in un numero interpretabile. In altri studi, invece, l'obiettivo è utilizzare i dati per stimare l'entropia di un processo: esistono diverse proposte in letteratura, che sono correzioni dello stimatore basato sulla sostituzione delle probabilità con proporzioni. In questo caso, si assume indipendenza e non è considerata la correlazione spaziale. Proponiamo un nuovo percorso per la stima dell'entropia: invece di intervenire sullo stimatore globale, miglioriamo la stima delle sue componenti, le probabilità, per tenere conto di effetti spaziali. Una volta che le probabilità sono accuratamente valutate, si può stimare direttamente l'entropia, essendo questa una funzione deterministica della distribuzione di probabilità. Con un approccio Bayesiano, deriviamo la distribuzione a posteriori per variabili categoriche spazialmente correlate. Si ottiene quindi una distribuzione a posteriori per l'entropia, che può essere sintetizzata a piacere e diventa un superficie di entropia per l'area di interesse.

Keywords: Spatial entropy, entropy estimation, CAR models, categorical variables

1 Introduction

Shannon's entropy is a successful measure in many fields, as it is able to synthesize several concepts in a single number: entropy, information heterogeneity, surprise, contagion. The entropy of a categorical variable X with $I < \infty$ outcomes is

$$H(X) = \sum_{i=1}^{I} \log(p(x_i)) \log\left(\frac{1}{p(x_i)}\right), \tag{1}$$

where $p(x_i)$ is the probability associated to the i-th outcome (Cover and Thomas, 2006). The flexibility of such index and its ability to describe any kind of data, including categorical variables, motivate its diffusion across applied fields such as geography, ecology, biology and landscape studies. Often, such disciplines deal with spatial data, and the inclusion of spatial information in entropy measures has been the target of intensive research (see, e.g., Batty, 1976, Leibovici, 2009, Leibovici et al., 2014). In several case studies, the interest lies in describing and synthesizing data. This is usually not a simple task: large amounts of data require advanced computational tools, qualitative variables have limited possibilities, and, when data are georeferenced, spatial correlation should be accounted for. When it comes to measuring the entropy of spatial data, we suggest an approach proposed in Altieri et al. (2018), which allows to decompose entropy into a term quantifying the spatial information, and a second term quantifying the residual heterogeneity.

In other cases, though, the aim lies in estimating the entropy of a phenomenon, i.e. in making inference rather than description. Under this perspective, a stochastic process is assumed to generate the data according to an unknown probability function and, consequently, an unknown entropy. One realization of the process is observed and employed to estimate such entropy. The standard approach relies on the so-called 'plug-in' estimator, presented in Paninski (2003), which substitutes probabilities with observed relative frequencies in the computation of entropy:

$$\widehat{H}_p(X) = \sum_{i=1}^{I} \log(\widehat{p}(x_i)) \log\left(\frac{1}{\widehat{p}(x_i)}\right), \tag{2}$$

where $\widehat{p}(x_i) = n_i/n$ is the relative amount of observations of category i over n data. It is the non-parametric as well as the maximum likelihood estimator (Paninski, 2003), and performs well when $I < \infty$ is known (Antos and Kontoyiannis, 2001). For unknown or infinite I, estimator (2) is known to be biased; the most popular proposals at this regard consist of corrections of the plug-in estimator: see, for example, the Miller-Madow (Miller, 1955) and the jack-knifed corrections (Efron and Stein, 1981). Recently, Zhang (2012) proposed a non-parametric solution with faster decaying bias and upper limit for the variance when $I = \infty$. Under a Bayesian framework, the most widely known proposal is the NSB estimator (Nemenman et al., 2002), improved by Archer et al. (2014) as regards the prior distribution. Other approaches, linked to machine learning methods, directly estimate entropy relying on the availability of huge amounts of data (Hausser and Strimmer, 2009). In all these works, independence among realizations is assumed.

Two main limits concern entropy estimation. Firstly, the above mentioned proposals only focus on correcting or improving the performance of (2). Secondly, no study is available about estimat-

ing entropy for variables presenting spatial association: the assumption of independence is never relaxed, while spatial entropy studies do not consider inference.

In this paper, we take a perspective to entropy estimation, which moves the focus from the index itself to its components. Entropy, as defined in equation (1), is a deterministic function of the probability mass function (pmf) of the variable of interest. Therefore, once the pmf is properly estimated, the subsequent steps are straightforward. In the case of categorical variables following, e.g., a multinomial distribution, the crucial point is to estimate the distribution parameters. A Bayesian approach allows to derive the pmf of such distribution and can be extended to account for spatial correlation among categories. After obtaining a posterior distribution for the parameters, this is used to compute the posterior distribution of entropy as a transformation. Thus, a point estimator of entropy can be, e.g., the mean of the posterior distribution of the transformation; credibility intervals and other syntheses may be obtained via the standard tools of Bayesian inference. This approach can be used for non-spatial settings as well; in the spatial context, coherently with standard procedures for variables linked to areal and point data, the estimation output is a smooth spatial surface for the entropy over the area under study.

The paper is organized as follows. Section 2 summarizes the methodology for Bayesian spatial regression and shows how to obtain the posterior distribution and the Bayesian estimator of entropy. Then, Section 3 assesses the performance of the proposed method on simulated data for different spatial configurations. Lastly, Section 4 discusses the main results.

2 Bayesian spatial entropy estimation

For simplicity of presentation, we focus on the binary case. Let X be a binary response variable with $x_1 = 1$ and $x_2 = 2$; consider a series of n realizations indexed by $u = 1, \ldots, n$, each carrying an outcome $x_u \in \{1, 2\}$. This may be thought of as a n-variate variable, or alternatively as a sequence of variables $X_1, \ldots X_n$, which are independent, given the distribution parameters and any effects modelling them. For a generic X_u, the simplest model is:

$$X_u \sim Ber(p_u) \tag{3}$$

$$logit(p_u) = z_u^{'}\beta \tag{4}$$

in absence of random effects, where z_u are the covariates associated to the u-th unit.

To the aim of including spatial correlation, consider n realizations from a binary variable over the two-dimensional space, where u identifies a unit via its spatial coordinates. Let us consider the case of realizations over a regular lattice of size $n = n_1 \times n_2$, where u identifies each cell centroid. The sequence $X_1, \ldots, X_n$ is now no longer independent, but spatially correlated. In order to define the extent of such correlation for grid data, the notion of neighbourhood must be introduced, linked to the assumption that occurrences at certain locations are influenced by what happens at surrounding locations, i.e. their neighbours. The simplest way of representing a neighbourhood system is via an adjacency matrix: for n spatial units, $A = \{a_{uu'}\}_{u,u'=1,\ldots,n}$ is a square $n \times n$ matrix such that $a_{uu'} = 1$ when unit u and unit u' are neighbours, and $a_{uu'} = 0$ otherwise; in other words, $a_{uu'} = 1$ if $u' \in \mathcal{N}(u)$, the neighbourhood of area u, and diagonal elements are all zero by default.

In the remainder of the paper, the word 'adjacent' is used accordingly to mean 'neighbouring', even when this does not correspond to a topological contact. The most common neighbourhood systems for grid data are the '4 nearest neighbours', i.e. a neighbourhood formed by the 4 pixels sharing a border along the cardinal directions, and the '12 nearest neighbours', i.e. two consequent pixels along each cardinal direction plus the four ones along the diagonals.

Auto-models provide a way of including spatial correlation, by explaining a response via the response values of its neighbours. They are thus developed by combining a logistic regression model with autocorrelation effects, and are initially developed for the analysis of plant competition experiments and then extended to spatial data in general. Besag (1974) proposed to model spatial dependence among random variables directly (rather than hierarchically) and conditionally (rather than jointly). The autologistic model for spatial data with binary responses emphasises that the explanatory variables are the surrounding array variables themselves; a joint Markov random field is imposed for the binary data. A recent variant of this model, substituting (4) with (5), is proposed by Caragea and Kaiser (2009):

$$logit(p_u) = z(u)'\beta + \sum_{u' \in \mathcal{N}(u)} \eta_{u'}(X_{u'} - \mu_{u'}) \tag{5}$$

where η parametrizes dependence on the neighbourhood and, in the simplest case, $z(u)'\beta = \beta_0$ only includes an intercept. Parameter $\mu_u = \exp(z(u)'\beta)/(1+\exp(z(u)'\beta))$ represents the expected probability of success in the situation of spatial independence.

An analogous rewriting in the form of a CAR model (Cressie, 1993), in absence of covariate information, is

$$\begin{aligned} &logit(p_u) = \beta_0 + \phi_u \\ &\phi \sim MVN_n(0, \Sigma) \\ &\Sigma = [\tau(D - \rho A)]^{-1} \end{aligned} \tag{6}$$

where $\phi = (\phi_1, \ldots, \phi_n)'$ is a spatial effect with a structured covariance matrix Σ, which depends on a precision parameter τ and a dependence parameter $\rho \in [-1, 1]$ quantifying the strength and type of the correlation between neighbouring units. The symbol A denotes the adjacency matrix reflecting the neighbourhood structure, and D is a diagonal matrix, where each element contains the row sums of A.

The estimation of the parameters for Bayesian spatial logit regression models may proceed via MCMC methods or the INLA approach. We exploit the latter (Rue et al., 2009) and obtain a posterior distribution for the parameters of the probability of success for each grid cell. A synthesis, such as the posterior mean, is chosen in order to obtain an estimate for p_u over each cell. Such estimate is used for the computation of a local estimated entropy value for each pixel:

$$\widehat{H}(X)_u = \hat{p}_u \log\left(\frac{1}{\hat{p}_u}\right) + (1 - \hat{p}_u) \log\left(\frac{1}{1 - \hat{p}_u}\right). \tag{7}$$

This way, an entropy surface is obtained for estimating the process entropy, whose smoothness may be tuned by the neighbourhood choice, or by the introduction of splines for the spatial effect. Any other surface can be obtained following the same approach for different aims, e.g. for plotting the entropy standard error or the desired credibility interval extremes.

3 Simulation study

To the aim of assessing the performance of the proposed entropy estimator, we generate binary data on a 40×40 grid under two spatial configurations: clustered and random. Figure 1 shows an example of the generated datasets. The underlying model is (6), with a 12 nearest neighbour structure for A, $\tau = 0.1$ and $\rho = \{0.999, 0.001\}$ for the two scenarios, respectively. For each scenario, 200 datasets are generated with varying values for β_0 so that the expectation of p_u in a situation of independence varies between 0.1 and 0.9; values for β_0 differ across replicates but are constant across scenarios, so that the proportion of pixels of each type is comparable.

Figure 1: Clustered scenario (left) and random scenario (right) - example with $\beta_0 = 0.57$.

Results show that fitting the model over the generated data leads to good estimates for the p_us. For all replicates on both scenarios, the estimated parameters are very close to the true ones, which are always included within the 95% credibility intervals. The proposed approach is able to produce good estimates for the probabilities of success, ensuring the goodness of the estimates for spatial entropy, which is a function of such probabilities.

Obtaining the entropy surface proceeds as follows. First, the posterior distribution for each p_u is synthesized with its posterior mean. This way, for each scenario and replicate we obtain a single number for $\hat{p}_u$ on every cell. Then, an entropy value is computed over each cell following (7) and a smooth spatial function is produced. An example is shown in Figure 2, where values range from 0 (dark areas in the figure) to $\log(2)$ (white areas). The clustered situation (left panel) shows a smoothly varying surface. By comparing the left panels of Figure 1 and 2, one can see that the entropy surface takes low values in areas where pixels are of the same type: white pixels in the top-left part in Figure 1, and black pixels in the top-right part of Figure 1, correspond to the darker areas of Figure 2 where entropy values are low. In the areas where white and black pixels mix, the entropy surface tends to higher values (whiter areas in Figure 2). The random configuration (right panel of Figure 2) has a constant entropy close to the maximum $\log(2)$; this is expected, as no spatial correlation influences the entropy surface in this scenario. Therefore, such spatial functions properly estimates the entropy of the underlying spatial process.

Thanks to the availablity of the marginal posterior distribution of all parameters, any other useful synthesis is straightforward to compute. An example is shown in Figure 3, where the standard error of the estimate over each cell is plotted. Again, it is possible to appreciate a smooth surface

for the clustered scenario, while the value is constant for the random one.

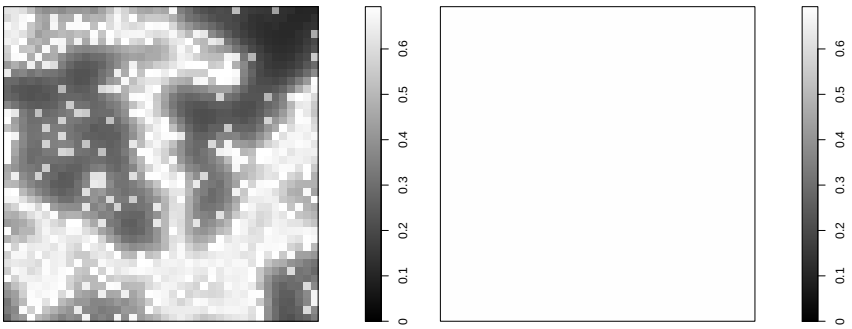

Figure 2: Example of estimated entropy surface for the two scenarios.

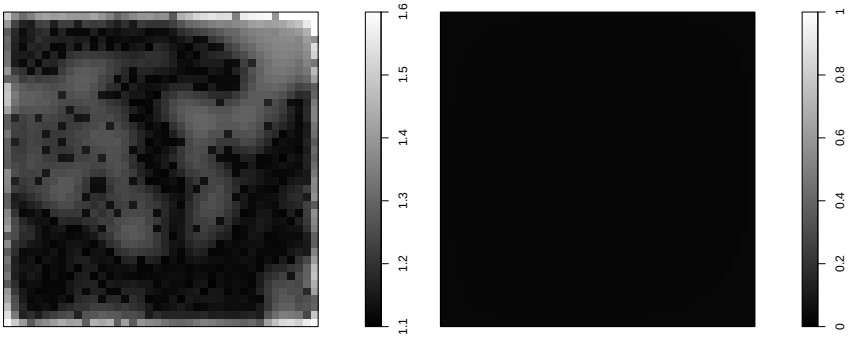

Figure 3: Example of the standard error of the estimate for the two scenarios.

4 Concluding remarks

In this paper, we describe an approach to entropy estimation which starts from rigorous posterior evaluation of its components, i.e. the probabilities. This way, we frame entropy within the theory of Bayesian models for spatial data, thus assembling the available results in this field.

Results from the simulation study enforce the validity of the approach in providing good estimates for the distribution parameters and, consequently, for entropy. The flexibility of the Bayesian paradigm allows to synthesize the posterior distribution of entropy as wished, in order to answer different potential questions.

Our procedure ensures realistic results, since, when the behaviour of a spatial process is under study, the basic hypothesis is that it is not constant but smoothly varying over space. In the same spirit, an appropriate spatial entropy measure is not a single number, rather it has to be allowed to vary over space as a smooth function.

The choice of the INLA approach allows to obtain results in a very reasonable time (minutes) for models including covariates and for very fine grids too, provided the model does not get extremely complicated in terms of random effects.

Acknowledgements
This work is developed under the PRIN2015 supported project 'Environmental processes and human activities: capturing their interactions via statistical methods (EPHASTAT)' [grant number 20154X8K23] funded by MIUR (Italian Ministry of Education, University and Scientific Research).
We thank an anonimous referee for the useful comments.

References

Altieri, L., D. Cocchi, and G. Roli (2018). A new approach to spatial entropy measures. *Environmental and Ecological Statistics 25*, 95–110.

Antos, A. and I. Kontoyiannis (2001). Convergence properties of functional estimates for discrete distributions. *Random structures and algorithms 19*, 163–193.

Archer, E., I. M. Park, and J. W. Pillow (2014). Bayesian entropy estimation for countable discrete distributions. *Journal of Machine Learning Research 15*, 2833–2868.

Batty, M. (1976). Entropy in spatial aggregation. *Geographical Analysis 8*, 1–21.

Besag, J. E. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B 36*, 192–236.

Caragea, P. and M. Kaiser (2009). Autologistic models with interpretable parameters. *Journal of Agricultural, Biological, and Environmental Statistics 14*, 281–300.

Cover, T. M. and J. A. Thomas (2006). *Elements of Information Theory. Second Edition*. Hoboken, New Jersey: John Wiley & Sons, Inc.

Cressie, N. A. C. (1993). *Statistics for spatial data (rev. ed.)*. New York, Wiley.

Efron, B. and C. Stein (1981). The jackknife estimate of variance. *Annals of statistics 9*, 586–596.

Hausser, J. and K. Strimmer (2009). Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research 10*, 1469–1484.

Leibovici, D. G. (2009). *Defining spatial entropy from multivariate distributions of co-occurrences*. Berlin, Springer: In K. S. Hornsby et al. (eds.): COSIT 2009, Lecture Notes in Computer Science 5756, pp 392-404.

Leibovici, D. G., C. Claramunt, D. LeGuyader, and D. Brosset (2014). Local and global spatio-temporal entropy indices based on distance ratios and co-occurrences distributions. *International Journal of Geographical Information Science 28*(5), 1061–1084.

Miller, G. (1955). *Note on the bias of information estimates*. Glencoe, IL free press: In H. Quastler (ed.) Information Theory in psychology II-B, pp. 95-100.

Nemenman, I., F. Shafee, and W. Bialek (2002). *Entropy and inference, revisited*. Cambridge, MA: MIT Press: In T. G. Dietterich, S. Becker, and Z. Ghahramani (eds.) Advances in neural information processing, 14.

Paninski, L. (2003). Estimation of entropy and mutual information. *Journal of Neural Computation 15*, 1191–1253.

Rue, H., S. Martino, and N. Chopin (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations (with discussion). *Journal of the Royal Statistical Society, Series B 71*, 319–392.

Zhang, Z. (2012). Entropy estimation in Turing's perspective. *Journal of Neural Computation 24*, 1368–1389.