Statistica & Applicazioni
Vol. III, n. speciale 1, 2005

Assessing Quality Using Routine Administrative Data: the Case of Preventable Hospitalizations

Marco Alfò [§]

Luciano Nieddu[‡]

Donatella Vicari[§]

Summary: *A vast literature has recently concerned the measurement of quality dimensions such as access, effectiveness, performance and outcome of health services supplied by national health care providers. The main concern is to achieve a classification of administrative areas with respect to observed quality indicators. We describe a simple and effective procedure to achieve this goal which allows powerful testing of the hypothesized cluster structure. We describe the performance of this method on a dataset on preventable hospitalizations (PPH) in Italy during 1998, in order to highlight clusters of regions with homogeneous relative risk.*

Keywords: *Unknown risk factors, Nonparametric ML, Spatial association, Statistical Mapping.*

1. Introduction

A wide class of models is now available to analyze univariate disease counts observed over a set of adjacent regions to identify areas with homogeneous relative risks by modelling the extra-Poisson variation. Proposals include Quasi-Likelihood (see, among others, Clayton and Kaldor, 1987, Breslow and Clayton, 1993, Langford et al., 1999) and Bayesian approaches (see e.g. Besag et al, 1991, Knorr-Held and Besag, 1998). Often, these approaches are based on restrictive parametric assumptions, which can be relaxed adopting the semiparametric approach proposed by Böhning et al. (1992) and Schlattmann and Böhning (1993). This approach is related to those developed by Aitkin (1999) for the analysis of overdispersed or clustered data and allows the classification of administrative areas on the basis of posterior probabilities of component membership, via a MAP (maximum a

[§] Dipartimento di Statistica, Probabilità e Statistiche Applicate – Università degli Studi di Roma "La Sapienza" – Piazzale A. Moro, 5, 00185 ROMA (e-mail: {marco.alfo;donatella.vicari}@uniroma1.it).

[‡] Facoltà di Economia. – Libera Università "S. Pio V". – via Delle Sette Chiese, 139, 00145 ROMA (e-mail: l.nieddu@luspio.it).

posteriori) approach. While powerful enough in general empirical conditions, such methods are not adequate to analyze spatial data where neighbouring areas can be dependent on each other. The semiparametric approach of Böhning et al. (1992) has been modified to take into account spatial dependence by Biggeri et al. (2003) and Alfò and Vitiello (2002).

In this paper, we will focus on semiparametric mapping models for the geographical analysis of preventable hospitalisation ratios in Italy. The observed data and the motivation for the working example will be introduced in the next section. The spatial feature will be considered via a standard mixture model that includes spatial information through a Gibbs distribution on mixture weights. The model is a likelihood version of the one proposed by Green and Richardson (2002) in the Bayesian context, with the major extension of explicitly allowing for class-specific strengths.

The paper is so organized: in section 2 the definition of PPH is discussed together with the motivation for using them as a measure of effectiveness of National Health Service. In sections 3 and 4, we describe finite mixture models and pseudo ML estimation of the proposed method. Experimental results are detailed in section 5 and concluding remarks are in section 6.

2. Working example

The foundation of the Italian National Health Service (NHS) in 1978 was inspired by ideas such as planning of health care, equity, public provision and decentralization. The promotion of primary care was the only route to achieve an effective and universal coverage, the goal being that of health for all. This approach was exclusively focused on perceived needs, giving little attention to health care demand. Since then, there has been a gradual shift of vision: rather than providing all possible care to everyone, delivering to all an effective and socially acceptable essential care.

We focus on the analysis of geographical variation in PPH rates recorded in Italy during 1998. Previous research in the UK suggests that some characteristics of primary care reflecting quality of practice are related to admission ratios for chronic diseases (see e.g. Griffith et al. 1997). The effects of socio-economic and racial/ethnic characteristics of the analyzed population (see e.g. Bindman et al., 1995; Fiscella et al., 2000) have raised severe questions about this interpretation. Statistical analyses of PPH rates have been discussed previously (see e.g. Parchman and Culler, 1999).

3. Finite mixture models

In this section a brief overview of standard semiparametric mapping methods will be given. Let us assume that the analyzed area can be

partitioned into n subregions and that counts o_i of observed cases have been recorded for each subregion. Conditionally on class membership k=1,…,K the observed counts O_i, $i=1,..,n$ are assumed to be Poisson random variables with class-specific parameter. Such an assumption can be motivated considering that, in the ideal case, areas belonging to each class should present a constant relative risk. In disease mapping the interest is usually focused on the relative risk vector $\boldsymbol{\lambda}=(\lambda_1,...,\lambda_n)^T$; using a log link, it can be modelled as follows:

$$\log \lambda_i = \log \lambda_0 + u_i + v_i \quad i=1,...,n \tag{1}$$

where λ_0 is the overall average risk and u_i and v_i represent individual heterogeneity and contextuality with respect to a specified neighborhood. This formulation, introduced by Besag et al. (1991), is referred to as the convolution model (Molliè, 1996). Various alternative specifications have been proposed for both random terms; our starting point is the semiparametric approach suggested, among others, by Böhning et al. (1992) and Aitkin (1999). Conditionally on class membership k, we assume that

$$O_i \,|\, \lambda_k \sim \text{Poisson}(E_i \lambda_k) \quad i=1,...,n \tag{2}$$

where E_i represents the number of expected cases in the $i-$th subregion after standardization with respect to known confounding factors, λ_k describes the departure of the standardized ratio for the k-th component from λ_0 and can be considered as a relative risk parameter for the $k-$th component. Adopting a discrete distribution $G(\cdot)$ for the λs results in at least two major advantages: robustness, and detection of clusters of subregions characterized by a homogeneous relative risk. The pseudolikelihood function can therefore be written as:

$$L(\cdot) = \prod_{i=1}^{n} \left\{ \sum_{k=1}^{K} f\left(O_i \,|\, \lambda_k\right) \pi_k \right\} \tag{3}$$

Here, locations λ_k and corresponding masses π_k represent unknown parameters, which have to be estimated. Denoting with $\boldsymbol{\delta}$ the parameter vector, we have:

$$\frac{\partial \log[L(\boldsymbol{\delta})]}{\partial \boldsymbol{\delta}} = \frac{\partial \ell(\boldsymbol{\delta})}{\partial \boldsymbol{\delta}} = \sum_{i=1}^{n} \sum_{k=1}^{K} w_{ik} \frac{\partial \log f_{ik}}{\partial \boldsymbol{\delta}} \tag{4}$$

where w_{ik} represents the posterior probability that the $i-$th area belongs to the $k-$th component of the mixture. We obtain pseudolikelihood equations which are weighted sums of those for an ordinary spatial GLM with weights w_{ik}; solving these equations for given w_{ik}s and updating weights from current parameters estimates is an EM algorithm. Differentiating the pseudolikelihood under the constraint $\sum_{k=1}^{K}\pi_k = 1$, we get $\hat{\pi}_k = \sum_{i=1}^{n} w_{ik}/n$, which is a standard result in finite mixture ML theory. The unknown number of locations K, can be estimated via penalized likelihood criteria, such as AIC or BIC. This approach has been modified to take into account an explicit spatial component by Biggeri et al. (2003) and it has been extended to the case of ecological regression by Alfò and Vitiello (2003).

4. Semiparametric Gibbs models

Since the autoregressive-type approach is unsatisfactory, in this section spatial dependence will be modelled through explicitly modelling prior probabilities following Green & Richardson (2002). Let N_i denote the neighbourhood for the $i-$th subregion; information on the ratios observed in neighbouring areas should be taken into account to determine the class the area belongs to, since it could be convenient to allocate adjacent areas to the same class. According to the work by Geman and Geman (1984), the class process can be modelled as a Markov Random Field and the Hammersley-Clifford theorem allows to impose a Gibbs distribution on the class process.

In the following, the conditional prior probability is denoted by π_{ik}. Given the neighbourhood system, the Gibbs distribution is defined by:

$$\pi_{ik} \propto \exp\left\{-\frac{1}{T_k}E_k\right\} = \exp\left\{\left[\tilde{\alpha}_k - \tilde{\beta}_k \sum_{s\in N_i} V(b_i(k)b_s(k))\right]\right\} \qquad (5)$$

where E_k represents the energy function, T_k denotes the temperature in the k-th class, and $b_i(k)$ is the indicator variable for the i-th subregion in the k-th class. The terms $\tilde{\alpha}_k$ and $\tilde{\beta}_k$ represent a class-specific constant related to the one-dimensional energy, and the smoothness (regularization) parameter, respectively. The potential function $V(\cdot)$ is usually defined as:

$$V(b_i(k)b_s(k)) = \begin{cases} -1 & \text{if } \sum_{k=1}^{K} b_i(k)b_s(k) = 1 \\ +1 & \text{otherwise} \end{cases} \quad (6)$$

in order to penalize neighbouring regions belonging to different classes and adopting the K-th class as reference category, the prior (conditional) probabilities π_{ik}, $k = 1,..,K-1$, can be written as:

$$\pi_{ik} \propto \exp\left[\alpha_k - \beta_k \sum_{s \in N_i} V(b_i(k)b_s(k)) \right] = \exp\left\{ \alpha_k + \beta_k \left[n_i(k) - \sum_{l \neq k} n_i(l) \right] \right\} \quad (7)$$

where $\alpha_k = \tilde{\alpha}_k - \tilde{\alpha}_K$ and $\beta_k = \tilde{\beta}_k - \tilde{\beta}_K$, $k = 1,..,K$. The term $n_i(l)$ represents the number of subregions in N_i belonging to the l-th class, $l = 1,..,K$. Expression (7) defines a multinomial logit for the prior probabilities of class membership. The class-specific intercept term α_k represents the log-prior probability of each class, regardless of the neighbouring subregions. The term β_k gives information on the strength of the class process describing how the class membership of each subregion is influenced by the class memberships of neighbouring subregions.

The proposed model represents an extension of standard models where parameters are assumed to be constant (and often arbitrarily fixed a priori); the choice of varying smoothness parameter should permit local rate stabilization without losing geographic resolution.

5. Working example (results)

As of 1998, there were 196 local health agencies (LHA) in Italy providing health care services to the population through independent public and private contracted structures working under their supervision. We used data on hospital discharges drawn from routine information collected in the national register of hospital discharges (the SDO register), maintained by the Ministry of Health. In particular, we considered admission standardized ratios for diabetes (ICD9cm codes 250.xx) aggregated at LHA level. Figure 1 shows the observed distribution of crude estimators for the relative risk, classified according to observed sextiles. We have first modelled observed ratios by using the standard finite mixture model (FM in the following). The model identifies K=7 components, where only 2 components show values for the estimated relative risk which are significantly higher than the average one (equal to 1.4157 and 2.4249 respectively). The results obtained by

adopting the FM approach and the extended approach here proposed are shown in Figure 2. Figure 2(a) displays the geographical representation of component membership; as can be observed, the estimated geographical distribution is highly scattered and shows no significant spatial correlation. To verify if spatial dependence could represent an additional source of extra-Poisson variation, we applied the proposed finite mixture model with Gibbs prior. The estimated distribution is quite similar to the one obtained by FM, as can also be argued looking at the significance test for the beta parameters in the multinomial regression model (7). These test statistics have to be used for fixed number of components K; care is needed when using standard asymptotic theory in a pseudo-ML contest. Values of relative risk parameters are not significantly different from those obtained via finite mixture and this shows that spatial dependence, if present, does not play a substantial role in the analysis. The only substantial effect is on the number of components which is selected via the BIC criterion. If the Gibbs prior distribution is used, only six components are estimated from the observed data; those components with higher values for the relative risk are left unchanged, while the spatial distribution affects only those components with smaller relative risk parameters. Components 1-3 (λ=0.836, 0.69 and 0.4814 respectively) are summarized by only two classes (with λ=0.7556 and 0.5118).

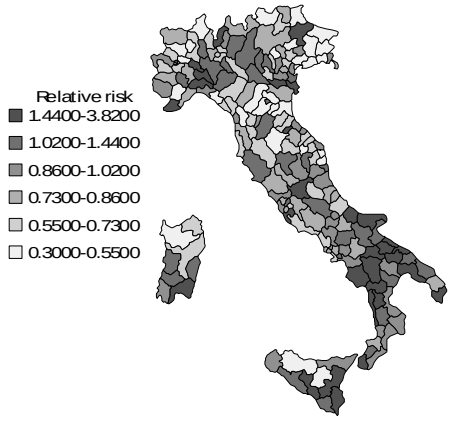
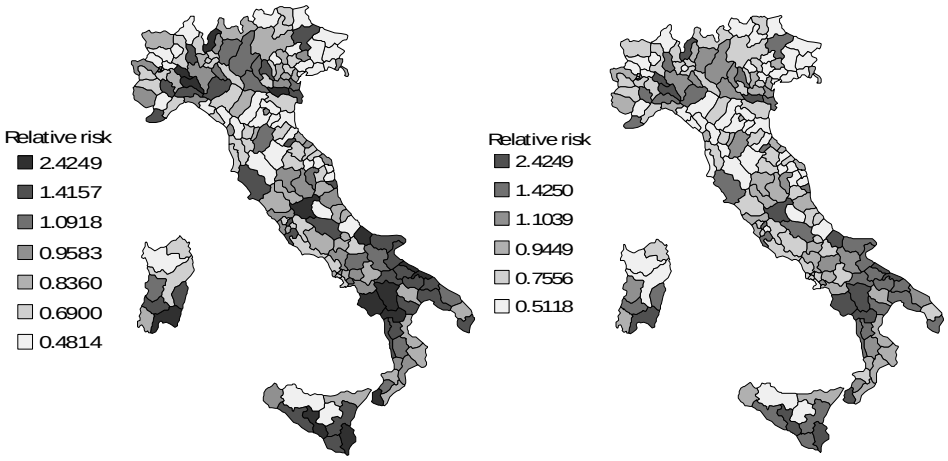


Figure 1. *Observed relative risks ratios classified by empirical sextiles***.**

Here the main focus is on assessing relevant differences between administrative areas in observed PPH hospitalisation ratios. While any methods based on mapping crude or estimated relative risk using thresholding fail to address this issue, the proposed model provides a classification of the LHAs, based on posterior probabilities of component membership, where the number of components can be formally tested. From a health policy perspective, one should simply select those LHAs in the first

(λ=2.4249) and/or second (λ=1.4250) component and try to proceed to a deeper analysis based on hospital and age-specific ratios.



(a) (b)

Figure 2. *Estimated relative risks ratios. Finite mixture model (a) and finite mixture model with Gibbs prior (b).*

6. Concluding Remarks

The suggested semiparametric approach allows to model both heterogeneity sources and spatial effects and represents a natural extension of usual finite mixture models for statistical mapping of spatial observations. The approach is computationally feasible and provides reliable estimates also when spatial dependence is weak, as in the working example. Thus, it could represent an interesting tool to assess geographical clusters which is less computationally intensive than proper Bayes approaches.

References

Aitkin, M. (1999). A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, **55**, 117–128.

Alfò, M. and Vitiello, C. (2003). Finite mixture approach to ecological regression, *Statistical Methods and Applications*, **12**, 93-108.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, **43**, 1–59.

Biggeri, A., Dreassi, E., Lagazio, C., and Böhning, D. (2003). A transitional non-parametric maximum pseudo-likelihood estimator for disease mapping. *Computational Statistics and Data Analysis*, **41**, 617–629.

Bindman, A., Grumbach, K., Osmond, D., et al. (1995). Preventable hospitalizations and access to health care. *Journal of the American Medical Association*, **274**, 305–311.

Böhning, D., Schlattmann, P., and Lindsay, B. (1992). Computer assisted analysis of mixtures (c.a.man): Statistical algorithms. *Biometrics*, **48,** 283–303.

Breslow, N. and Clayton, D. (1993). Approximate inference in generalized linear models. *Journal of the American Statistical Association*, 88, 9–24.

Clayton, D. and Kaldor, J. (1987). Empirical Bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43, 671–681.

Fiscella, K., Franks, P., Gold, M., and et al. (2000). Inequality in quality. addressing socioeconomic, racial and ethnic disparities in health care. *Journal of the American Medical Association*, **283**, 2579–2584.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721 – 741.

Green, P.J. and Richardson, S. (2002). Hidden Markov Models and Disease Mapping, *Journal of the American Statistical Association*, **97**, 1-16.

Griffiths, C., Sturdy, P., Naish, J., and et al. (1997). Hospital admissions for asthma in east London: associations with characteristics of local general practices, prescribing, and population. *British Medical Journal*, 314, 482–486.

Knorr-Held, L. and Besag, J. (1998). Modelling risk from a disease in time and space. *Statistics in Medicine*, **17**, 2045–2060.

Langford, I., Leyland, A. H., Rasbash, J., and Goldstein, H. (1999). Multilevel modelling of the geographical distribution of diseases. *Journal of the Royal Statistical Society C*, **48**:253–268.

Molliè, A. (1996). Bayesian mapping of disease. In Gilks, W., Richardson, S., and Spiegelhalter, D., editors, *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London.

Parchman, M. and Culler, S. (1999). Preventable hospitalizations in primary care shortage areas. an analysis of vulnerable medicare beneficiaries. *Archives of Family Medicine*, **8**, 487–491.

Schlattmann, P. and Böhning, D. (1993). Mixture models and disease mapping. *Statistics in Medicine*, 12:943–950.