Bias correction for underreported data in small area mapping.

Serena Arima, Loreto Gesualdo, Giuseppe Pasculli, Francesco Pesce, Silvia Polettini and Deni Aldo Procaccini.

Abstract Data quality is emerging as an essential characteristics of all data driven processes. The problem is particularly severe when health or vital statistics are concerned, with important consequences on government intervention policies and distribution of financial resources. In this paper, we deal with the underreporting issue with particular attention on its effects on the estimation of the prevalence of a phenomenon. We propose a non parametric compound Poisson model that allows for the estimation of underreporting probabilities. We will apply the proposed model to original data about the incidence of Chronic Kidney Disease (CKD) in Apulia.

Keywords: Underreporting probability disease mapping, non parametric model, MCMC.

1 Introduction

Data quality is an essential prerequisite for taking appropriate data driven decisions. As experienced in the last year, inaccurate data collection leads to inappropriate conclusions even when accurate and complex statistical methodologies had been performed. The problem is particularly severe when health or vital statistics are concerned, with important consequences on government intervention policies and distribution of financial resources. For example, the area-specific prevalence of a particular disease is the first criterion considered for distributing financial resources to hospitals and health devices. However, most often the number of individuals affected by the disease is inferred from patient registers, usually compiled upon registration by the health services, e.g. in hospitals when the medical examination occurs. A similar argument applies when it is of interest the geographical distribution

Serena Arima
Department of history, society and social sciences, University of Salento, Lecce, Italy, e-mail: serena.arima@unisalento.it

of different crimes or the regional distribution of forest fires. In all these cases, data are affected by underreporting and estimation of the prevalence of the phenomenon under study is substantially biased. The problem of underreporting is well known in the literature. [2] propose a bias correction method based on a compound Poisson model for count data that includes an area-specific reporting probability, whose uncertainty is accounted for in the model. In the proposal, areas are clustered according to their data quality. Since underreporting probabilities might reflect socio-economic, political and/or demographic characteristics of each region, we focus on modelling the underreporting rates in different areas. Extending the idea in [2], we consider the compound Poisson model. Following the approach described in [4], we propose to introduce covariates to define the clustering structure for the underreporting probability ϵ_i.

We will apply the proposed method to unpublished data coming from a retrospective study conducted between the 1st of January 2011 and the 31st of December 2013 for evaluating the incidence of Chronic Kidney Disease (CKD) in Apulia. To our knowledge, no prior studies have investigated the underreporting issue with respect to CKD prevalence in Italy and specifically in Apulia. But since the seminal paper by [3], it is clear that the availability of medical care tends to vary inversely with the need for it in the population served. Hence the need to further investigate possible underreporting in Apulia also considering that it is one of the most deprived regions in Italy. Moreover, Apulia is also characterized by very different geographical as well social conditions that we will consider in accounting for underreporting in CKD disease mapping.

2 Modelling underreported data

Consider a region consisting of m areas and denote by Y_i the observed counts in area i ($i = 1, ..., m$). Let E_i denote a known offset representing the expected number of events in the $i-$th area. The observed counts are modeled as a compound Poisson model (CPM)

$$Y_i|\theta_i\epsilon_i \sim Poisson(E_i\theta_i\epsilon_i)$$

and the relative risks are related to a set of covariates $X_1, ..., X_p$:

$$log(\theta_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}$$

The parameter ϵ_i defines the reporting probability in the $i-$th area: low values of ϵ_i indicate areas whose observed counts are underreported. As in [2], we assume that areas can be clustered according to their data quality. Moreover, we assume that the reporting probabilities are equal for areas where the covariates related to the reporting process take similar values. In the aforementioned paper, this goal is achieved by using a-priori information that induces a clustering structure among the areas: in particular, they fix the number of cluster and model the probability of underreporting according to a-priori knowledge of data quality. Very informative priors are defined

especially for the areas that are supposed to be characterized by the best data quality. In this work we consider an alternative approach: we specify a clustering structure for the underreporting probability ϵ_i following a non parametric approach based on a dependent Dirichlet process, that allows the aggregating property of the DP to depend on covariates. Although such a specification is more complex from a theoretical as well as computational point of view, it significantly increases the flexibility of the model since it does not require a-priori knowledge of the number of clusters and it defines the clustering structure in a complete nonparametric way. Indeed, the clustering is induced by introducing covariates in the stick breaking construction of the Dirichlet process.

2.1 The proposed model

Let $\epsilon^n = (\epsilon_1, ..., \epsilon_m)$ and $Z^n = (z_1, ..., z_m)$ denote, respectively, the entire vector of the underreporting probabilities and the covariate Z used as predictor for ϵ.

A simple nonparametric model can be defined by introducing a DP model on the underreporting probabilities:

$$p(y_i|\theta_i, \epsilon_i) = \prod_{j=1}^{k_n} e^{-(E_i\theta_i\epsilon_j)} \frac{(E_i\theta_i\epsilon_j)^{y_i}}{y_i!} \quad (1)$$

$$log(\theta_i) = \beta_i + \beta_1 X_{1i} + ... + \beta_p X_{pi} \quad (2)$$

$$\epsilon_i \sim iid\ G \quad (3)$$

$$G \sim DP(\alpha, G_0) \quad (4)$$

For any measurable set B, the DP process has the well known stick-breaking representation [5]

$$G(B) = \sum_{j=1}^{\infty} w_j \delta_{\eta_j}(B)$$

where $\delta_{\eta_j}(\cdot)$ is the Dirac measure at η_j and $w_j = V_j \prod_{l<j}[1 - V_l]$ with $V_j|\alpha \overset{i.i.d.}{\sim} Beta(1, \alpha)$

[1] propose a modification of the well known stick-breaking representation of the DP in which the weights are made dependent on covariates, this is achieved replacing the Beta random variables by normally distributed random variables transformed through the normal cdf. The resulting measure is defined as the probit-stick breaking (PSB) process, see also []. As described by [4], [1] allow for dependence on covariates via the introduction of independent Gaussian processes indexed by the covariates as specified in the following formula:

$$G_z(\cdot) = \sum_{j=1}^{\infty} \left\{ \Phi(\eta_j(z)) \prod_{l<j} [1 - \Phi(\eta_l(z))] \right\} \delta_{\epsilon_j}(\cdot)$$

where $\eta_j(z) = z'\gamma$,

The proposed model will be applied to the data described in Section 1 and results will be presented during the conference.

References

1. Chung, Y. and Dunson, D. B. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association* 104,1646–1660
2. de Oliverira, G.L., Argiento, R., Loschi, R.H. (2020) Bias correction in clustered underreported data, *Bayesian Analysis*, TBA, 1–32.
3. Hart,J.T. (1971) The inverse care law, *Lancet*, (7696): 405–412
4. Quintana, F., Mueller, P., Jara, A., MacEachern, S. (2021) The dependent Dirichlet process and related models. arXiv:2007.06129
5. Sethuraman, J. (1994). A constructive definition of Dirichlet prior. *Statistica Sinica*, 2, 639–650.