

Material Deprivation and Incidence of Lung Cancer: a Census Block Analysis.

Laura Grisotto, Dolores Catelan and Annibale Biggeri¹

Abstract We study the relationship between incidence for lung cancer in males in the Tuscan Region and material deprivation defined at census block level. We developed a bi-variate hierarchical Bayesian model to assess completeness of registration of incidence data and we proposed a series of random effect hierarchical Bayesian models to estimate the degree of association with material deprivation. Model comparison is addressed by a modified Deviance Information Criterion. We estimated a percent increase in risk of lung cancer for an increase of one standard deviation of material deprivation at census block level of 3.36% vs 5.87% at municipality level. The random slope models reported a paradoxically negative effect of material deprivation at census block level in some areas. Spatially structured random intercept models behaved better and random slope models were penalized by their extra complexity.

To our knowledge this is the first Italian study which analyzed cancer incidence at regional level.

Key words Spatial data analysis, Random effect, Hierarchical models; Lung cancer, Material deprivation.

¹ Laura Grisotto, Department of Statistics “G.Parenti”, University of Florence, Viale Morgagni, 59, 50134 Florence, Italy; Biostatistics Unit, ISPO Cancer Prevention and Research Institute, Via Cosimo il Vecchio, 2, 50139 Florence, Italy; email: grisotto@ds.unifi.it

Dolores Catelan, Department of Statistics “G.Parenti”, University of Florence, Viale Morgagni, 59, 50134 Florence, Italy; Biostatistics Unit, ISPO Cancer Prevention and Research Institute, Via Cosimo il Vecchio, 2, 50139 Florence, Italy; email: catelan@ds.unifi.it

Annibale Biggeri, Department of Statistics “G.Parenti”, University of Florence, Viale Morgagni, 59, 50134 Florence, Italy; Biostatistics Unit, ISPO Cancer Prevention and Research Institute, Via Cosimo il Vecchio, 2, 50139 Florence, Italy; email: abiggeri@ds.unifi.it

1 Introduction

The relationship between socioeconomic factors and health has been studied in many circumstances, mainly on aggregated data. A strong association with mortality for lung cancer is reported in Van der Heyden et al. (2009).

In the epidemiological literature material deprivation indexes defined at census block level have been used as a proxy of individual socio-economic status (e.g. Krieger et al., 2002, and in Italy Michelozzi et al., 1999). In fact census block is considered a small enough aggregation level to assume that individuals be homogeneously exposed and then to avoid the problem of ecological bias (Morgenstern, 2008).

The aim of this work is to study the relationship between incidence for lung cancer in males in the Tuscan Region and material deprivation defined at the census block level.

We take advantage of individual incidence data made available by the Tuscan Cancer Registry for the year 2004.

We developed a hierarchical Bayesian model to assess completeness of registration of incidence data and we proposed a series of random effect hierarchical Bayesian models to estimate the association with material deprivation. We also addressed model selection issue.

2 Data

2.1 *Incidence and Mortality data*

Incidence data come from the Tuscan Cancer Registry (Regione Toscana, 2005). Individual records report information on disease code and demographic characteristic of the person including the address at diagnosis. For the present analysis we considered the incident cases of lung cancer in male in the year 2004 for the whole Region (a total of 2,097 patient records).

Through the Iter.net system (Regione Toscana, 2006) 1,545 cases have been linked to census block indicator, because the Arezzo Province was not covered by the system. The corresponding number of analyzed municipalities is 248 (over the whole set of 287 Tuscan municipalities).

The expected number of cases for each census block was computed under indirect standardization applying a set of age-specific (16 age classes, 0-4,...,75 or more) reference rates (Tuscany 2004) to the population of each area.

To assess completeness of registration ISTAT death certificates of lung cancer in males in the period 2003-2005 for the 248 municipalities for which incidence data is available were recovered from the Regional Mortality Register.

2.2 *Material deprivation data*

Data on socio-economic factors derived from individual records of the 2001 census made available by the Tuscan Longitudinal Study (SLTo) (Regione Toscana, 1999; Accetta et al., 2007). Four variables are used as indicator of socio-economic inequalities:

- unemployment,
- low education (less than 6 years of schooling),
- house tenure,
- overcrowding.

A material deprivation index at census block level has been computed as the sum of z-scores (Caranci et al., 2010).

3 Methods

3.1 *Bayesian Analysis of Completeness of Registration*

To check for completeness of registration we conducted a Bayesian analysis on the mortality/incidence ratio (M/I) (Parkin et al., 1994). Let event counts Y_{lk} be distributed as $Poisson(\theta_{lk} \times Pop_k)$, where θ_{lk} is the rate parameter, Pop_k person years at risk, $l=1,2$ indexes death or incidence event respectively, and k indexes areas. A log linear bivariate model is specified on the rate parameter:

$$\log \theta_{lk} = \alpha + \mu_k + \delta_k I_l$$

where α is the intercept, μ_k is an area-specific random term, I_l and indicator variable which is 1 for mortality and 0 for incidence, δ_k is the parameter for the M/I ratio. We compared different prior specifications for the μ_k and δ_k parameters: spatially unstructured (heterogeneity) Gaussian($0, \sigma^2$) (with uninformative Gamma hyperprior for σ^2) or spatially structured (clustering) Gaussian conditional autoregressive $c_k | c_{v \sim k} \sim Normal(\bar{c}_k, \lambda_c n_k)$ where $v \sim k$ denotes areas adjacent to the k -th one, $\bar{c}_k$ is the mean over the adjacencies and n_k their number (Besag, 1974).

The expected M/I ratio is one. In fact, when the median survival time is very short as in the case of lung cancer, the number of incident cases will be almost equal to the number of deaths. A value of the M/I ratio greater than one suggests incompleteness of registration of incident cases; a M/I ratio smaller than one suggests misclassification of prevalent cases as incident cases because of poor definition of the date of diagnosis (Parkin et al., 1994).

3.2 Fixed-effect Models

Let x_{jk} be the material deprivation index for the j -th ($j=1, \dots, 23182$) census block in the k -th ($k=1, \dots, 248$) municipality and $\bar{x}_k$ the mean deprivation for the k -th municipality.

Let Y_{jk} , the number of incident cases, follows a Poisson distribution with mean $\theta_{jk} x_{jk}$ where θ_{jk} is the Relative Risk and E_{jk} the expected number of cases under indirect standardization. We specified a series of log linear models for θ_{jk} :

3.2.1) ecological model at census block level: $\log \theta_{jk} = {}^0\beta + {}^1\beta x_{jk}$, where ${}^1\beta$ is the ecological effect of the census block material deprivation value;

3.2.2) ecological model at municipality level: $\log \theta_{jk} = {}^0\beta + {}^2\beta \bar{x}_k$, where ${}^2\beta$ is the ecological effect of the average material deprivation at municipality level;

3.2.3) Cronbach model: $\log \theta_{jk} = {}^0\beta + {}^{1*}\beta (x_{jk} - \bar{x}_k) + {}^2\beta (\bar{x}_k - \bar{x})$, where ${}^{1*}\beta$ is the effect of the average census block material deprivation within municipality and ${}^2\beta$ is the ecological effect of the average material deprivation at municipality level (Cronbach and Webb, 1975).

3.3 Random-effect Models

The previous models do not consider random intercept, i.e. variability of the baseline risk among areas, or random slope. We defined:

3.3.1) a random intercept model: $\log \theta_{jk} = {}^0\beta_k + {}^{1*}\beta (x_{jk} - \bar{x}_k) + {}^2\beta (\bar{x}_k - \bar{x})$ where ${}^0\beta_k \sim \text{Normal}(\mu_{{}^0\beta}, \sigma^2_{{}^0\beta})$

3.3.2) a random intercept and slope model: $\log \theta_{jk} = {}^0\beta_k + {}^{1*}\beta_k (x_{jk} - \bar{x}_k) + {}^2\beta (\bar{x}_k - \bar{x})$ where $({}^0\beta_k, {}^{1*}\beta_k) \sim \text{MVNorm}(\boldsymbol{\gamma}, \mathbf{T}^{-1})$, $\boldsymbol{\gamma} \sim \text{MVNorm}(\mathbf{a}_\gamma, \mathbf{b}_\gamma)$ and $\mathbf{T} \sim \text{Wishart}(\mathbf{a}_T, \mathbf{b}_T)$.

3.3.3) a spatially structured random intercept model: we used the conditional autoregressive specification $\log \theta_{jk} = {}^0\beta_k + {}^{1*}\beta (x_{jk} - \bar{x}_k) + {}^2\beta (\bar{x}_k - \bar{x})$ where ${}^0\beta_k | {}^0\beta_{v \sim k} \sim N({}^0\bar{\beta}_k, \lambda_{{}^0\beta} n_k)$, where $v \sim k$ indicates the adjacent areas to the k -th municipality, ${}^0\bar{\beta}_k$ is the mean over the adjacencies, n_k their number, and $\lambda_{{}^0\beta} \sim \text{IGamma}(a_{{}^0\beta}, b_{{}^0\beta})$ is the precision parameter.

3.3.4) a spatially structured random intercept and a random slope model: as before we used a conditional autoregressive specification for the intercept and a spatially unstructured prior for the random slope $\log \theta_{jk} = {}^0\beta_k + {}^{1*}\beta_k (x_{jk} - \bar{x}_k) + {}^2\beta (\bar{x}_k - \bar{x})$ where ${}^0\beta_k | {}^0\beta_{v \sim k} \sim N({}^0\bar{\beta}_k, \lambda_{{}^0\beta} n_k)$ and ${}^{1*}\beta_k \sim \text{Normal}(\mu_{{}^{1*}\beta}, \sigma^2_{{}^{1*}\beta})$.

Non informative distribution specifications were assumed for all the hyperparameters.

3.4 Model selection

Comparison of several competing hierarchical Bayesian models is a difficult task and requires special consideration. When comparing models which involve random terms of growing complexity usually the Deviance Information Criterion (DIC) is used (Spiegelhalter et al., 2002). DIC is defined as the sum of two components $DIC = \bar{D} + p_D$; $\bar{D}$ is the posterior deviance expectation and summarizes the fit of the model, p_D is the expected deviance minus the deviance evaluated at the posterior expectations. It represents the “effective number” of model parameters and reflects the complexity of the model.

Plummer (2008) discussed the validity of DIC when the number of independent observation is less than the number of parameters (as common in geographical studies). He proposed to adjust for the “degree of optimism” present in the DIC statistics when evaluating model complexity. A penalized loss function based on cross-validation or a simpler modified version of the DIC was suggested. Here we used the modified DIC.

In particular:

$$DICc = \bar{D} + r_{opt}$$

$$r_{opt} \approx \sum_{j=1}^J p_{Dj} / (1 - p_{Dj})$$

where j indexes census block as before, and r_{opt} denotes the residual optimism.

4 Results

Figure 1a shows the spatial distribution of lung cancer standardized incidence ratios (SIR) at municipality level for the Tuscan region (excluding the Province of Arezzo) and Figure 1b shows the spatial distribution of material deprivation index at municipality level. Both maps are highly spatially structured and suggest a positive association between lung cancer incidence and material deprivation.

4.1 Completeness of Cancer Registration

Before addressing the analysis of the association between socio-economic factors and lung cancer incidence we must evaluate potential flaws in cancer registration which could bias the results. The mortality/incidence ratio for each municipality was estimated by means of a series of bi-variate hierarchical Bayesian models. Figure 2a shows the predicted M/I ratios by the best fitting model with spatially structured terms. The distribution is slightly skewed. A slightly incompleteness of registration is seen in the north-west part of the region and a misclassification of prevalent cases in the south-eastern areas. However we cannot exclude selective survival by covariates such as socio-economic factors. Eventually the observed M / I ratio may correlate with the material deprivation index.

4.2 Deprivation and Incidence of Lung Cancer

The estimate of the rate ratio (RR) for a unit increase of material deprivation at municipality level on incidence of lung cancer ranged from $RR = 1.16$ to $RR = 1.18$ depending on the hierarchical Bayesian random effect model considered. To compare the estimated rate ratios for material deprivation at census level to those at municipality level we express the rate ratio for an increase of one standard deviation and we report here for brevity only the best fitting model results. The percent increase in risk for an increase of one standard deviation of material deprivation at census block level was 3.36% vs 5.87% at municipality level. Notice that introducing a spatially structured random intercept deflated the risk estimates from those under the fixed-effect Cronbach model (3.87% and 6.72% respectively). The random slope models reported a paradoxically negative effect of material deprivation at census block level in some areas (data not shown). Table 1 reports model comparison statistics. Despite a poorer fit the spatially structured random intercept model behaves better in term of DICc. Random slope models are penalized by their extra complexity.

Figure 2b shows the spatial distribution of random intercepts from the best fitting model. A lower risk in the north-western part of the region might be spurious as highlighted by the predicted distribution of the M/I ratio.

Figure 1:a) Spatial distribution of the Standardized Incidence Ratio. Lung cancer. Males, Tuscany (without Arezzo). Tuscany Cancer Registry, 2004.

b) Spatial distribution of Material Deprivation index. Tuscany municipalities (without Arezzo), ISTAT Census 2001.

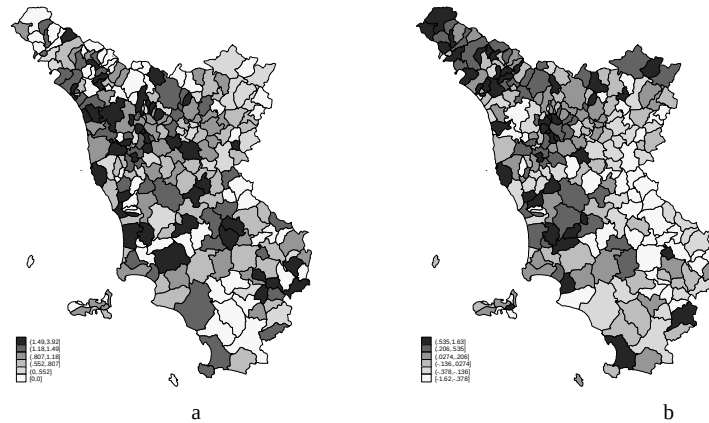
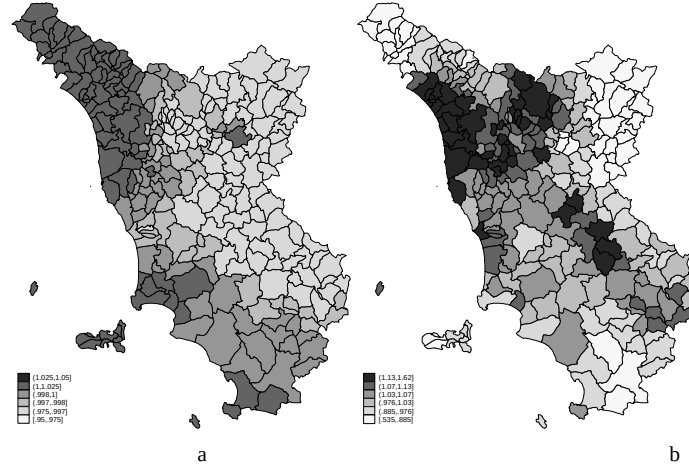


Figure 2a: Spatial distribution of the ratio M / I estimated.**2b:** Spatial distribution of baseline relative risks from the spatially structured random intercept.

Males, Tuscany (without Arezzo), Tuscany Cancer Registry, 2004.

**Table 1:** Indices of goodness of fit: $\bar{D}$, DIC, p_D , corrected DIC (DICc), r_{opt} , EDP.

<i>Model</i>	$\bar{D}$	DIC	p_D	DICc	r_{opt}	EDP
Random intercept	4578	4670	92	4666	88	8469
Random intercept and slope	4580	4680	100	4684	104	8466
Spatially structured random intercept	4583	4646	63	4648	65	8471
Spatially structured random intercept and random slope	4570	4663	93	4667	97	8471

5 Discussion and Conclusions

In Italy there are few studies that compare results from analysis at the individual and the census block level (Biggeri et al., 2004). In the present paper we assumed that census block information may be used as a surrogate for the individual level. However since a contextual effect is plausible we expect the effect estimates be larger than those obtained having the complete information.

The results showed that municipality level has a larger impact than census block level even when we adjust for baseline risk variability. We must be cautious in interpreting this finding, since the ecological bias may menace the validity of this comparison.

Bayesian approach allows flexible random effect modelling. Future work will address a joint modelling of completeness of registration and disease risk.

To our knowledge this is the first Italian study which analyzed cancer incidence at regional level.

Acknowledgements: Italian Ministry of Education, University and Scientific Research Project PRIN20072S2HT8.

References

1. Accetta, G., Grisotto, L., Terni, G., Biggeri, A.: Mortalità per condizione socio-economica. Studio Longitudinale Toscana 2001-2005. Informazioni Statistiche, Studi e Ricerche. Edizioni Regione Toscana. 1-116 (2007)
2. Besag, J.: Spatial interaction and the statistical analysis of lattice systems. *JRSS, series B* **36**, 192-236 (1974)
3. Biggeri, A., Dreassi, E., Marchi, M.: A Multilevel Bayesian Model for Contextual Effect of Material Deprivation. *Statistical Methods & Applications*. **13**, 88-102 (2004)
4. Caranci, N., Biggeri, A., Grisotto, L., Pacelli, B., Spadea, T., Costa, G.: L'indice di deprivazione italiano a livello di sezione di censimento: definizione, descrizione e associazione con la mortalità. *Epidemiologia e Prevenzione*. (2010) (to appear)
5. Cronbach, L.J., Webb, N.: Between class and Within class Effects in a Reported Aptitude $\times$ Treatment Interaction: A reanalysis of a study by G.L. Anderson. *Journal of Educational Psychology*. **67**, 717-724 (1975)
6. Krieger, N., Chen, J.T., Waterman, P.D., Soobader, M.J., Subramanian, S.V., Carson, R.: Geocoding and monitoring of US socioeconomic inequalities in mortality and cancer incidence: does the choice of area-based measure and geographic level matter?: the Public Health Disparities Geocoding Project. *Am J Epidemiol*. **156**, 471-82 (2002)
7. Michelozi, P., Perucci, C.A., Forastiere, F., Fusco, D., Ancona, C., Dell'Orco, V.: Inequality in health: socioeconomic differentials in mortality in Rome, 1990-95. *J Epidemiol Community Health*. **53**, 687-93 (1999)
8. Morgenstern, H.: Ecologic Studies. In *Modern Epidemiology*, 3rd Edition, Rothman KJ et al. (eds) Boston: Lippincott, Williams, & Wilkins. (2008)
9. Parkin, D.M., Chen, V.W., Ferlay, J., Galceran, J., Storm, H., Whelan, S.: Comparability and Quality Control in Cancer Registration. IARC Technical Report No. 19. Lyon: IARC. (1994)
10. Plummer, M.: Penalized loss functions for Bayesian model comparison. *Biostatistics*. **9**, 523-539 (2008)
11. Regione Toscana.: Studio Longitudinale Toscana. DGR 423 19-04-1999. (1999)
12. Regione Toscana.: Registro Tumori della Regione Toscana. DGR 1200 12-12-2005. (2005) <http://www.ittumori.it/ITA/chisiamo/registro-tumori.shtml>
13. Regione Toscana. Progetto ITER-NET sistema regionale delle strade e indirizzi. DGR 44 30-01-2006. (2006) <http://www.iternet.fi.eng.it/SigmaPortal2>
14. Van der Heyden, J.H., Schaap, M.M., Kunst, A.E., Esnaola, S., Borrell, C., Cox, B., Leinsalu, M., Stirbu, I., Kalediene, R., Deboosere, P., Mackenbach, J.P., Van Oyen, H.: Socioeconomic inequalities in lung cancer mortality in 16 European populations. *Lung Cancer*. **63**, 322-30 (2009)
15. Spiegelhalter, D.J., Best, N., Calrlin, B.P., van der Linde, A.: Bayesian measures of model complexity an fit (with discussion). *Journal of the Royal Statistical Society B*. **64**, 583-639 (2002)