Dipartimento Jonico in Sistemi Giuridici ed Economici del Mediterraneo: Società, Ambiente, Culture

Jonian Department - Mediterranean Economic and Legal Systems: Society, Environment, Cultures

ANNALI 2015 – ANNO III

(ESTRATTO)

MASSIMO BILANCIA - GIUSI GRAZIANO

Measuring geographical disparity in lung cancer mortality in Apulia, Italy. Results for the 2002-2009 period

http://www.annalidipartimentojonico.org **ISBN:** 978-88-909569-4-2

Redazione:
Prof. Francesco Mastroberti
Dipartimento Jonico in Sistemi Economici e Giuridici del Mediterraneo: Società, Ambiente, Culture
Convento San Francesco, Via Duomo, 259 - 74123 Taranto, Italy
E-mail: francesco.mastroberti@uniba.it
Telefono: + 39 099 372382
Fax: + 39 099 7340595
http://www.annalidipartimentojonico.org

Massimo Bilancia[†], Giusi Graziano[††]

MEASURING GEOGRAPHICAL DISPARITY IN LUNG CANCER MORTALITY IN APULIA, ITALY. RESULTS FOR THE 2002-2009 PERIOD[*]

ABSTRACT	
This paper reports on the presence of an excess mortality for lung cancer among resident males in the southeast part of the Salento peninsula. This increasead level of risk was first documented in a report of the Registro Tumori Puglia (formerly known as Registro Tumori Jonico Salentino), and further analyzed by one of the authors of this paper, using mortality data for the 1992-2001 period. Here, we present an updated analysis conducted using data over two distinct time intervals (2002-2005 and 2006-2009). The obtained results show that the excess mortality is still present, and that the causes of such geographical disparity remain somewhat obscure.	Il presente lavoro mira a documentare la presenza di un eccesso di mortalità per tumore al polmone tra i maschi residenti nella parte sud-orientale della penisola salentina. L'aumento del livello di rischio per tale patologia era già stato individuato in un report del Registro Tumori Puglia (già noto come registro Tumori Jonico Salentino), ed ulteriormente analizzato da uno degli autori di questa pubblicazione, utilizzando dati di mortalità per il periodo 1992-2001. In questa sede presenteremo una analisi aggiornata basata su due distinti intervalli temporali (2002-2005 e 2006-2009). I risultati ottenuti dimostrano che l'eccesso di mortalità è tuttora presente, e che le cause di una siffatta disparità geografica rimangono ancora oscure.
Spatial epidemiology – Bayesian hierarchical modeling – Lung cancer mortality	**Epidemiologia spaziale – Modellistica bayesiana gerarchica – Mortalità per tumore al polmone**

[†] Correspondence: Ionian Department of Law, Economics and Environment, Università degli Studi di Bari Aldo Moro, Via Lago Maggiore angolo Via Ancona - 74121 Taranto – Italy. Email: massimo.bilancia@uniba.it.

[††] Correspondence: Oncologic Hospital Giovanni Paolo II, Viale Orazio Flacco 65 – 70124 Bari – Italy. Email: giusi.graziano78@gmail.com.

[*] This paper has been reviewed under double-blind peer review.

1. - Despite many nations have seen a general improvement in cancer survival, and a decline in incidence and mortality rates for some cancers, notable inequalities in these outcomes still persist between people of differing race, ethnicity, socioeconomic status, and area of residence, as witnessed by many researchers and numerous international studies[1].

Of course, inequalities need to be quantified before they can be addressed, and since the mid-1800s maps have been used to provide a visual representation of cancer outcomes. Cancer mortality and incidence maps often offer clues of the many different forms a cancer inequality can take[2]. For example, cancer patients living in disadvantaged areas are more likely to be diagnosed with advanced cancer and to have poorer survival. Similarly, patients living in rural areas with a reduced access to health care services, and lower socioeconomic groups living in deprived areas have a higher prevalence of cancer risk factors such as smoking, obesity and low level of physical activity.

When the Apulia region is taken into consideration, it must be stressed that public health investigations are dominated by the analyses of relationships existing between environment and cancer mortality/incidence in two contaminated sites. The first site of interest is the city of Taranto, because of several polluting sources, such as a large steel plant, a refinery, a harbor and illegal waste dumps[3]. The second is the Brindisi area, characterized by the presence of industries with high environmental impact, located along its eastern border. Several epidemiological studies have revealed critical situations[4].

However, the purpose of this paper is to document a different phenomenon, i.e. the presence of an excess mortality for lung cancer among resident males in the southeast part of the Salento peninsula. This increasead level of risk was first documented in a report of the Registro Tumori Puglia (formerly known as Registro Tumori Jonico Salentino), and further analyzed by one of the authors of this paper, using mortality data for the 1992-2001 period[5]. Here, we present an updated analysis

[1] A. WEINBERG, P.M. JACKSON, C. DECOURTNEY, K. CRAVATT, J. OGO, M.M. SANCHEZ, G. TORTOLERO-LUNAS, R.L. ROLLINS, *Progress in addressing disparities through comprehensive cancer control*, in "Cancer Causes & Control", 21:12 (2010), pp. 2015–2021.

[2] S.M. CRAMB, K.L. MENGERSEN, P.D. BAADE, *Developing the atlas of cancer in Queensland: methodological issues*, in "International Journal of Health Geographics", 10:9 (2011), doi:10.1186/1476-072X-10-9.

[3] R. PIRASTU, P. COMBA, I. IAVARONE, A. ZONA, S. CONTI, G. MINELLI, V. MANNO, A. MINCUZZI, S. MINERBA, F. FORASTIERE, F. MATALONI, A. BIGGERI, *Environment and Health in Contaminated Sites: The Case of Taranto, Italy*, in "Journal of Environmental and Public Health", Article ID 753719 (2013), doi:10.1155/2013/753719.

[4] C. MANGIA, A. BRUNI, M. CERVINO, E.L. GIANICOLO, *Sixteen-year air quality data analysis of a high environmental risk area in Southern Italy*, in "Environmental Monitoring and Assessment", 183:1-4 (2011), pp. 555–570.

[5] M. BILANCIA, A. FEDESPINA, *Geographical clustering of lung cancer in the province of Lecce, Italy: 1992-2001*, in "International Journal of Health Geographics", 8:40 (2009). doi:10.1186/1476-072X-8-40.

conducted using data over two distinct time intervals (2002-2005 and 2006-2009). The obtained results show that the excess mortality is still present, and that the causes of such geographical disparity remain somewhat obscure.

The paper is organized as follows. Section 2 introduces the necessary notation and the basic Poisson model when the outcome of interest is the area-level count of an adverse health event. In Section 3 the basic concepts of Bayesian hierarchical modeling are introduced, and the spatio-temporal model used to analyze the data is shortly explained. In Section 4, some mathematical difficulties associated with priors for the spatial distribution of relative risks are highlighted. Section 5 contains a brief introduction to local cluster detection using the Kulldorff's circular scan statistic. Section 6 contains a pointer to the numerical facilities used to map the relative risk estimates. Section 7 contains results and maps depicting the geographical variation in risk. The paper ends in Section 8 with some clues for future research.

2. - First, we introduce the necessary notation. Let Y_{it} denote the number of deaths observed within the i-th area ($i = 1,2,\ldots,N$) during the time window t ($t = 1,2,\ldots,T$). We can broadly assume that, independently in each area and time windows, death counts follow a Poisson distribution:

$$Y_{it} \overset{ind}{\sim} \text{Poisson}(N_{it} p_{it}), \tag{1}$$

where p_{it} represents the mortality rate within area i and time windows t and, in a similar vein, N_{it} is the amount of person-years at risk in each time interval considered. With the help of a suitable reference rate p, we are able to re-parametrize to area-specific relative risks $\theta_{it} = p_{it}/p$ such that:

$$N_{it} p_{it} = N_{it} p \theta_{it} = E_{it} \theta_{it}, \tag{2}$$

in which $N_{it} p = E_{it}$ represents the number of expected deaths under the reference rate p. When the reference population is the same as the population under study, we set the reference rate as:

$$p = \sum_{i,t} Y_{it} \Big/ \sum_{i,t} N_{it}. \tag{3}$$

This *internal standardization* centres the data with respect to the current maps. Therefore, the re-parametrized Poisson model assumes the form:

$$Y_{it} \overset{ind}{\sim} \text{Poisson}(E_{it} \theta_{it}), \tag{4}$$

in which the E_{it}s are (improperly) regarded as fixed, whereas Y_{it}s are regarded as random and conditionally indipendent of one another given the relative risks θ_{it}s. The obvious summaries of the Poisson model (4) are the maximum likelihood estimates (MLE) of area-specific relative risks $\hat{\theta}_{it} = SMR_{it} = Y_{it}/E_{it}$ (*SMR* = Standardized Mortality Ratio). The areas where there is an excess of risk are those in which the number of observed deaths exceeds the expected ones (and hence $SMR_{it} > 1$).

Estimates derived from the Poisson model are ready for mapping and GIS analyses, although *SMR*s tend to be unreliable estimators of the underlying risks, as $Var(\hat{\theta}_{it}) = \theta_{it}/E_{it}$. Since the expected counts depend on the underlying population at risk, the precision of the *SMR*s will vary inversely with the size of the population. As a consequence, any map suggesting outliers may be spurious, since the extreme values may result from a large degree of variability of the estimates. For example, sparsely populated areas can visually dominate the map, but provide the least reliable estimates.

3. - A way for overcoming the above mentioned issue consists in borrowing information from sources other than the observation at hand, in order to improve the properties of the relative risk estimates, achieving a better mean squared error. For example, under the Poisson model (4) we are not able to exploit the information that disease outcomes in spatial units are often not independent of each other. As a consequence, risk levels of areas that are close to each other tend to be positively correlated, as they share a number of spatially varying characteristics.

From this perspective, the most commonly used models are derived from the full Bayesian approach. This class of models allow for the augmentation of the Poisson likelihood function, in order to incorporate a spatially correlated structure of the relative risk, resulting in a Bayesian hierarchical model. This hierarchical structure promotes the concept of Bayesian smoothing, whereby the estimation of relative risk is based upon not only the data observed for the selected region, but from neighbouring regions as well. Following this approach, the log-relative risks of the Poisson likelihood are modeled as a suitable combination of fixed and random effects, leading to the specification[6,7,8]:

$$\eta_{it} = \log(\theta_{it}) = \alpha + \upsilon_i + \nu_i + (\beta + \delta_i) \times t \,. \qquad (5)$$

[6] L. BERNARDINELLI, D. CLAYTON, C. PASCUTTO, C. MONTOMOLI, M. GHISLANDI, M. SONGINI, *Bayesian analysis of space-time variation in disease risk*, in "Statistics in Medicine", 14:21-22 (1995), pp. 2433–2443.
[7] B. SCHRÖDLE, L. HELD, *A primer on disease mapping and ecological regression using INLA*, in "Computational Statistics", 26:2 (2011), pp. 241–258.
[8] N. WHITE, *Review of statistical methods for disease mapping*, 2012, http://eprints.qut.edu.au/56859/.

In this classical formulation, introduced to analyse the variation of a given disease in space and time, random effects υ_i allow for spatially structured risk patterns, being defined from the following set of conditional distributions:

$$\upsilon_i \left| \upsilon_j \sim \mathrm{N}\left(\frac{1}{m_i} \sum_{j:j\sim i} \upsilon_j, \frac{\sigma_\upsilon^2}{m_i} \right), \right. \tag{6}$$

where $\{ j : j \sim i \}$ is the set of first-order neighbours of area *i*, whereas m_i is the cardinality of this set. In the context of disease mapping, the prior specification (6) has been proposed as an efficient solution to model spatial dependencies[9], and it is commonly referred to *intrinsic conditional autoregression* (ICAR). On the contrary, random effects ν_i describe sources of extra-Poissonian variation that are not spatially structured (for example, data inaccuracies that inflate the marginal variance to a level greater than that expected under the standard Poisson model). A zero-mean exchangeable Gaussian prior is commonly used for unstructured effects:

$$\nu_i \overset{iid}{\sim} \mathrm{N}\left(0, \sigma_\nu^2\right). \tag{7}$$

The terms δ_i represent local deviations between the global time trend β and the area specific trend. When $\delta_i < 0$ the area-specific trend is steeper ?? than the global trend, whilst $\delta_i > 0$ implies that the area-specific trend is steeper than the global trend. We assume that random effects δ_i follow a Gaussian exchangeable prior with zero mean and variance σ_δ^2. Other non-parametric specifications are possible, such as a random-walk stochastic trend[10], but this approach was not considered here as only a short time interval is taken into account.

Finally, Bayesian hierarchical modeling envisages specification of the prior distribution of the random effect variance parameters $\sigma_\upsilon^2, \sigma_\nu^2$. These distributions are parametrized by hyperparameters which control the variability of the relative risks across the map. In Section 7, we will briefly discuss a few common alternatives to eliciting these priors, minimizing the risk of over-smoothing of the estimated relative risks.

4. - A few technical details concerning the prior distribution (6) are worth noting. A broad class of spatially structured priors is obtained by implicitly specifying a Gaussian Markov Random Field[11] (GMRF) with the following set of conditional distributions:

[9] J. BESAG, J. YORK, A. MOLLIÈ, *Bayesian image restoration, with two applications in spatial statistics*, in "Annals of the Institute of Statistical Mathematics" 43:1 (1991), pp. 1–20.
[10] SCHRÖDLE, HELD, *op. cit.*, p. 4
[11] H. RUE, L. HELD, *Gaussian Markov Random Fields: Theory and Applications* (Monographs on Statistics and Applied Probability,Vol. 104), 2005, London, Chapman & Hall.

$$\upsilon_i \left| \upsilon_j \sim \mathrm{N}\left(\sum_{j:j\sim i} w_{ij}\upsilon_j, \sigma_v^2 q_{ii}^{-1} \right), \right. \tag{8}$$

for $i = 1,2,\ldots,N$, and for some set of coefficients $\left\{ w_{ij} \text{ for } j \neq i, \text{and } w_{ii} = 0 \right\}$ whose nonzero terms implicity define the relation $\sim$ (which might be more general, in this context, than the first order neighbourliness). Of course, even though the conditional distributions (8) are well defined, we are not sure that a joint GMRF exists with the prescribed Markov properties. By using the Brook's expansion, it can be proved that the set of Gaussian conditionals (8) defines a GMRF for the vector $\upsilon = \left(\upsilon_1,\ldots,\upsilon_N\right)^T$ having the following joint specification:

$$\upsilon \sim N\left(0, \sigma_v^2 Q^{-1}\right), \tag{9}$$

where the precision matrix Q has diagonal elements q_{ii}, whereas $q_{ij} = q_{ii} w_{ij}$ for $i \neq j$, provided that Q is positive definite and that the compatibility conditions $q_{ii} w_{ij} = q_{jj} w_{ji}$ hold (to ensure that $Q^T = Q$). When these two conditions are satisfied, it is a simple task to verify that $Q = D^{-1}(I - W)$, with $W = \left\{ w_{ij} \right\}$ and $D = diag\left\{ q_{11}^{-1},\ldots,q_{NN}^{-1} \right\}$.

The ICAR prior (6) is obtained as a special case of the model defined above, by taking $w_{ij} = 1$ if areas *i* and *j* are neighbours, and $w_{ij} = 0$ otherwise (including the case that $w_{ii} = 0$). Under these circumstances $w_{ij} = m_i^{-1}$ and $q_{ii} = m_i$, finding again the conditional specification (6). However, simple inspection reveals that the corresponding precision matrix Q has rank $N - 1$ and so is not positive definite. Hence the ICAR model can be considered as a limiting form of the CAR model, in which σ_v^2 is only interpretable conditionally and no longer as a marginal variance, because the joint specification no longer exists. However, from a Bayesian viewpoint we can still write the joint version of the conditional specification (6) as an improper prior over the space of spatially structured effects:

$$\pi\left(\upsilon \left| \sigma_v^2 \right.\right) \propto \exp\left(-\frac{1}{2\sigma_v^2} \sum_{j:j\sim i} \left(\upsilon_i - \upsilon_j\right)^2 \right). \tag{10}$$

With this prior specification, Bayesian posterior learning may become problematic as the improper ICAR prior only informs about contrasts $\upsilon_i - \upsilon_j$ for $j \sim i$, but it does not identify the overall mean α for the log-relative risks. With an improper prior over α, it can be formally proved that the joint posterior distribution of model parameters is not integrable[12]. This result just reflects the fact that the baseline effect cannot be resolved neither in the likelihood nor in the prior structure

[12] M. GHOSH, K. NATARAJAN, L.A. WALLER, D. KIM, *Hierarchical Bayes GLMs for the analysis of spatial data: An application to disease mapping*, in "Journal of Statistical Planning and Inference", 75:2 (1999), pp. 305–318.

of the model. A common solution to generate a proper posterior is to identify the overall mean by adding the constraint:

$$\sum_{i=1}^{N} \upsilon_i = 0\,, \tag{11}$$

which results in a proper embedded posterior distribution over a constrained lower-dimensional parameter space[13]. Of course, for the same identifiability purpose a sum-to-zero constraint is imposed on random effects ν_i and δ_i as well.

5. - The assessment of spatial variation in risk is aimed at producing a global map of important spatial effects, while simultaneously removing any disturbing noise. On the other side, in studying local area clustering we are concerned with 'a geographically and/or temporally bounded group of occurrences of sufficiente size and concentration to be unlikely to have occurred by chance'[14].

While a wide range of methods have been proposed to test for local spatial clustering, the spatial scan statistic (introduced in a remarkable series of papers published during the mid '1990s [15,16]) is by far the most popular. A circular window is imposed on the map and its centre is allowed to move at any position, so that this circular window defines a potential cluster by including different sets of neighbouring areas. For practical purposes, the centre of each circular window is placed only at the small area centroids, and the radius is allowed to vary from zero to a maximum radius so that the window never includes more than a fixed percentage of the population at risk. A very large number of distinct circular windows will be created by this method, and each Z defined by a circle consists of all those areas whose centroids lie inside the circle.

Throughout this section study region will be denoted as G. For simplicity, we suppress the explicit time dependency as well. For a given window Z, we assume that events are generated by an inhomogeneous Poisson process having the following intensity function for all $x \in G$:

$$\lambda(x) = p_Z \mu(x) I_Z(x) + p_{\bar{Z}} \mu(x) I_{\bar{Z}}(x)\,, \tag{13}$$

where $\bar{Z} = G/Z$, $I_Z(\cdot)$ is the indicator function over the window Z, p_Z and $p_{\bar{Z}}$ indicate the probability that one individual at risk – living respectively inside or outside Z – has a given disease, and $\mu(x)$ models the spatial distribution of the

[13] A.E. GELFAND, S.K. SAHU, *Identifiability, Improper Priors, and Gibbs Sampling for Generalized Linear Models*, in "Journal of the American Statistical Association", 94:445 (1999), pp. 247–253.
[14] E.G. KNOX, *Detection of clusters*, in P. Elliot (Ed.), *Methodologies of Enquiry into Disease Clustering,* 1989, pp. 17–22, Small Area Health Statistics Unit, London.
[15] M. KULLDORFF, N. NAGARWALLA, Spatial disease clusters: Detection and inference, in "Statistics in Medicine", 14:8 (1995), pp. 799–810.
[16] M. KULLDORFF, *A spatial scan statistic*, in "Communications in Statistics - Theory and Methods", 26:6 (1997), pp.1481–1496.

population at risk over G. Under the Poisson likelihood implied by these premises, to test the alternative hypothesis of local raised incidence $H_1 : p_Z > p_{\bar{Z}}$ versus the null $H_0 : p_Z = p_{\bar{Z}}$ we can consider the following generalised likelihood-ratio statistic[17] for a specific circular window Z:

$$\Lambda_Z = \frac{\max_{p>q} L(Z, p_Z, p_{\bar{Z}})}{\max_{p=q} L(Z, p_Z, p_{\bar{Z}})} = \left(\frac{Y(Z)}{E(Z)}\right)^{Y(Z)} \left(\frac{Y(\bar{Z})}{E(\bar{Z})}\right)^{Y(\bar{Z})} I\left(\frac{Y(Z)}{E(Z)} > \frac{Y(\bar{Z})}{E(\bar{Z})}\right), \quad (14)$$

where $Y(Z)$ (resp.: $Y(\bar{Z})$) represents the random number of deaths occurred inside (resp.: outside) Z, and $E(Z)$ is the expected number of deaths inside Z under the reference rate p. The spatial scan statistic Λ is the maximum of the likelihood ratio (14) over all the candidate clusters Z:

$$\Lambda = \max_Z \Lambda_Z, \quad (15)$$

which identifies the most likely cluster (MLC), that is the least likely to have occurred by chance. The distribution of the test statistic (14) under H_0 has no simple closed form, even though an approximate p-value can be obtained using a Monte Carlo approach to evaluate MLC's statistical significance[18]. In the jargon of spatial statistics literature, a significant MLC is often referred to as a 'primary' cluster.

6. - The calculations reported in this paper were made possibile by the use of R 3.2.1[19]. In particular, Bayesian estimation of space-time model (5) was carried out by using the Integrated Nested Laplace Approximation (INLA[20,21]), which has been developed as a computationally efficient alternative to Markov Chain Monte Carlo (MCMC) numerical schemes. The INLA approach is implemented in the `R-INLA`[22] package, which substitutes a standalone program built around the `GMRFLib` library. For the interested reader, the web-site http://www.r-inla.org/ provides documentation and many worked examples. Finally, the Kulldorff method for finding the MLC was

[17] KULLDORFF, NAGARWALLA, *op. cit.*, p. 7

[18] Z. ZHANG, R. ASSUNÇÃO, M. KULLDORFF, *Spatial Scan Statistics Adjusted for Multiple Clusters*, in "Journal of Probability and Statistics", Article ID 642379 (2010), doi: 10.1155/2010/642379.

[19] R CORE TEAM, *R: A Language and Environment for Statistical Computing*, 2015, Vienna, Austria. Retrieved from http://www.r-project.org/

[20] H. RUE, S. MARTINO, N. CHOPIN, *Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations*, in "Journal of the Royal Statistical Society: Series B (Statistical Methodology)", 71:2 (2009), pp. 319–392.

[21] M. BLANGIARDO, M. CAMELETTI, G. BAIO, H. RUE, *Spatial and spatio-temporal models with R-INLA*, in "Spatial and Spatio-Temporal Epidemiology", 4 (2013), pp. 33–49.

[22] T.G. MARTINS, D. SIMPSON, F. LINDGREN, H. RUE, *Bayesian computing with INLA: New features*, in "Computational Statistics & Data Analysis", 67 (2013), pp. 68–83.

implemented using an ad hoc function available in the `SpatialEpi`[23] package. Source code is available from authors upon request.

7. - Updated mortality data used in this paper are distributed by the Apulian Epidemiological Regional Centre[24], and cover periods 2002-2005 ($t=1$) and 2006-2009 ($t=2$). We consider the number of deaths occurred among male residents in the 258 municipalities of Apulia, Italy, for malignant neoplasm of trachea, bronchus, and lung (ICD-IX: 162). The amounts of person-years at risk N_{it} were based on ISTAT resident population estimates, available as of January 1st of each year between 2002 and 2009[25]. The internally standardized reference rate (3) was $p=75.784$ deaths for 100,000 males at risk per year.

Following the methodological recommendations of NHCS Methodological Issues in Measuring Health Disparities[26], we compared disparities, across geographic areas and over time, converting any set of estimates of the relative risks $\tilde{\theta}_{it}$ to rates $\tilde{p}_{it}=p\tilde{\theta}_{it}$, using the global rate p as a reference point. Our relative disparity measure (*RD*) is defined as:

$$RD_{it}=\frac{\text{Rate of interest } (p\tilde{\theta}_{it})-\text{reference rate } (p)}{\text{reference rate } (p)}\bullet 100\%=\left(\tilde{\theta}_{it}-1\right)\bullet 100\%. \qquad (16)$$

Figure 1 shows map of geographical relative disparity based on the maximum likelihood estimates $\hat{\theta}_{it}=Y_{it}/E_{it}$, corresponding to area-specific SMRs of lung cancer among male residents. Interpretation is not easy due to the sampling variability that is inherent in such estimates (see also Figure 2, showing the large variability of the distribution of SMRs, for both time intervals considered here). However, despite the presence of this disturbing noise, it seems possibile to conclude that: 1) an increasing trend in the relative disparity in the North-South direction is clearly seen. A large region including both the southern part of the province of Brindisi and the southeast part of the Salento peninsula presents higher mean relative disparities, with many areas having a relative disparity greater than 200% of the reference rate. Some areas showing increased disparities are present along the border of the province of Foggia as well, although these values are obtained in small areas having a low population at

[23] C. CHEN, A. Y. KIM, M. ROSS, J. WAKEFIELD, *SpatialEpi: Methods and Data for Spatial Epidemiology*, 2014, R package version 1.2.1, http://CRAN.R-project.org/package=SpatialEpi.
[24] S. BARBUTI, M. QUARTO, C. GERMINARIO, R. PRATO, P. LOPALCO, E. COVIELLO, G. CAPUTI, D. MARTINELLI, S. TAFURI, M.T. BALDUCCI, L. LAMARINA, F. FORTUNATO, R. BERARDINO, A.M. ARBORE, M.D. PALMA, *Atlante delle Cause di Morte della Regione Puglia: Anni 2000–2005*, 2006, Regione Puglia - Osservatorio Epidemiologico della Regione Puglia.
[25] http://demo.istat.it/archivio.html
[26] K. KENNETH, E. PAMUK, J. LYNCH, O. CARTER-POKRAS, I. KIM, V. MAYS, J. PEARCY, V. SCHOENBACH, J.S. WEISSMANN, *Methodological issues in measuring health disparities*, 2005, U.S. Department of Health and Human Services. Retrieved from http://stacks.cdc.gov/view/cdc/6654.

risk and elevated variability of the corresponding SMRs; 2) the time evolution of adverse health events seems to be stationary, and no significant increase in the risk estimates is apparent when we examine the 2006-2009 period against 2002-2005.

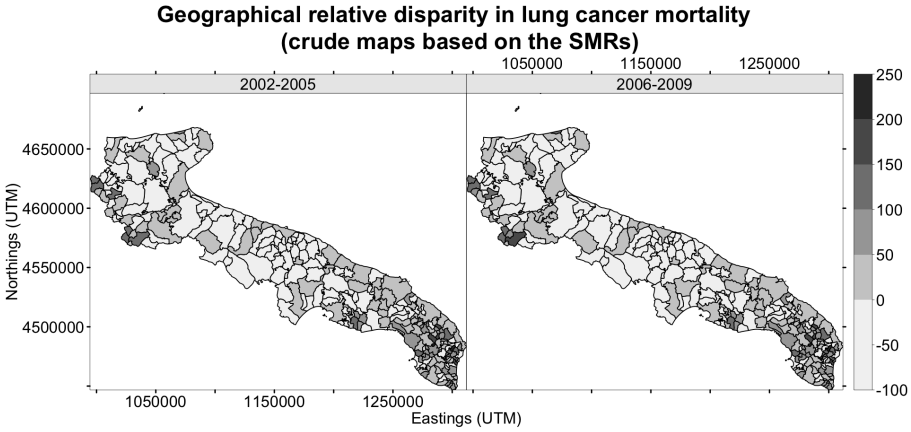

FIGURE 1 - Spatial distribution of relative mortality in lung cancer of males in Apulia, Italy, during the 2002-2005 and 2006-2009 periods. Maps are based on maximum likelihood estimates (MLE) of the relative risks (SMRs) from the saturated Poisson likelihood (4).

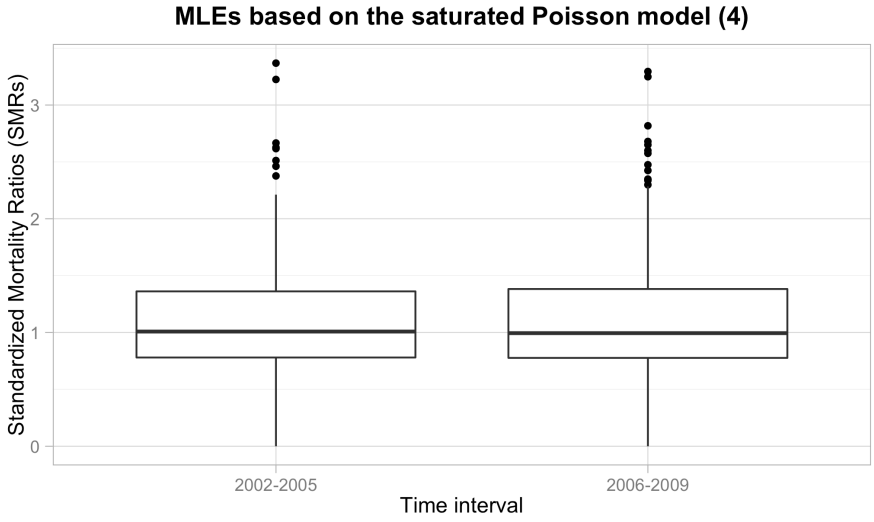

FIGURE 2 - Boxplots comparing the distribution of maximum likelihood estimates (MLE) of the relative risks (SMRs) for lung cancer of males in Apulia, Italy, during the 2002-2005 and 2006-2009 periods.

In order to confirm these findings, we fitted the Bayesian spatio-temporal model specified by the Poisson likelihood (4) and the linear predictor (5) plus a suitable

prior specification, a somewhat detailed account of which is provided in both Section 3 and 4. As we said before, the variability of Bayesian smoothed estimates across the map crucially depends on hyperparameters of the prior distribution of σ_v^2, σ_v^2. By default, we have specified weakly informative priors on the log of both structured and unstructured effect precision[27], exploring different hyperprior settings as follows:

`Model 1`: $\log \tau_v \sim \text{Gamma}(0.5, 0.0005)$, $\log \tau_v \sim \text{Gamma}(0.5, 0.0005)$;
`Model 2`: $\log \tau_v \sim \text{Gamma}(1, 0.0005)$, $\log \tau_v \sim \text{Gamma}(1, 0.0005)$;
`Model 3`: $\log \tau_v \sim \text{Gamma}(0.1, 0.1)$, $\log \tau_v \sim \text{Gamma}(0.1, 0.1)$.

We set $\log \tau_\delta \sim \text{Gamma}(1, 0.0005)$ for the log-precision of the local effects δ_i and a diffuse prior over the trend parameter β. A sound albeit informal criterion to decide among several competing Bayesian specifications is a generalization of the Akaike Information Criterion (AIC), based on the posterior distribution of the deviance statistic:

$$D(\eta) = -2 \log f(\mathbf{y}|\eta) + 2 \log h(\mathbf{y}), \tag{17}$$

where the log-likelihood of the current model is compared to the saturated log-likelihood $h(\mathbf{y})$, which is a function of the observed data vector $\mathbf{y}$ alone and does not affect posterior inference. For the Poisson likelihood (4) written in terms of the vector η of area-specific log-relative risks η_{it}, such that $\exp(\eta_{it}) = \theta_{it}$, the deviance statistic (17) assume the form:

$$D(\eta) = 2 \sum_{i=1}^{N} \sum_{t=1}^{2} \left\{ Y_i \left[\log \frac{Y_{it}}{\exp(\eta_{it}) E_{it}} \right] - \left[Y_{it} - \exp(\eta_{it}) E_{it} \right] \right\}. \tag{18}$$

As a measure for comparing complex hierarchical models in which the number of parameters is not clearly defined, we can define the Deviance Information Criterion[28] (DIC):

$$\text{DIC} = \bar{D} + p_D, \tag{19}$$

where $\bar{D}$ is the posterior mean of the saturated deviance (18) and:

$$p_D = \bar{D} - D\left[E(\eta|\mathbf{y}) \right], \tag{20}$$

[27] The precision is defined as the inverse variance $\tau = 1/\sigma^2$.

[28] D.J. SPIEGELHALTER, N. BEST, B.P. CARLIN, A. VAN DER LINDE, *Bayesian measures of model complexity and fit*, in "Journal of the Royal Statistical Society: Series B (Statistical Methodology)", 64:4 (2002), pp. 583–639.

equals the posterior expectation of deviance minus the deviance evaluated at the posterior expectation of log-relative risks. The proposed criterion can be justified by several arguments, on which $\bar{D}$ can be considered as a posterior summary of the goodness of fit of the current model, whilst p_D (the *effective number of parameters*, obtained after removing the interdependencies across parameter in the likelihood introduced by random effects in higher levels) is a penalty term measuring the complexity of the model. It is clear that smaller values of DIC will indicate a better-fitting model, after penalizing it for the effective dimension of the parameter space.

Model	$\bar{D}$	p_D	DIC
`Model1`	2764.25	110.01	2874.26
`Model2`	2766.18	106.93	2873.12
`Model3`	2752.89	124.88	2877.76

TABLE 1 – Deviance Information Criterion (DIC) for the three Bayesian spatio-temporal specifications defined by the likelihood/linear predictor (4)-(5), and the prior/hyperprior setting detailed in Section 3 and 7. $\bar{D}$ is the posterior mean of the saturated deviance (18), measuring model fit; p_D is the effective number of parameters, representing model complexity.

Table 1 presents the DIC components for the three specifications. It is apparent that `Model 2` has a consistently lower number of effective parameters, resulting in the smaller DIC. Bayesian smoothed estimates of the area-speficic relative risks provided by `Model 2` are showed in Figure 3. By visual eximination, the presence of a strong overall spatial trend in the disparity estimates is revealed, with spatial effects dominating the variability across the map and the highest disparities being estimated across the municipalities of the southeast part of the Salento peninsula.

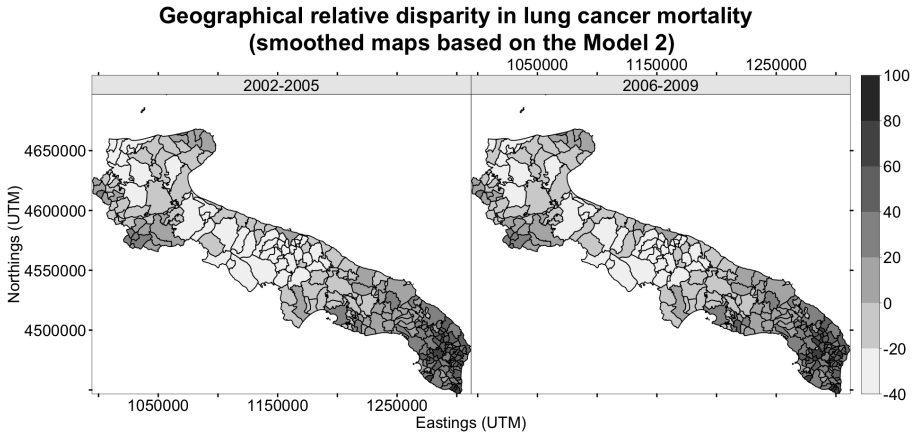

FIGURE 3 - Spatial distribution of relative mortality in lung cancer of males in Apulia, Italy, during the 2002-2005 and 2006-2009 periods. Maps are based on smoothed posterior estimates of the relative risks (SMRs) from the Bayesian spatio-temporal specifications defined by the likelihood/linear predictor (4)-(5), and the prior/hyperprior setting described in Section 3 and 7 (`Model 2`).

Also in this case, the existence of a global temporal trend in the relative risk surface cannot be confirmed. The fixed effects estimated by INLA are presented in Table 2, and if exponentiated they can be interpreted as relative risks. In particular, the global increase in risk from time period 1 to period 2 amounts to a modest 1.23%, $\exp(0.0122) \cong 1.0123$. In addition, the 95% posterior credible interval (-2.4%, 4.99%) includes 1%, meaning equal risk surfaces between the two time intervals. The same conclusion is apparent if we examine the boxplot of smoothed relative risk posterior estimates presented in Figure 4. For the 2002-2005 period we have a mean of 1.10, (interquartile range = 0.4677), versus the 2006-2009 period for which the mean is 1.1120 (interquartile range = 0.4643).

	Mean	SD	2.5%	50%	95%
α	0.0368	0.0320	-0.0262	0.0368	0.0994
β	0.0122	0.0186	-0.0243	0.0123	0.0487

TABLE 2 – Summary statistics for `Model 2`: posterior mean, posterior standard deviation (SD) and posterior 95% credible interval for the fixed effects of the spatio-temporal model defined by the specification (4)-(5) and the prior/hyperprior setting described in Section 3 and 7.

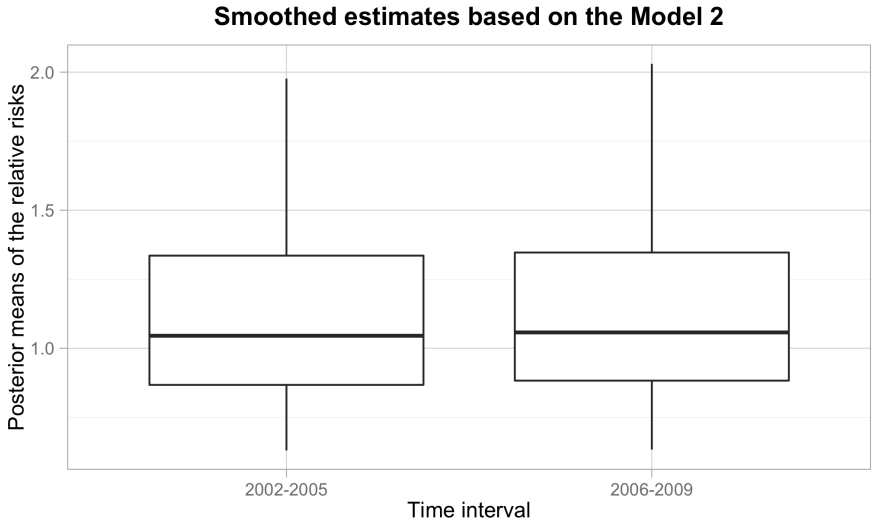

Figure 4 - Boxplots comparing the smoothed posterior estimates of the relative risks (SMRs) from the Bayesian spatio-temporal specifications defined by the likelihood/linear predictor (4)-(5), and the prior/hyperprior setting described in Section 3 and 7 (`Model 2`), for lung cancer of males in Apulia, Italy, during the 2002-2005 and 2006-2009 periods.

The detection of MLCs using the Kulldorff's circular scan statistic with 5% of the total population at risk to define the maximum circle size is presented in Figure 5. In

both time periods, the primary cluster includes a collection of municipalities across the central-eastern part of the Salento area. Hence, the results of the disease mapping exercise are strenghtened by the detection of spatial clusters. For the 2002-2005 period the p-value of MLC and the relative risk estimate inside the cluster were respectively $p = 0.001$ and $RR = 1.5669$, whilst we had $p = 0.001$ and $RR = 1.5994$ for the 2006-2009 period.

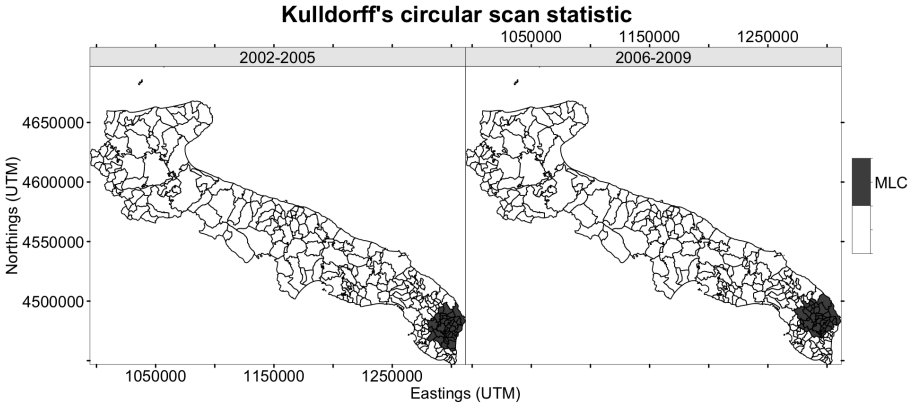

FIGURE 5 - Most likely cluster (MLC) of lung cancer mortality of males in Apulia, Italy, during the 2002-2005 and 2006-2009 periods.

8. - Although not directly comparable, the results presented in this paper confirm the previous findings[29]. The study identified a significant spatial cluster of excess male lung cancer mortality, located in a collection of municipalities lying across the southeast part of the Salento peninsula. The search for those risk factors responsible of the increase in risk is still underway. Exposure to environmental carcinogens (such as radon), as well as tobacco inhalation, are major risk factors for developing lung cancer. Familial heredity also increases an individual's risk, although the impact of complex genetic factors on the development of lung cancer remains unclear[30].

Previous research has focused on further aspects, such as the relevance of an important meteorological phenomenon, frequently observed on the Salento peninsula, consisting in the convergence of sea breezes in the middle of the peninsula, which would be responsibile for the transport of the emissions resulting primarily from the large industrial plants localized near the coasts in Brindisi and Taranto[31,32,33]. Of

[29] BILANCIA, FEDESPINA, *op. cit.*, p. 2
[30] R.W. LOGAN, D.K. SARKAR, *Genetic variation of the natural killer gene complex has a role in lung cancer susceptibility*, in "Journal of Thoracic Disease", 5:1 (2013), pp. 3–5.
[31] C. MANGIA, P. MARTANO, M.M. MIGLIETTA, A. MORABITO, A. TANZARELLA, *Modelling local winds over the Salento peninsula*, in "Meteorological Applications", 11:3 (2004), pp. 231–244.

course, at the present state of knowledge, it is a questionable theorem to affirm the existence of a direct causal link between pollutants originating from a distant single industrial source and the excess mortality for lung cancer observed among resident males.

The absence of a similar spatial pattern in risk among resident females[34] suggests that the role of occupational risk factors should be further considered. In particular, the 2009 IARC work group determined that there was sufficient evidence in humans for lung carcinogenicity of occupational exposures occurring during work activities in the following 6 occupational categories[35]: 1) coal gasification; 2) coke production; 3) iron and steel founding; 4) aluminum production; 5) painting; 6) rubber production industry. Others IARC group 1 lung carcinogen metals, such as arsenic and arsenic compounds, beryllium, cadmium, chromium and nickel are important for numerous common products, including textile products. The link between human health and the local industrial system present in the southern part of the Salento peninsula has not yet been investigated.

From a purely statistical point of view, a reasonable way to identify factors associated with spatial distribution of disease consists in representing the outcome variabile as a function of some selected area-specific explanatory variables:

$$\eta_{it} = \log(\theta_{it}) = \alpha + \gamma_1 x_{i1} + \cdots + \gamma_p x_{ip} + \text{random effects}. \qquad (21)$$

For example, the relationship between socioeconomic factors and mortality for lung cancer in Tuscany has been investigated[36], confirming the presence of an association with a latency time of 10 years. Researchers should be aware that regression modeling of aggregated data may inadequately represent the exposure-response relationship at the individual level, an effect known as ecological bias[37]. However, these analyses may identify areas where the incidence/mortality level is not accounted for by the explanatory variables (i.e. known risk factors). These areas, once indentified, might be targeted for further investigations. Future research will undoubtely benefit from explicitly investigating all the above-mentioned issues.

[32] C. MANGIA, I. SCHIPA, A. TANZARELLA, D. CONTE, G.P. MARRA, G. P., M. M. MIGLIETTA, U. RIZZA, *A numerical study of the effect of sea breeze circulation on photochemical pollution over a highly industrialized peninsula*, in "Meteorological Applications", 17:1 (2010), pp. 19–31.

[33] C. MANGIA, M. CERVINO, A.E.L. GIANICOLO, *Secondary Particulate Matter Originating from an Industrial Source and Its Impact on Population Health*, in "International Journal of Environmental Research and Public Health", 12:7 (2015), pp. 7767–7781.

[34] BILANCIA, FEDESPINA, *op. cit.*, p. 2

[35] R.W. FIELD, B.L. WITHERS, *Occupational and Environmental Causes of Lung Cancer*, in "Clinics in Chest Medicine", 33:4 (2012), pp. 681–703.

[36] E. DREASSI, *A space–time analysis of the relationship between material deprivation and mortality for lung cancer*, in "Environmetrics", 14:5 (2003), pp. 511–521.

[37] H. MORGENSTERN, *Ecologic Studies in Epidemiology: Concepts, Principles, and Methods*, in "Annual Review of Public Health", 16:1 (1995), pp. 61–81.