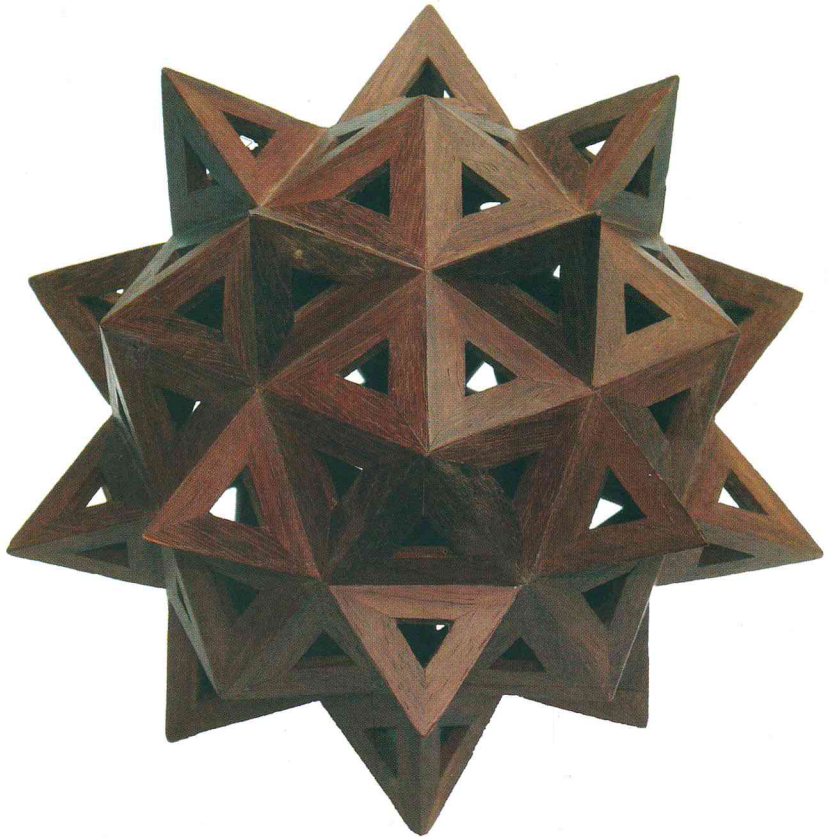


Classification and Data Analysis 2007

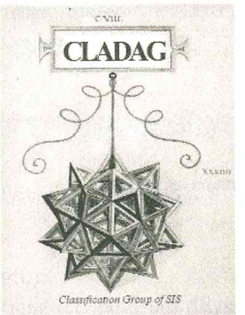
Book of Short Papers

Meeting of the Classification and Data Analysis Group of the Italian Statistical Society

CLASSIFICATION AND DATA ANALYSIS 2007

Book of Short Papers

Meeting of the
CLAssification and Data Analysis Group
of the Italian Statistical Society



MACERATA - SEPTEMBER, 12TH-14TH, 2007

eum

Isbn 978-88-6056-020-9
Prima edizione: settembre 2007

Vicolo Tornabuoni, 58 - 62100 Macerata
info.ceum@unimc.it
http://ceum.unimc.it
Realizzazione e distribuzione:
Quodlibet società cooperativa
Via S. Maria della Porta, 43 - 62100 Macerata
www.quodlibet.it
Stampa: Grafica Edirice Romana s.r.l., Roma

This book was realized with the financial support of

Università degli Studi
di Macerata

REGIONE
MARCHE

SISTEMA
INFORMATIVO
STATISTICO

Table of contents

Specialized Session 2
Statistical models with errors-in-covariates
Organizer: A. Pastore

Specialized Session 3
Multivariate Analysis for Microarray Data
Organizer: M Vichi

Specialized Session 4
Cluster Analysis of complex data
Organizer: R. Verde

Specialized Session 5
Educational Processes Assesment by means of Latent Variables Models
Organizer: S. Mignani

Specialized Session 6
Classification of complex data
Organizer: D.A. Zighed

Specialized Session 7
Multidimensional Scaling
Organizer: A. Okada

Specialized Session 8
Statistical Models for Public Policies
Organizer: S. Ingrassia

Solicited Papers Session 4
Customer Relationship Management
Organizer: C. Davino

Solicited Papers Session 5
Missing Data
Organizer: A. Plaia

Solicited Papers Session 6
Statistical methods in decision-making
Organizer: R. Amenta

Solicited Papers Session 7
Multivariate methods in social research
Organizer: D. F. Iezzi

Solicited Papers Session 8
Statistical Models with Latent Features
Organizer: L. Tardella

Solicited Papers Session 9
Classificaion Trees
Organizer: R. Siciliano, M. Aria

An algorythm for earthquakes clustering based on maximum likelihood [1]

Giada Adelfio, Marcello Chiodi
Dipartimento di Scienze Statistiche e Matematiche "Silvio Vianelli"

Università degli studi di Palermo
Viale delle Scienze, Ed.13, 90144 Palermo
adelfio@dssm.unipa.it, chiodi@unipa.it

Dario Luzio
Dipartimento di Chimica e Fisica della Terra ed Applicazioni alle Georisorse e ai Rischi Naturali
Università degli studi di Palermo
Via Archirafi, 36 - 90123 Palermo
luzio@unipa.it

Abstract: We propose a clustering technique to separate the two main components of seismicity: the background seismicity and the triggered one. We suppose that a seismic catalogue is the realization of a non homogeneous space-time Poisson clustered process, with an intensity function obtained mixing a Poisson-type component and clustered components. The proposed method assigns each earthquake to a cluster or to the set of independent events according to the increment to the overall likelihood and iteratively changing the assignment of the events; after a change of partition, MLE of parameters are computed again;the process is iterated until there is no more improvement in the likelihood. The algorythm revealed quite sensible to some initial choices: nevertheless we obtained satisfactory results when applied to the South-Tyrrenian catalogue.

Keywords: Earthquakes clustering, point process, intensity function, MLE.

1 Introduction

Because of the different seismogenic features controlling the kind of seismic release of background and clustered seismicity (Adelfio and Chiodi, 2006), to describe the seismicity of an area in space, time and magnitude domains, sometimes it is useful to study separately the features of *independent* events and *triggered* ones: different kinds of sets of events give different information on the seismicity of an area. For the estimation of parameters of phenomenological laws useful for the description of seismicity, we need a good definition of *earthquake cluster*, based on the definition of suitable point processes mechanism; furthermore the prediction of the occurrence of large earthquakes, related to the assessment of seismic risk in space and time, is complicated by the presence of clusters of aftershocks, that are superimposed to the background seismicity, according to some (unknown) mixing parameter, and shade its principal characteristics.
For this purposes the preliminary subdivision of a seismic catalog in background seismicity (represented by isolated events, that do not trigger any further event, and the mainshock of each seismic sequence) and clustered events is sometimes required. At this regard, a seismic sequences detection technique is presented; it is based on MLE of parameters that identify the conditional intensity function of a model describing seismic activity as a clustering-process (Adelfio and Chiodi, 2006) and representing a slight modification

[1] This research has been partially funded by the University of Palermo, year 2005

of the ETAS model (Epidemic Type Aftershocks-Sequences model; Ogata (1988), Ogata and Zhuang (2004)).

2 Conditional intensity function in point processes

A seismic catalogue contains information about seismic events occurred in a region, in a given time interval: we suppose to have a seismic catalogue of n events, and the i-th row of the catalogue reports quantitative information about a seismic event U_i, $(i = 1,\dots,n)$, where x_i, y_i, z_i are the estimated latitude, longitude and depth, t_i is the time of occurrence and m_i the magnitude of the event.
We suppose that the catalogue is the realization of a space-time point process. In this paper, as usual when modelling observed seismicity, the depth z will not be considered, since the high level of its measurement error.
The conditional intensity function of a space-time point process can be defined as:

$$\lambda(t,x,y|H_t) = \lim_{\Delta t,\Delta x,\Delta y \to 0} \frac{Pr_{\Delta t \Delta x \Delta y}(t,x,y|H_t)}{\Delta t \Delta x \Delta y} \tag{1}$$

where H_t is the space-time occurrence history of the process up to time t; $\Delta t, \Delta x, \Delta y$ are time and space increments; $Pr_{\Delta t \Delta x \Delta y}(t,x,y|H_t)$ is the history-dependent probability that an event occurs in the volume $\{[t,t+\Delta t) \times [x,x+\Delta x) \times [y,y+\Delta y)\}$. The conditional intensity function completely identifies the features of the associated point process (Schoenberg and Bolt, 2000) (i.e. if it is independent of the history but dependent only on the current time and the spatial locations (1) supplies a nonhomogeneous Poisson process; a constant conditional intensity provides a stationary Poisson process).
ETAS model is a self-exciting point process describing earthquakes catalogs as a realization of a branching or epidemic-type point process. In particular it could be considered as an extension of the Hawkes model (Hawkes, 1971), which is a generalized Poisson cluster process associating to cluster centers a branching process of descendants.
Indeed in our formulation, we suppose that the seismic catalog is the realization of a clustered inhomogeneous Poisson process, that is obtained (Daley and Vere-Jones, 1988) supposing that points of the background seismicity come from a space-time Poisson process (spatially inhomogeneous) and that among these there is a number k of *mainshocks* that can generate aftershocks sequences, inhomogeneous both in space and times, and with an intensity also linked to the magnitude of the main event. We define the intensity function by:

$$\lambda(x,y,t;\theta) = \lambda_t \mu(x,y) + K_0 \sum_{\substack{j=1 \\ (t_j<t)}}^{k} g_j(x,y) \frac{\exp[\alpha(m_j - m_0)]}{(t-t_j+c_j)^{p_j}} \tag{2}$$

where $\theta = (\lambda_t, K_0, c_j, p_j, \alpha)$. In (2) t_j and m_j are time of the first event and magnitude of the mainshock of the cluster j, $g_j(x,y)$ is the space intensity of the cluster j and $\mu(x,y)$ is the background one; K_0 and λ_t are the weights of the clustered seismicity and of the background one, respectively. Background seismicity is assumed stationary in time, while time aftershock activity is represented by a non stationary Poisson process according to the modified Omori formula (Utsu, 1961), of parameters c_j and p_j, relating the occurrence rate of aftershocks to the mainshock magnitude (with α measuring the influence on the relative weight of each sequence and m_0 the completeness threshold of magnitude, that is the lower bound for which earthquakes with higher values of magnitude are surely recorded in the catalogue).
In our approach space intensity both of background seismicity and of each cluster, is estimated by a bivariate kernel estimator: it is computed either using only the independent events (isolated and mainshocks) or points belonging to the cluster, including the mainshock, respectively. In both cases the smoothing constant, as a first approximation, is evaluated with Silverman's formula (Silverman, 1986).
In the evaluation of (2), different kinds of parametrization, that allow to take into account for different assumptions on the seismicity of an area, are considered (e.g. Omori's law parameters p_j of the k clusters can be assumed equal or distinct in each cluster). The choices can be compared at the end of the procedure comparing the likelihood values obtained.
The hypotesis underlying the ETAS model are slightly different, because each point (either mainshock or aftershock) can generate an offspring, with an intensity varying according to distance in space and from the triggering event: there is not a distinction between isolated points, mainshocks and aftershocks, because each point can be either an offspring (aftershock) or a generating event (mainshock). However a stochastic declustering technique has been proposed for this method (Ogata and Zhuang, 2004) in order to build catalog that should maintain the property of the background seismicity, even if it will not afforded in this paper.

3 The proposed clustering method

The clustering technique that we propose leads to an intensive computational procedure, implemented by software R (R Development Core Team, 2005). It identifies a partition of a catalogue of seismic events, $\mathscr{P}_{k+1}$, formed by $k+1$ sets: the background seismicity and the clustered events (k disjoint sets). It iteratively changes the partition assigning events either to the background seismicity or to the $j-th$ cluster $(j = 1,\dots,k)$, on the basis of the likelihood function variation due to their moving from a set to another one.
Let $[T_0, T_{max}]$ and Ω_{xy} time and space domains of observation respectively; the likelihood function to be maximized is:

$$\log L(\theta) = \sum_{i=1}^{n} \log \lambda(x_i,y_i,t_i;\theta) - \int_{T_0}^{T_{max}} \int_{\Omega_{xy}} \lambda(x,y,t;\theta) dx dy dt \tag{3}$$

where $\lambda(x,y,t;\theta)$ is defined by (2).
To start, the proposed method needs an initial partition, found by a single-linkage method procedure (Frolich and Davis, 1991) or choosing a threshold parameter (Resenberg, 1985), an initial number of clusters is determined, but then the classification of the events can change, because each event could change its position moving to the set for which the likelihood function is maximized.
The space densities and the ML estimates of the parameters of the intensity function (2) are evaluated. Then we try to move all events (one by one) from their current position and we compute the change in (3) due to these movements, using the current value of the estimated parameters. If the current value of the likelihood increases, then the movements become effective. If at least one event changes its position the partition $\mathscr{P}_{k+1}$ is updated and the algorithm restarts from the intensity estimation step; the number of clusters could decrease during the iterative optimization. The iterative procedure stops when the current classification does not change after a whole iteration.

4 Real data application and final remarks

Finally we report here some brief results of the application of our method to a catalogue of 1080 seismic events (Adelfio *et al.*, 2006) occurred in the south-tyrrenyan area in approximately fifteen years. Four main clusters were found, starting from an initial partition founded on a single-linkage type method; two different parametrizations for the intensity function (2) have been considered: one with a common value of p and one with distinct values p_j for each cluster; evidence for the second model has been observed; we experimented a computing time of about thirthy minutes on a common PC with a 2,66GHz CPU.
The proposed method has some critical aspects that have been afforded (the computational burden and the problem of the initial choice); it could be the basis to carry out an analysis of the complexity of the seismogenetic processes relative to each sequence and to the background seismicity, separately: when applied to Italian and South-Tyrrenian catalogues it returns a plausible separation of the different components of seismicity and clusters that have a good interpretability, estimating the space pattern of the induced seismicity through non parametric methods, using only the events of each cluster.

References

Adelfio G. and Chiodi (2006) Earthquakes clustering based on maximum likelihood estimation of point process conditional intensity function, *4th International Workshop on statistical seismology, Hayama Campus, Japan, January 9-13,2006*, 2–5.

Adelfio G., Chiodi M., De Luca L., Luzio D. and Vitale M. (2006) Southern-tyrrhenian seismicity in space-time-magnitude domain, *Annals of geophysics, 49*, 1139–1151.

Daley D.J. and Vere-Jones D. (1988) *An introduction to the theory of point processes*, New York: Springer-Verlag.

Frolich C. and Davis S.D. (1991) Single-link cluster analysis, synthetic earthquake catalogues, and aftershocks identification, *Geophysical Journal International, 104*, 289–306.

Hawkes A. (1971) Spectra of some self-exciting and mutually exciting point processes, *Biometrika, vol. 58, No. 1*, 83–90.

Ogata Y. (1988) Statistical models for earthquake occurrences and residual analysis for point processes, *Journal of the American Statistical Association, 83, 401*, 9–27.

Ogata Y. and Zhuang J. (2004) Analyzing earthquake clustering features by using stochastic reconstruction, *Journal of Geophysical Research, 109*, B05301:1–17.

R Development Core Team (2005) *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Resenberg P. (1985) Second-order moment of central california seismicity, 1969-1982, *Journal of Geophysical Research, vol. 90, No. B7*, 5479–5495.

Schoenberg F.P. and Bolt B. (2000) Short-term exciting, long-term correcting models for earthquake catalogs, *Bulletin of the Seismological Society of America, vol. 90, No. 4*, 849–858.

Silverman B.W. (1986) *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, London.

Utsu T. (1961) A statistical study on the occurrence of aftershocks, *Geophys. Mag., 30*, 521–605.

Hospital performance comparison: assessing inappropriate stay in the hospitals of Palermo

Francesca Di Salvo
Depart. Stat. Mat. Science
University of Palermo
viale delle Scienze, 13 (PA)
disalvo@dssm.unipa.it

Nicola Ferotti
Hosp. Activ. Depart.
AUSL 6 Palermo
via Cusmano, 24 (PA)
nicolaf18@tin.it

Sergio Consagra
Progr. Manag. Unit
AUSL 6 Palermo
via Cusmano, 24 (PA)
s.consagra@tiscali.it

Abstract: Increasing attention is being focused on evaluating and improving inhospital health care quality and efficiency at both the local and national levels. The present study is based upon 2006 hospitalization data of the eight hospitals of the Health Authority AUSL 6 located in the province of Palermo. The main critical aspects concern the appropriatness of hospital admissions, based on the definition of "Inappropriateness High Risk DRGs" and of "sentinel DRGs"; the meaning of such inappropriate hospitalization is the inability to refer the patients to a more appropriate day hospital or alternative care structures.
Our purpose is to investigate *Overutilization* (medical unnecessary) practice in the AUSL 6 and to identify where unnecessary hospital use is likely to occurr. Having computed inappropriate utilization rates by mean of AP *P*RO method in each of the eight hospitals, we compare them in order to assess their performances.

Keywords: Appropriate hospitalization, High Risk DRG, AP *P*RO, APR-DRG.

1 Introduction

The disease management is the collection of the processes aimed to control the health care and improving the quality at same time reducing the overall cost of the procedures. Efforts are being made to provide standardized, useful and valid information about hospitals, to contain costs and also to improve the efficiency using capacity of existing healt care facilities as optimally as possible. The Italian SSN promoted the reduction of the use of hospital beds through the definition of a list of 43 "Inappropriatness High Risk" DRGs in the D.P.C.M. 29/10/2001, resulting inappropriate when they are registred as long stay. Moreover a list of 53 "sentinel DRGs" are defined in the *"Testo Unico in materia di compensazione interregionale della mobilità sanitaria"* as medical DRGs correctly performed on a day care basis. The aim is to achieve fewer hospitalizations, medical procedures and medical errors, without sacrificing quality of care. We use the term *inappropriate* to refer to those admissions for which hospital-level care is not medically justified. Results presented here are a first step in this ongoing process to evaluate health care quality in AUSL 6 hospitals operating in the province of Palermo. Data for this report were collected on patients who had been in the hospitals of the Health Authority AUSL 6 during the period from January 1, 2006 through September 30, 2006. In that period, 26550 patients were admitted with an average length of stay of 4 .77 days. Hospital structures of the AUSL 6 consist in nine hospitals: in this study only eight hospitals are considered, excluding the one with only one Unit and less than 10 admissions; they will be indicated by mean of the respective ISTAT codes. On the basis of the D.P.C.M. 29/10/2001 and the *"Testo Unico in materia di compensazione interregionale della mobilità sanitaria"* , data on hospital dis-

charges from the SDO register are analysed in detail to assess the inappropriate setting. Results indicate that in overall AUSL 6 a percentage of 75,95% of hospital admissions was appropriate, while 14,79% were hospital admission with high risk of inappropriateness and 9,25% were sentinel hospital stay; but a more in-depth analysis highlights important differences between medical and surgical DRGs. Whether inappropriateness explains much of the variation in admission rates among hospitals is an important issue for public policy; to compare hospital performance, the weighted mean of the inappropriatness rates for DRGs treated in each hospital is computed.

Table 1: *Distribution of the hospital admissions by typologies of DRGs*

	Medical	**Surgical**	**Total**
High Risk	18.886%	7.133%	14.796%
Sentinel	13.412%	0.000%	9.350%
Non High Risk	67.702%	92.867%	75.954%

2 The admissibility thresholds for Diagnosis-Related Groups (DRGs)

The diagnosis-related groups, DRGs, are defined by diagnosis and other factors.
DRG assignment is based on diagnosis and treatment information of the patient's SDO; the coded information includes the principal diagnosis, up to additional diagnoses and procedures, the patient age, sex, and discharge status.
Each of the DRGs has an official weight that determines payment. The array of patients across DRGs in a hospital is the hospital's case-mix, and the average DRG weight for these patients is the hospital's case-mix index. Reduction of hospitalization of "Inappropriatness High Risk DRGs " and "sentinel DRGs " is a necessary step of the cost containment efforts planned by Regional Health Care Agency of Sicily, that has recently focused on reducing expenditures for acute-level care and incentivizing day hospital practice or alternative medical care processes. Interventions can include comparing appropriateness rates of different hospitals, in order to encourage more efficient practice, and using sanctions as cuts in hospital remunerations taking into account the typologies of admission (day hospital/long stay) and the distinction between medical and surgical DRGs.
Assessment of admissibility thresholds for DRGs may depend from regional laws or cost containment strategies. For our purpose, we consider three different thresholds: for the 'High Risk' DRGs, the National thresholds indicated in Baglio et al. (2003); Regional thresholds defined by Regional Health Care Agency of Sicily (decree n. 70 6732, 22/07/2002); a third threshold is here computed at AUSL level as the percentage of hospital long stay (≥ 2 days), on the basis of the overall DRGs with a low severity of illness.

3 Analysis of Appropriateness

In Italy a widely used clinical-based tool to evaluate hospital appropriateness is the PRUO (Review of Hospital usage Protocol) deriving from the AEP (Appropriatness Evaluation Protocol) developed in USA (Gertman and Restuccia, 1981). PRUO evaluates either the admission and all days during the patient's stay in according to a dichotomous variable (appropriate /inappropriate) and make a summary judgment to assess whether or not the services which the patient received justified acute-level care; aggregated evaluations are used to judge appropriateness of hospitals and Units.
The Health Authority of Lazio has developed the AP *P*RO method for assessing organizational inappropriateness of hospital care using administrative data (Fortino et al., 2002); the evaluation procedure, that aims to identification of standard hospital long stays that could be efficiently assisted in a lower level, is based on statistical criteria and requires moderate resources.
The AP*P*RO method computes the proportions of inappropriate hospital stay through a preliminary selection of DRGs with a low severity of illness, for which long stay is potentially to avoid or at least to keep at minimum. The selection is based on the system of APR-DRG (*All Patient Refined* DRG) (Baglio et al., 2001), a classification of DRGs in classes characterized by different degrees of severity. Selection of the specific methods or instruments to use in identifying inappropriate utilization depends on the overall goals and scope of the program and on resources. In this study, inappropriate episodes were estimated on the basis of AP *P*RO method.
To determine inappropriateness with respect to AUSL thresholds all DRGs with more than 20 patients are considered; the selection of cases included in the analysis is based on following criteria: patient's age ≥ 2 and ≤ 69; length of stay lower than thresholds; exclusion of inhospital death.
The final dataset consists of 16398 cases, for which Case-Mix index is reported in Table 2; results relating inappropriateness are resumed in Table 3.

Table 2: *Hospital case-mix index for selected dataset*

Hospital	**180**	**181**	**182**	**183**	**184**	**190**	**191**	**193**	**All**
admissions	802	2048	921	446	3192	5102	481	3406	16398
Case Mix Index	0.83	0.88	0.85	0.82	0.83	0.92	0.68	0.71	0.84

To determine inappropriateness with respect to Regional and National thresholds, SDO relating to the 43 'High Risk' are selected from the eight hospitals; for these DRGs the proportion of long stays on the total (long stays + Day Hospitals) is computed whithin the hospital; then the proportions of inappropriate admissions are computed as the percentages of excesses with respect to Regional and National admissibility thresholds described in Section 2. The results are resumed in Table 4.

4 Conclusions

For each Hospital a performance rate is the weighted mean of inappropriateness rates of the respective DRGs, where the weights are the number of patients across DRGs.
A comparison of performance rates of the hospitals is performed by means of ANOVA test, using results of Tables 3 and 4; differences among the hospitals are statistically significantly ($p < 0.001$).
The analysis highlights a substantial amount of the potentially avoidable use of hospital resources and describes differences in admission practice among the hospitals of the AUSL 6. The distinction between medical and surgical DRGs is maintained since an appreciable difference is revealed among percentages: it indicates that the practice of day surgery is well-established, while medical DRGs keep to be performed in long stay hospitalization even if it's not strictly necessary. As urgency of admission, residence, age and

comorbidity may cause appropriate long stay, they may be relevant factors in the analysis, and further investigations may concern the study of independent predictive factors of inappropriate stays. To asses the validity of the results obtained an useful step may be the investigation of the relation between APR-DRG severity subgroups and PRUO assessment, although examining the findings from different studies needs some cautions.

Table 3: *Excess of inappropriate admission with respect to AUSL threshold*

Hospital	Medical	Surgical	All
180	0.000	−2.590	−2.072
181	4.984	−7.131	−2.419
182	11.283	5.718	8.810
183	20.771	−6.649	7.061
184	7.414	10.683	8.404
190	−6.772	3.867	−3.295
191	24.354	–	24.354
193	−4.254	7.499	−2.227

Table 4: *Excess of inappropriate admission with respect to national and regional thresholds for high risk DRGs.*

	National threshold			Regional threshold		
	Medical	**Surgical**	**All**	**Medical**	**Surgical**	**All**
180	41.74	−21.56	15.08	97.73	31.40	69.80
181	23.69	−42.55	−2.80	82.54	16.18	56.00
182	31.49	−5.54	22.23	89.79	51.91	80.60
183	42.01	−6.66	27.19	97.67	51.90	83.74
184	27.42	−38.56	5.43	82.59	18.64	61.27
190	1.20	12.40	4.94	54.81	68.88	59.36
191	15.72	–	15.71	83.69	–	83.69
193	5.31	−48.08	−8.74	70.11	14.67	55.52

References

D.P.C.M. 29 NOVEMBRE 2001. Definizione di livelli essenziali di assistenza. Supplemento ordinario n.26 alla Gazzetta Ufficiale n.33, 8 febbraio 2002.

Baglio G., Di Domenicantonio R., Materia E., Guasticchi G. (2003) La Valutazione della appropriatezza organizzativa per modulare le tariffe della Compensazione Interregionale della Mobilita Sanitaria. *Tendenze Nuove*,2: 108-120.

Baglio G., Materia E., Vantaggiato G., Perucci C. (2001) Valutare appropriatezza dei ricoveri con dati amministrativi: ruolo degli APR-DRG. *Tendenze Nuove*,1: 51-70.

Fortino A., Lispi L., Materia E., Di Domenicantonio R., Baglio G.(2002) La valutazione della appropriatezza dei ricoveri ospedalieri in Italia con il metodo APPRO. *Panorama Sanità*, 32: 42-49.

Gertman P.M., Restuccia J.D.(1981) Th appropriateness evaluation protocol: a technique for assessing unnecessary days of hospital care. *Medical Care*, 8: 855-871.

Weighted Likelihood Inference in Gaussian Spatial Linear Models

Carlo Gaetan
Dipartimento di Statistica
Università Ca'Foscari - Venezia
S.Giobbe, Cannaregio, 30121 Venice, ITALY
gaetan@unive.it

Luca Greco
Dipartimento PE.ME.IS.
Università del Sannio
P. Arechi II, 82100 Benevento, ITALY
luca.greco@unisannio.it

Abstract: The presence of anomalous observations can badly affect inference in Gaussian spatial linear models. Therefore, we propose a robust procedure which, on the one hand, allows us to take into account possible departures of the data from the specified model, and on the other hand can help in identifying spatial outliers. This procedure is based on weighted likelihood methodology.

Keywords: Outliers, Profile likelihood, Spatial linear model, Weighted likelihood

1 Introduction

Let $Y(s)$ be a random variable measured on different locations $s_i \subset \mathbb{R}^2$, $i = 1, \ldots, n$ representing points or regions in the Euclidean plane. As an illustrative example of our proposal, we focus on the mixed regressive spatial autoregressive model (SAR) introduced by Cliff and Ord (1981):

$$Y(s_i) = \rho \sum_{j \neq i} w_{ij} Y(s_i) + X(s_i)'\beta + \sigma\epsilon(s_i)\,, \qquad \sigma > 0\,, \tag{1}$$

$X(s_i)$ is a p-dimensional vector of covariates and $\epsilon(s_i) \sim \mathrm{N}(0,1)$, $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$, $i \neq j$. The spatial relationships between sites are described by a known connectivity matrix $W = [w_{ij}]$, whose elements are zero unless sites s_i and s_j are neighbors. The unknown vector parameter is $\theta = (\beta, \sigma, \rho)$. As well as in the case of independence or in time series, the presence of anomalous observations can badly affect a likelihood based inference, both on the significance of any trend parameter β and the strength of the spatial dependence ρ. In view of this, there is the need of a robust procedure which, on the one hand, allows us to take into account possible departures of the data from the specified model, basically due to the presence of outliers in the sample, and on the other hand can help in detecting anomalous situations. In this work, we propose to extend the weighted likelihood based algorithms for linear regression models with independent errors (Agostinelli and Markatou, 2001) to the model (1). The method is described in Section 2. The finite-sample performance of weighted likelihood estimates (WLEs) is studied numerically in Section 3. Finally a real data example is given in Section 4.