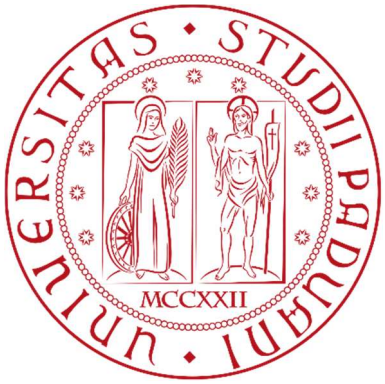
Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Corso di Laurea Triennale in Statistica per le Tecnologie e le Scienze

MODELLI STATISTICI APPLICATI



CONFLITTI NEL CONTINENTE AFRICANO

A cura di

Rebecca Esegio (matr: 20111389),

Gicomo Filippin (matr: 2003009),

Emma Lovato (matr: 2008153)

INTRODUZIONE

I dati provengono dal sito ACLED D (The ACLED Conflict Alert System), un sito dedicato a strumenti interattivi di previsione degli eventi politici violenti nel futuro. Uno studio osservazionale socio-economico-climatologico ha raccolto 38216 osservazioni sugli eventi relativi ai conflitti politici ti tipo violento nel continente Africano

OBIETTIVO

Obiettivo principale dello studio di questi dati è vedere come vari aspetti descritti in questi dataset possano influenzare la formazione di conflitti tra le regioni africane. È inoltre importante capire la tipologia di conflitto.

SOFTWARE UTILIZZATO

Software R - Version 4.2.2 (2022-10-31), Copyright (C) 2010 The R Foundation for Statistical Computing, http://cran.r-project.org/bin/windows/base/old/4.2.2

DATI E ASSUNZIONI

I dati sono contenuti nel file 'conflict10'.csv.
Il dataset contiene le seguenti variabili:

- Year: anno dell'evento di conflitto
- Status: vale 1 se l'evento è una *Riots*, vale 0 se l'evento è una *Battles* o *Violence against civilians*
- EventType: tipo dell'evento di conflitto (Riots, Battles, or Violence against civilians)
- Actor1: primo attore del conflitto
- Actor2: secondo attore del conflitto
- Country: Paese dove è avvenuto l'evento di conflitto
- Region: Regione dove è avvenuto l'evento di conflitto
- Location: localizzazione del conflitto
- ConflictLat: latitudine geografica del luogo in cui è avvenuto il conflitto
- ConflictLong: longitudine geografica del luogo in cui è avvenuto il conflitto
- StationID: codice identificativo della stazione metereologica
- YrMoDy: data dell'evento di conflitto nel fomato Year/Month/Day
- MaxTemp: temperatura massima (in gradi Fahrenheit) del giorno in cui è avvenuto il conflitto
- StationName: nome della stazione metereologica
- StationLong: longitudine geografica della stazione metereologica
- StationLat: latitudine geografica della stazione metereologica 1
- TempCat: livello di temperatura (C = freddo, M = mite, H = caldo) assegnato sulla base delle temperature registrate dalla stazione StationID

Il dataset 'weeklydata.txt' contiene dei dati aggiuntivi sul clima rilevati da tutte le stazioni metereologiche in Africa.
Il dataset contiene le seguenti variabili:

- Year: anno dei conflitti
- Week: numero d'ordine della settiana nell'anno d'interesse. (da 1 a 53)
- StationID: codice identificativo della stazione metereologica
- BattlesCount: numero totale di conflitti di tipo 'Battles' accaduti nella settimana e relativi alla stazione StationID
- VACCount: numero totale di conflitti di tipo 'Violence against civilians' accaduti nella settimana e relativi alla stazione StationID
- RiotsCount: numero totale di conflitti di tipo 'Riots' accaduti nella settimana e relativi alla stazione StationID

- TotalCount: numero totale di conflitti di ogni tipo (somma di 'Violence against civilians', 'Battles', 'Riots')
- HighMaxTemp: la temperatura massima giornaliera (in gradi Fahrenheit) più alta rilevata nella settimana e relativa alla stazione stationID (corrisponde al valore massimo della variabile MaxTemp nel dataset degli eventi di conflitti)
- TempCat: livello di temperatura (C = freddo, M = mite, H = caldo) assegnato sulla base delle temperature registrate dalla stazione StationID, costruito a partire dalla variabile HighMaxTemp (uguale alla variabile 'TempCat' nel dataset degli eventi di conflitti)

Si assume un livello di significatività fissato al 5%.

ANALISI DI SOPRAVVIVENZA

Analisi univariata

Si procede quindi con l'analisi dal punto di vista univariato di *Year* unita a *status* che, nell'analisi di sopravvivenza, costituisce la variabile risposta.

```
Call: survfit(formula = Surv(Year, status) ~ 1, data = dati)
```

time	n.risk	n.event	survival	std.err	lower 95% CI	upper 95% CI
1997	2763	104	0.962	0.00362	0.955	0.969
1998	2603	9	0.959	0.00377	0.952	0.966
1999	2397	9	0.955	0.00395	0.948	0.963
2000	2194	16	0.948	0.00428	0.940	0.957
2001	1985	15	0.941	0.00463	0.932	0.950
2002	1807	8	0.937	0.00484	0.928	0.947
2004	1413	2	0.936	0.00493	0.926	0.946
2005	1227	19	0.921	0.00587	0.910	0.933
2006	913	5	0.916	0.00625	0.904	0.929
2008	656	5	0.909	0.00694	0.896	0.923
2009	345	8	0.888	0.01001	0.869	0.908
2010	76	1	0.877	0.01525	0.847	0.907

Si può notare che nel 2001 i soggetti a rischio sono 1985, con il rispettivo numero di eventi uguale a 15 e probabilità di sopravvivenza stimata pari a 0.94. Nell'anno successivo il numero dei soggetti a rischio scende a 1807 data la presenza di un elevato numero di dati censurati che vengono rappresentati nella Figura 3 dal simbolo + sulla curva di K-M.

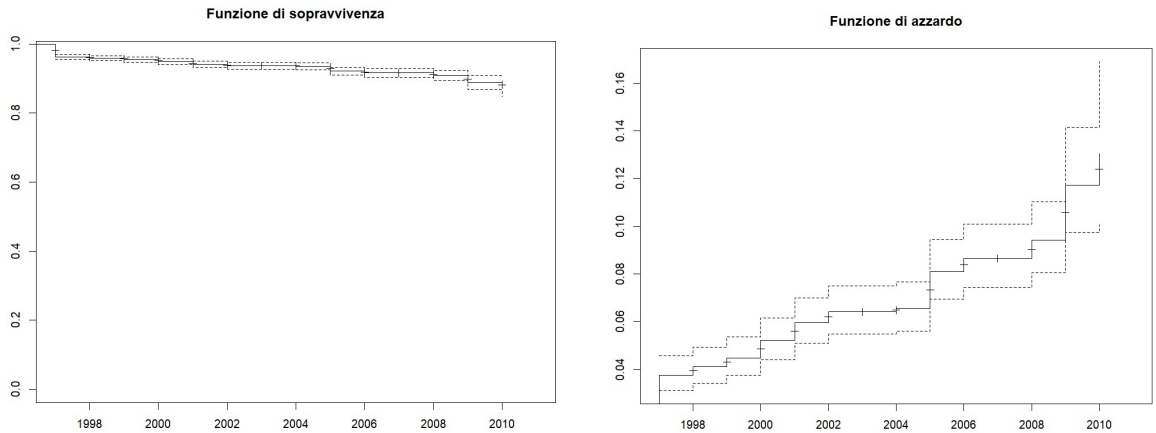


Figura 3: Curva di sopravvivenza e di rischio cumulato per Year.

Variabile risposta e TempCat

La Figura 4 rappresenta la stima delle curve di sopravvivenza di K-M e la funzione di azzardo in relazione al livello di temperatura (C = freddo, M = mite, H = caldo) assegnato sulla base delle temperature registrate dalla stazione StationID.

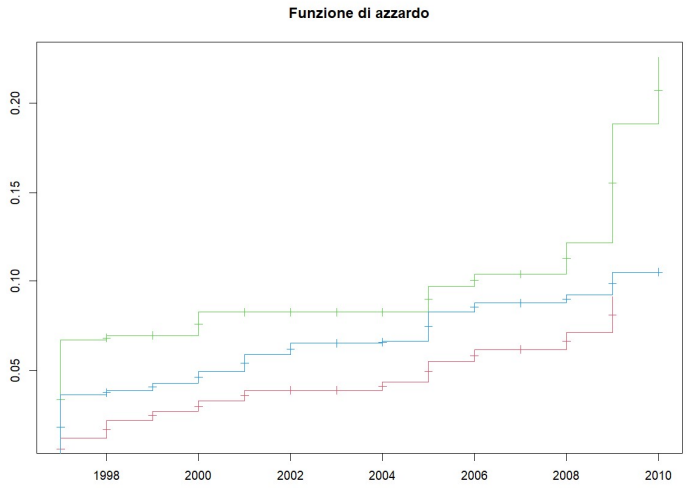


Figura 4: Funzione di rischio cumulato per la variabile TempCat.

La curva verde continua si riferisce alla temperatura “Hot”, la curva azzurra continua si riferisce alla temperatura “Mite”, mentre la curva rossa continua si riferisce alla temperatura “Cold”.

Dalla Figura 4 si può notare una pendenza crescente di tutte e tre le curve, indice di un aumento del rischio nel tempo. Questo indica un aumento dell'incidenza dell'evento di interesse. Inoltre, la curva verde presenta un’altezza maggiore rispetto alle altre due curve, si evince quindi una maggior probabilità che si verifichi l’evento studiato.

Log-Rank test

Applicando il test log-rank si ottiene che il p-value (0.01) risulta minore dell’alpha fissato = 0.05, quindi si rifiuta l’ipotesi nulla di uguaglianza delle funzioni di sopravvivenza e si conclude che vi è una differenza nella sopravvivenza tra i tre gruppi.

```
Call:
survdiff(formula = Surv(Year, dati$status) ~ dati$TempCat)

                  N Observed Expected (O-E)^2/E (O-E)^2/V
dati$TempCat=C  411       21     30.4     2.930     3.539
dati$TempCat=H  433       45     31.5     5.835     7.098
dati$TempCat=M 1919      135    139.1     0.121     0.402

 Chisq= 9.1  on 2 degrees of freedom, p= 0.01
```

Modello di Cox

Nell'analisi della variabile status si procede con la stima del modello di Cox. Nello schema seguente viene riportato il risultato ottenuto:

```
Call:
coxph(formula = Surv(dati$Year, dati$status) ~ dati$TempCat,
    data = dati, ties = "breslow")

  n= 2763, number of events= 201

                coef exp(coef) se(coef)     z Pr(>|z|)
dati$TempCatH 0.7315    2.0781   0.2646 2.765  0.00569 **
dati$TempCatM 0.3422    1.4081   0.2346 1.459  0.14470
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

              exp(coef) exp(-coef) lower .95 upper .95
dati$TempCatH     2.078     0.4812     1.237      3.49
dati$TempCatM     1.408     0.7102     0.889      2.23
```

Nell'output si trovano le stime dei coefficienti β, il loro valore esponenziale, l'errore standard e la significatività. Si nota che il coefficiente di TempCatH è significativo contro l'ipotesi nulla. Ne consegue che ad un aumento del coefficiente di TempCatH corrisponde un aumento del rischio che si verifichi una rivolta. Il coefficiente di TempCatH è in valore assoluto maggiore del coefficiente di TempCatM, questo sta ad indicare che TempCatH ha un effetto più forte sull'hazard ratio (exp(coef) = 2.0781) rispetto a TempCatM (exp(coef) = 1.4081), ovvero ad un aumento del rischio di evento.

Analisi dei residui

La Figura 5 illustra i residui di Cox-Snell calcolati per la covariata TempCat.

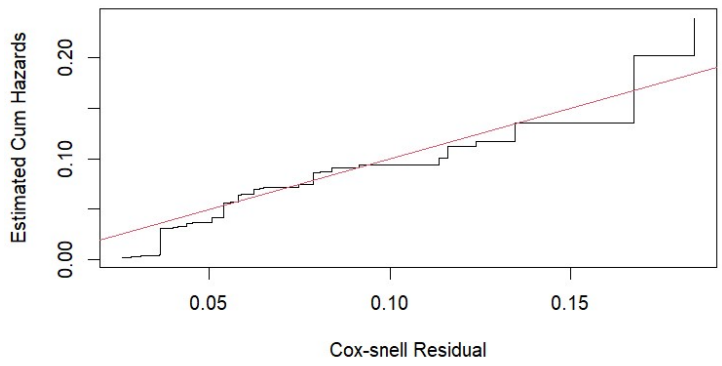


Figura 5: Residui di Cox-Snell per il modello di Cox.

L'andamento dei residui di Cox-Snell nel tempo fornisce ulteriori informazioni sulla bontà di adattamento del modello di Cox. I residui sono costanti nel tempo, ciò suggerisce che il modello di Cox si adatta in modo adeguato ai dati nel lungo periodo.

Con il test basato sui residui di Schoenfeld riportato nell'output sottostante si verifica l'indipendenza dal tempo dei coefficienti ipotizzata dal modello di Cox. Esso testa questa ipotesi per entrambe le variabili del modello ridotto e per il modello nel suo complesso.

```
             chisq df    p
dati$TempCat  1.77  2 0.41
GLOBAL        1.77  2 0.41
```

Si ottiene così la conferma che l'ipotesi di rischio proporzionale su cui si basa il modello di Cox non può essere respinta infatti, come si può vedere dall'output sotto riportato, nel complesso viene stimato un p-value GLOBAL = 0.41.

Considerazioni finali sull'analisi di sopravvivenza

Il modello che si adatta meglio ai dati è il modello di Cox, il tutto è confermato da opportune analisi grafiche e dall'analisi dei residui del modello.
La variabile EventType è formata da tre livelli: Riots (rivolte), Battles (battaglie) o Violence against civilians (atti di violenza contro i civili), l'avvenimento di questi tre tipi di conflitti è influenzato solamente dalla variabile TempCat, che indica il livello di temperatura (C = freddo, M = mite, H = caldo) assegnato sulla base delle temperature registrate dalla stazione StationID.

Da un'analisi preliminare condotta sui dati si è visto che l'evento maggiormente influenzato dal livello di temperatura è *Riots,* in particolare si evince che in corrispondenza delle zone più calde degli Satati presi in considerazione si verifica un'aumento delle rivolte.

Si procede dunque con l'analisi sulla geostatistica che, come ci si aspetta, si adatta meglio ai dati presenti in questi dataset.

ANALISI SULLA GEOSTATISTICA

Analisi esplorativa dei dati

Nell'analisi esplorativa viene svolta una preliminare descrizione delle informazioni date dalle variabili osservate.

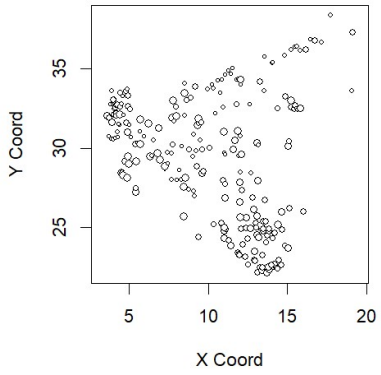


Figura 6: Grafico a cerchi (circle plot) delle osservazioni.

Dalla Figura 6 e 7 si può notare che le osservazioni si dispongono in maniera abbastanza casuale, sono presenti però due cluster in corrispondenza delle coordinate (5, 30) e (15, 25) circa. Inoltre, vi sono dei punti isolati in alto a destra. Questa struttura spaziale è probabilmente dovuta alle caratteristiche delle variabili presenti nel dataset che verranno studiate in seguito.

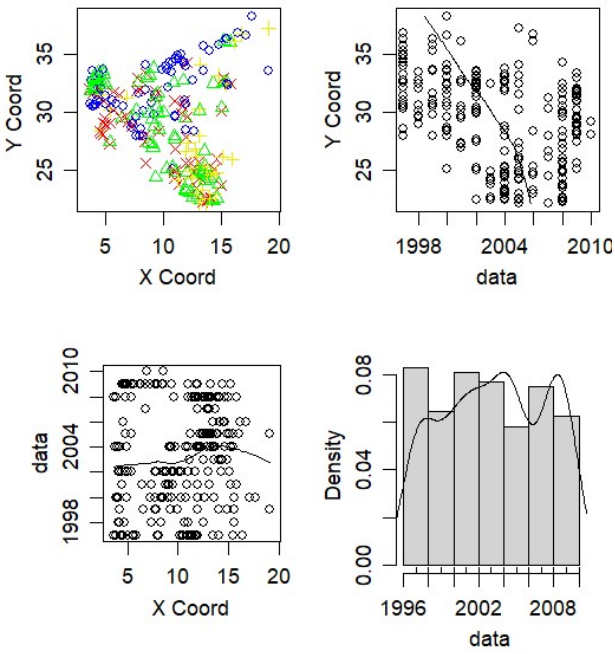


Figura 7: Informazioni sintetiche sui dati.

Variogramma

In Figura 8 viene riportato il variogramma empirico a nuvola

Osservando il grafico le semivarianze più alte tendono ad essere più concentrate a distanze ridotte indicando la presenza di una dipendenza spaziale ravvicinata.

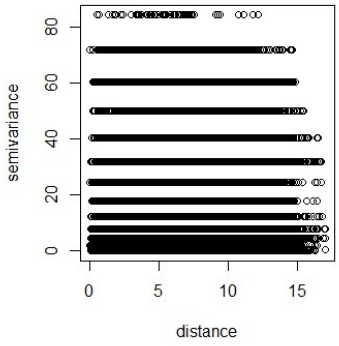


Figura 8: Variogramma empirico a nuvola dei dati.

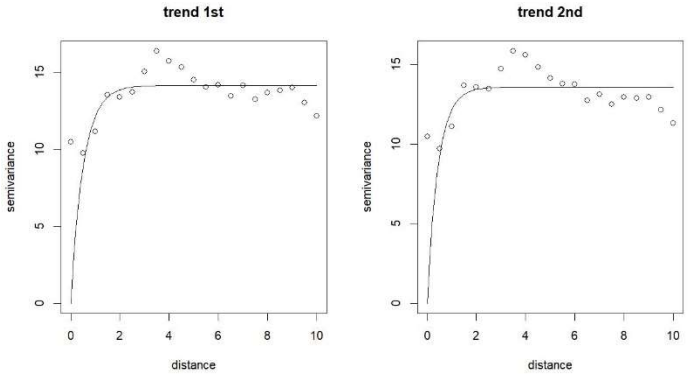


Figura 11: Grafici delle stime dei parametri di covarianza dei modelli parametrici sui variogrammi empirici.

Analisi dei residui del variogramma empirico

In Figura 9 vengono rappresentati i residui per trend, nel grafico a destra i punti neri sono per il trend di primo grado, mentre i punti rossi rappresentano il trend di secondo grado.

Dall'analisi condotta risulta preferibile utilizzare il trend di primo grado per rappresentare i dati, infatti l'andamento dei residui risulta molto simile, quindi non verrebbe giustificato l'over-fitting di utilizzare un modello con trend quadratico.

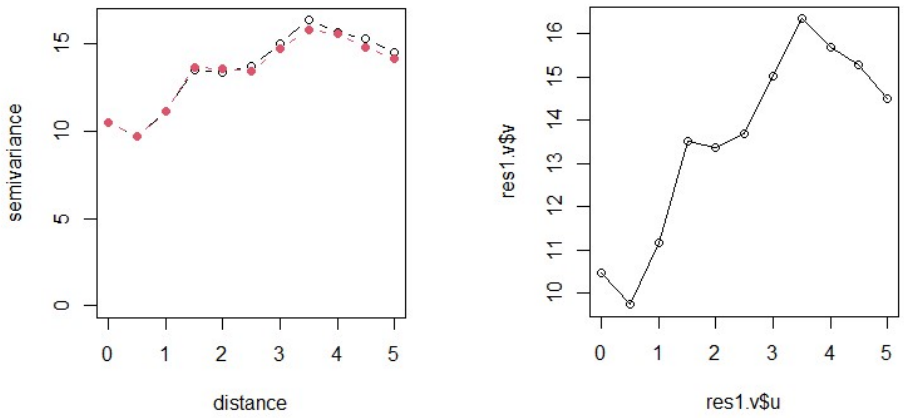


Figura 9: Studio dei residui dei dati geostatistici per trend.

Modelli parametrici

Nell'analisi delle variabili si procede con la stima del modello parametrico. Nello schema seguente viene riportato il risultato ottenuto:

```
Call:
lm(formula = datiG$data ~ datiG$coords[, 1] + datiG$coords[,
    2])

Residuals:
    Min      1Q  Median      3Q     Max
-6.8887 -2.7534 -0.3622  2.4913  7.4695

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.016e+03  2.052e+00 982.307  < 2e-16 ***
datiG$coords[, 1]  9.552e-03  6.187e-02   0.154    0.877
datiG$coords[, 2] -4.336e-01  6.101e-02  -7.107 1.37e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.642 on 238 degrees of freedom
Multiple R-squared:  0.185,    Adjusted R-squared:  0.1781
F-statistic:    27 on 2 and 238 DF,  p-value: 2.694e-11
```

Nell'output si trovano le stime dei coefficienti β, l'errore standard e la significatività . Si nota che il secondo coefficiente è significativo contro l'ipotesi nulla. Ne consegue che il cambiamento nella longitudine geografica influisce sul verificarsi dell'evento.

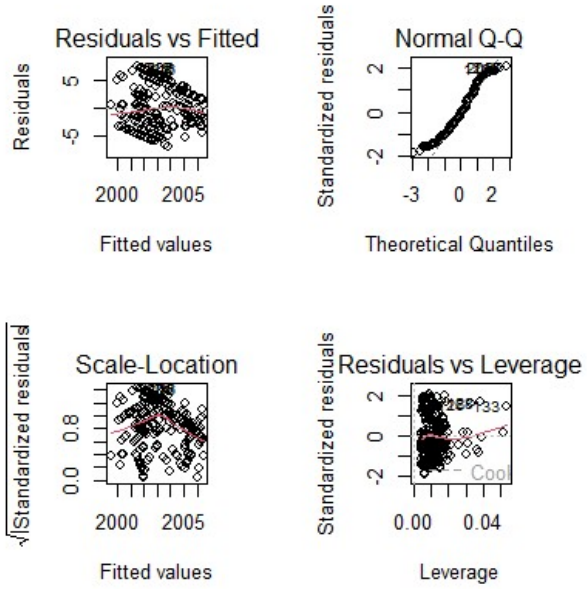


Figura 10. Informazioni sintetiche sul modello con trend lineare.

Dal diagramma quantile - quantile si può notare che tutto sommato i residui seguono l'andamento della retta di riferimento, indice che il modello si adatta bene ai dati.

RISULTATI

Dall'analisi sulla geostatistica emerge che la variabile di interesse EventType, formata dai tre livelli Riots (rivolte), Battles(battaglie) e Violence Against Civilians (atti di violenza contro i civili), viene influenzata dal cambiamento della coordinata riguardante la longitudine geografica.

Dalle due analisi condotte risulta quindi che la variabile EventType venga maggiormente influenzata dal livello di temperatura e dalla longitudine geografica.

BIBLIOGRAFIA E SITOGRAFIA

- Statistics for Spatial Data, N. Cressie, 2003
- Model-based Geostatistics, P. Diggle, P. Ribeiro, 2007
- SURVIVAL ANALYSIS: Techniques for Censored and Truncated Data, Klein & Moeschberger, 2003
- http://www.ricercasit.it/public/documenti/Dottorato/Formazione/Geostatistica_e-learning/Lezione_Noti_23set2011_slide_parte_c.pdf