ALCUNI METODI PER L'ANALISI DELLE SERIE STORICHE IN AGROMETEOROLOGIA

SOME METHODS FOR TIME SERIES ANALYSIS IN AGROMETEOROLOGY

Luigi Mariani

Università degli Studi di Milano, Dipartimento di Produzione Vegetale, Via Celoria 2, 20100 Milano MI
Corresponding author: e-mail luigi.mariani@unimi.it

Ricevuto 15 giugno 2006, accettato16 ottobre 2006

Riassunto

La presente nota tecnica descrive alcuni metodi di analisi di serie storiche. In particolare ad una disa-mina circa le problematiche connesse al recupero di serie storiche in Italia segue la descrizione delle attività preliminari di controllo di qualità (controlli di consistenza assoluta, relativa, spaziale e tempo-rale). A ciò seguono alcune considerazioni sui metodi di analisi statistica utili per verificare la stazio-narietà delle serie, caratteristica preliminare all'esecuzione di molti test statistici. In particolare vengo-no discusse l'analisi di discontinuità e l'analisi di autocorrelazione, presentando anche alcuni strumenti informatici di pubblico dominio utili per eseguire tali analisi.

Parole chiave: analisi statistica, serie storiche, stazionarietà, trend, discontinuità

Abstract

This technical note describes some methods for time series analysis. A preliminary discussion about the recovery of climatic time series for Italian area and quality check (check of different kinds of con-sistency: absolute and relative – spatial and temporal) is followed by a general description of some sta-tistical methods, useful in order to detect the stationarity of time series, a characteristic that is an as-sumption needed for the execution of many statistical tests. In particular are presented here some methods for analysis of discontinuity, autocorrelation and trend and procedures to perform these analysis using free software.

Keywords: *statistical analysis, time series, stationarity, trend, discontinuity*

Dati meteo-climatici e problemi connessi al loro recupero

La caratterizzazione climatica di un sito o territorio deve fondarsi su dati osservativi strumentali (e cioè rilevati con strumenti meteorologici) e sensoriali (frutto cioè di osservazioni "a vista"). Ad esempio i dati sulla visibilità o sulla copertura nuvolosa sono di norma frutto di osservazioni sensoriali mentre i dati di temperatura, umidità relativa o vento sono frutto di misure strumentali eseguite con specifici sensori.

Le **condizioni meteorologiche** di un sito o di una particolare area sono rappresentate dall'insieme delle grandezze meteorologiche che vi si determinano in un certo istante (la temperatura e l'umidità dell'aria ad una data altezza, le precipitazioni, la radiazione, il vento, la copertura nuvolosa, ecc.).

Per cogliere invece il clima di un particolare sito o territorio occorre considerare l'insieme delle condizioni meteorologiche su lunghi periodi di tempo, dell'ordine dei decenni. Il clima è dunque un'astrazione ottenuta elaborando statisticamente l'insieme delle variabili meteorologiche (es: dati giornalieri o orari) in modo da ricavarne opportuni indici climatici (valori medi ed estremi, valori cumulativi, frequenze, ecc.). Parliamo così di temperature o di precipitazioni medie ed estreme sull'anno, sul mese o sulla decade (per decade intendiamo un periodo di 10 giorni così ripartiti: 1–10, 11–20, 21–ultimo giorno del mese) o ancora sulla pentade (periodo di 5 giorni). "Spremendo" a fondo i dati di cui si dispone si può ottenere una notevole messe di indici statistici ed avere così una caratterizzazione sempre più approfondita del clima.

Secondo le normative dell'Organizzazione Meteorologica Mondiale il **clima attuale** di un certo sito o territorio e cioè la normale climatica da utilizzare come termine di paragone per valutare il livello di anomalia dei dati attuali, si ottiene elaborando statisticamente i dati degli ultimi 30 anni. Perché trent'anni e non 10 oppure 100? Questa scelta vuole sfuggire a due tipi di rischio:

1. quello di considerare periodi troppo brevi che non consentano di descrivere adeguatamente il clima, rischio tanto più rilevante quanto più le variabili in esame presentano un variabilità interannuale accentuata (es: copertura nuvolosa, precipitazioni);
2. quello di considerare periodi eccessivamente lunghi ricadendo così in errori legati a fluttuazioni del clima verificatesi nell'arco di tempo considerato.

Anche se più avanti si vedrà che anche il concetto di normale climatica può essere sottoposto ad un esame critico, si può in prima istanza affermare che chi è chiamato a caratterizzare il clima di un dato sito o territorio deve

Tab. 1 – Dati meteorologici di un giorno di aprile per una stazione della pianura padana occidentale in condizioni di tempo anticiclonico e confronto con il clima attuale

***Tab. 1** – Meteorological data of an April day with anticyclonic weather conditions for a station of the Po plain and match with present climate data.*

	Condizioni meteorologiche del 15 aprile 2000	***Valori climatici (trentennio 1961 – 1990) (seconda decade di aprile)***
Temperatura dell'aria	Massima: 22.5°C Minima: 11,7°C	Media delle massime: 18,5°C Media delle minime: 8,1°C (valori estremi del trentennio: +2,3 / +28,9)
Umidità relativa	Massima: 89% Minima: 52%	Media delle massime: 95% Media delle minime: 67% (valori estremi del trentennio: 6% / 100%)
Soleggiamento	7,5 ore di sole	Soleggiamento medio: 5,9 ore di sole
Vento	Velocità media: 1,2 m/s	Velocità media: 1,4 m/s Velocità massima: 19,8 m/s
Precipitazioni	0 mm	Nella seconda decade Valore medio di 22 mm su 2,6 giorni Valore massimo: 163 mm su 9 giorni
Evapotraspirazione da coltura di riferimento (ET0)	3,3 mm	Et0 media: 2,8 mm Et0 massima: 11,2 mm

recuperare grossomodo un trentennio di dati meteorologici. Posta tale premessa vediamo come egli debba operare per fare ciò. A questo proposito non possiamo nasconderci che l'Italia non è un Paese ben organizzato dal punto di vista della meteorologia. Ciò a causa della patologica molteplicità di soggetti coinvolti nelle attività di gestione delle reti osservative, fenomeno questo che affratella nella confusione tutta l'Italia, dimostrando al di sopra d'ogni dubbio quello che costituisce uno dei principali caratteri nazionali.

Aldilà di ogni considerazione connessa allo spreco di risorse – ogni rete comporta un proprio manipolo di gestori – e alla scarsa qualità dei dati – una seria validazione di dati così variegati è del tutto improponibile – emerge nettissima la complessità connessa al reperimento di dati meteorologici in Italia.

Come muoversi allora di fronte ad una tale realtà? Anzitutto occorre armarsi di moltissima pazienza e seguire la seguente procedura:

1. individuare le reti che posseggono dati per il territorio indagato;
2. contattare i gestori di tali reti ed acquisire indicazioni circa le modalità e il prezzo di rilascio dei dati. Circa il prezzo si va dalla gratuità completa a casi estremi di prezzi elevatissimi
3. verificare se i gestori pubblicano annuari cartacei ovvero informatizzati in cui siano raccolti i dati da loro prodotti;
4. attraverso i gestori acquisire indicazioni ulteriori circa altri eventuali detentori di dati che siano sfuggiti alla vostra prima ricognizione;
5. ricercare i dati disponibili su Internet. In particolare si segnalano i siti dell'Ufficio Centrale di Ecologia Agraria (www.ucea.it), quello del Servizio Meteorologico dell'Aeronautica (www.meteoam.it) e quelli dei Servizi meteorologici regionali, reperibili al sito www.agrometeorologia.it;
6. consultare gli annali idrologici del Servizio Idrografico nonché le serie storiche presenti sulla Rivista di Meteorologia Aeronautica:
7. consultare alcuni "testi sacri", vere scialuppe si salvataggio allorché non si sa più a che santo votarsi. In particolare un testo spesso criticato per la multiforme varietà delle fonti a cui l'autore ha attinto ma tuttavia essenziale è ancor oggi *Il clima d'Italia* di Cristofaro Mennella (1972).

Questa la realtà con cui ci confrontiamo tutti i giorni. Domandiamoci allora cosa dice la legge, ed in proposito è interessante segnalare il contenuto del DLgs n. 39, 1997, pubblicato sulla Gazzetta Ufficiale n.54 del 6/3/97 (con la solita utile postilla "È fatto obbligo a chiunque spetti di osservarlo e farlo osservare") e intitolato "Attuazione della direttiva 90/313/CEE concernente la libertà di accesso alle informazioni in materia di ambiente" (Ferrari, 1998).

Il Dlgs. è di diretto interesse per le "autorità pubbliche" quali Stato, Regioni, Amministrazioni locali, Enti pubblici o Concessionari di pubblici servizi che in qualche modo "producono" o sono in possesso di informazioni relative all'ambiente ed è molto chiaro. Questi i punti fondamentali:

- assicura a chiunque la libertà di accesso alle informazioni relative all'ambiente in possesso delle autorità pubbliche;
- definisce in modo chiaro cosa si deve intendere per "informazioni relative all'ambiente" (vi include esplicitamente anche quelle contenute nelle "basi di dati") e cosa si deve intendere per "autorità pubbliche";
- ribadisce che "le autorità pubbliche sono tenute a rendere disponibili le informazioni a chiunque ne faccia richiesta";
- definisce i casi di esclusione dal diritto di accesso quali ad esempio il caso di materiale fornito da terzi non tenuti giuridicamente a fornirlo oppure il caso in cui dalla divulgazione possano derivare danni all'ambiente stesso, oppure alle relazioni internazionali, alla difesa nazionale, alla sicurezza pubblica, ad inchieste o azioni investigative preliminari, alla riservatezza commerciale e industriale o alla riservatezza di dati personali. La sottrazione all'accesso dei dati o il suo differimento è regolata in modo preciso, con esplicita motivazione: può ad esempio essere rifiutato "quando la generica formulazione della richiesta non consente l'individuazione dei dati da

mettere a disposizione". Il DLgs fissa in 30 giorni. la conclusione del procedimento di accesso. "Trascorso inutilmente detto termine la richiesta si intende rifiutata".

- "Il diritto di accesso consiste nella possibilità di duplicazione o di esame delle informazioni". Per le modalità e le forme di accesso si applica il DPR. n.352, 27/6/92 che a sua volta esclude ben pochi casi (segreto di stato, sicurezza e difesa nazionale, politica monetaria e valutaria, ordine pubblico, prevenzione e repressione della criminalità, privacy). "La visione e l'esame delle informazioni deve essere disposta a titolo gratuito" mentre "il rilascio di copie è subordinato al rimborso dei costi relativi alla riproduzione".

Controlli di consistenza e la ricostruzione di dati errati o mancanti

L'analisi di una serie storica di dati dev'essere anzitutto preceduta dalla **verifica di consistenza** (correttezza) dei dati disponibili. I dati dovranno essere sottoposti ai controlli qui di seguito descritti e ad ogni dato controllato dovrebbe essere poi associata una etichetta di validità del tipo: 0 = dato in attesa di validazione; 1= dato validato; 2 = dato sospetto; 3 = dato errato.

Nel rinviare a Rana *et al.* (2004) per un'analisi di dettaglio sulle regole di validazione dei dati meteorologici, si ricorda l'opportunità di distinguere la consistenza assoluta da quella relativa, che a sua volta potrà essere di tipo temporale o spaziale.

La verifica di **consistenza assoluta** prevede in genere i seguenti controlli:

1. che i dati esaminati ricadano all'interno dei range strumentali (che potrebbero ad esempio essere di –20/+50 °C per i termometri e di 0/100 % per gli igrometri)
2. che i dati ricadano entro i range previsti dalla climatologia del luogo di installazione per quel particolare periodo dell'anno (ad esempio in pianura padana l'analisi di lunghe serie storiche ci dice che è da escludere la discesa delle temperature minime al di sotto degli 0 °C in luglio mentre una discesa al di sotto della soglia dei 10 °C non è da escludere pur rappresentando un fenomeno che si verifica con una frequenza dell'ordine di un caso ogni 8-10 anni, in genere a seguito di violente manifestazioni temporalesche innescate da irruzioni di aria artica).

La verifica di **consistenza relativa** è costituita invece dal confronto con dati prodotti dalla stessa stazione (consistenza temporale) ovvero con dati prodotti da stazioni climaticamente simili (consistenza spaziale).

La consistenza temporale prevede regole del seguente tipo: una stazione la cui temperatura media del giorno n è stata di 20 °C non può presentare una temperatura media del giorno successivo inferiore a 10 °C ovvero superiore a 30 °C (si tratta di valori del tutto indicativi, da sostituire con valori ricavati dall'analisi delle serie storiche di riferimento per la stazione in esame).

Il controllo di consistenza spaziale si baserà invece su regole del tipo: se la stazione B ha presentato una temperatura media del giorno n di 35 °C, la stazione A, posta in zona climaticamente omogenea rispetto a B, non potrà presentare una temperatura media inferiore a 30 °C.

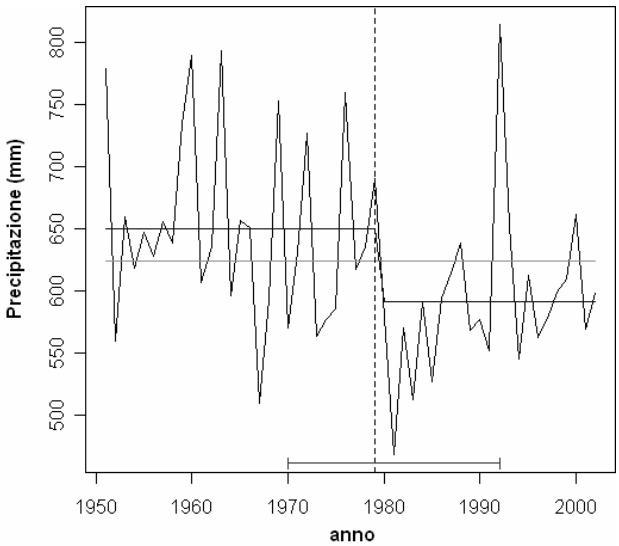

Fig. 1 – Andamento delle precipitazioni sul Mediterraneo centro occidentale per il periodo 1951-2003. La linea tratteggiata mostra il change point stimato (si tratta del 1978). La linea orizzontale più in basso, i cui limiti sono il 1970 ed il 1992, indicano l'intervallo di confidenza al 90% della ripartizione della serie in due segmenti. Le tre linee orizzontali sovrapposte alla spezzata che collega i valori rappresentano rispettivamente la media dell'intero periodo e le medie del periodo che precede e che segue il change point.

***Fig. 1** – Behavior of precipitation for the Central Occidental Mediterranean (1951-2003 period). Dotted line shows a change point in 1978. The lowest horizontal line, witch limits are 1970 and 1992, shows the confidence range of 90% for the change point. The other 3 horizontal lines superimposed to the diagram represent respectively the mean of the whole period and the means for the period before and after the change point.*

Come già visto nel caso della consistenza assoluta, è cruciale la definizione di un set di regole specifiche da individuare per ogni singola stazione ovvero per ogni area climaticamente omogenea attraverso l'analisi di serie storiche sufficientemente lunghe.

Ultimati i controlli di consistenza potrebbe risultare necessario procedere alla **ricostruzione dei dati errati** (ed eventualmente di quelli sospetti) in vista delle elaborazioni. Per fare ciò sarà importante far ricorso a procedure geostatistiche operanti a partire da dati di stazioni climaticamente omogenee ovvero da dati delle stessa stazione che precedano o seguano quello errato.

Stazionarietà e analisi statistica

L'analisi delle serie storiche non è un'attività banale e dev'essere fondata su metodi statistici rigorosi. Scopo di questo paragrafo è quello di introdurre il lettore a tale tematica, rimandandolo a testi specifici ovvero al supporto di specialisti della statistica per i necessari approfondimenti.

Una serie storica è statisticamente **stazionaria o omogenea** quando presenta una costanza dei momenti ed in particolare della media e della varianza (Munn, 1970). La

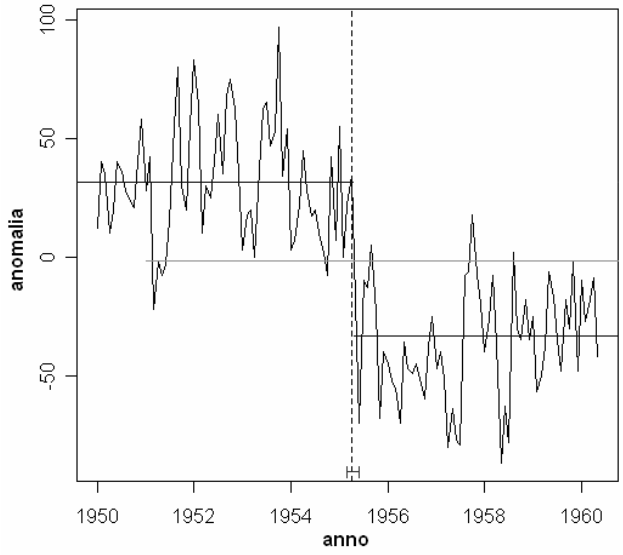

Fig. 2 – Andamento dell'anomalia di altezza del geopotenziale a 200 hPa a Hong Kong per il periodo 1950-1996 (per il significato delle linee che compaiono nel grafico si veda il commento a figura 1). Il change point individuato appare definito in modo assai netto, come mostra l'intervallo di confidenza (in riferimento a limiti di confidenza del 90% il breakpoint ricade fra marzo e giugno 1955) . I dati cui è applicato l'esempio sono tratti da Lanzante (1996).

Fig. 2 *– Behavior of the anomaly of geopotential height at 200 hPa for* Hong Kong - *1950-1996 period (the meaning of the different lines is explained in figure 1). Change point is very well identified, with a confidence level of 90% between march and june 1955. The data used for this example are reported by Lanzante (1996)*

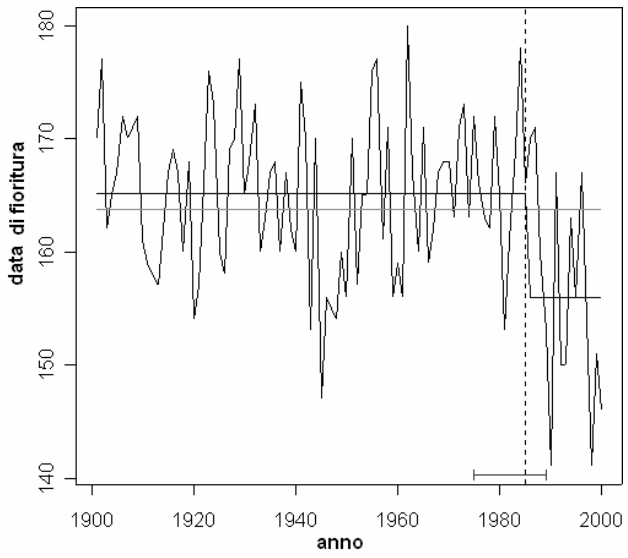

Fig. 3 – Serie storiche delle date giuliane (1..366) di fioritura del sambuco a De Bilt (Olanda) per il periodo 1901-2000 (per il significato delle linee che compaiono nel grafico si veda il commento a figura 1). La discontinuità ricade nel periodo 1975-1989 (assumendo una limiti di confidenza del 90%) e l'anno più probabile di discontinuità è il 1985.

Fig. 3 *– Time series of flowering of Sambucus nigra at De Bilt - The Netherlands for 1901-2000 (the meaning of the different lines is explained in figure 1). Change point, considering a confidence level of 90% is between 1975 and 1989; the most probable year of discontinuity is 1985.*

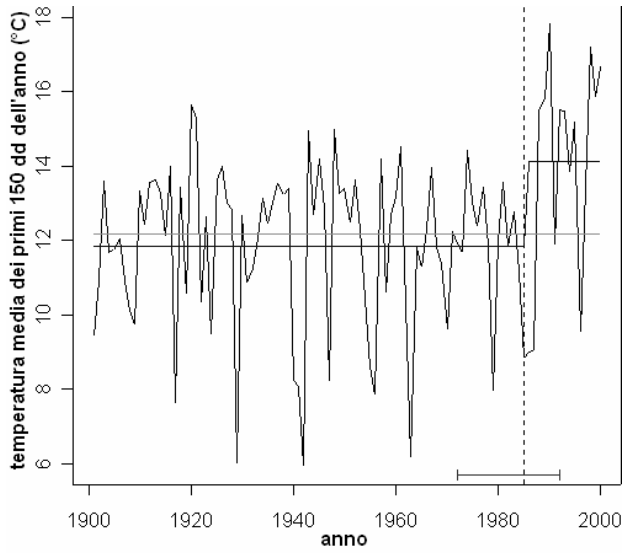

Fig. 4 – Serie storiche delle temperature medie dei primi 150 giorni dell'anno a De Bilt (Olanda). questa discontinuità ricade fra 1972 e 1992 (assumendo una limiti di confidenza del 90%). Come per la fioritura del sambuco (figura 3), l'anno più probabile di discontinuità è il 1985.

Fig. 4 *– Time series of mean temperature of first 150 days of each year at De Bilt - The Netherlands for 1901-2000 (the meaning of the different lines is explained in figure 1). Change point is well identified, with a confidence level of 90% between 1972 and 1992. The most probable year of change point is 1985, the same year obtained for the flowering of Sambucus nigra in figure 3.*

stazionarietà si verifica di rado in geofisica ed in proposito Lamb (1966) suggeriva di sostituire consueto il concetto di "normale climatica" enunciato nel primo paragrafo di questa nota e da lui ritenuto non appropriato, con quello di "valore medio per un periodo dato", periodo che andrebbe ovviamente sempre indicato. Alla luce di ciò non è sempre vero che una serie trentennale sia quella più efficace in termine di serie di riferimento con cui confrontare i dati attuali. Ad esempio per aree soggette a crescente urbanizzazione può risultare più rappresentativa la serie degli ultimi 10 anni (Munn, 1970), così come a fronte del cambiamento climatico che ha interessato l'area europea nella seconda metà degli anni '80 del 20° secolo per effetto di un brusco cambio di regime delle grandi correnti occidentali (Werner, 2000) e che si è tradotto nel brusco aumento delle temperature si dovrebbe mirare ad utilizzare come serie di riferimento quelle successive a tale cambiamento climatico.

Il trattamento statistico delle serie storiche si rivela allora fondamentale per porre in evidenza la stazionarietà delle serie**,** la quale è una delle più fondamentali assunzioni che stanno alla base di molti test statistici.

Fra le cause di non stazionarietà discuteremo in questa sede i trend, le discontinuità e le varianze non costanti.

Un trend è dato dalla presenza di medie non stazionarie. E' questo il caso che deriva da cambiamenti nel sito di installazione dei sensori ovvero da modifiche nell'uso

del suolo della zona circostante i sensori; tale fenomeno è particolarmente grave nel caso degli osservatori storici sorti fra il 1700 e il 1800 in aree urbane che poi nel tempo si sono trasformate, intensificando alcuni effetti quali la cosiddetta "isola di calore". Si pensi inoltre all'effetto di incremento delle temperature prodotto dalla costruzione di una strada asfaltata in vicinanza di una stazione ovvero dalla sostituzione del tappeto erboso (su cui secondo le normative internazionali devono essere installate le stazioni meteorologiche) con ghiaietto ovvero dal trattamento periodico del tappeto erboso stesso con un disseccante.

Altra causa di trend è la deriva strumentale: svariati sensori in assenza di manutenzione tendono a deteriorare gradualmente la loro accuratezza; assai noto a tale proposito è il caso di taluni igrometri elettronici.

Fra le cause di discontinuità in una serie storica climatica possiamo rammentare ad esempio i cambiamenti nella tecnica osservativa ovvero la sostituzione della strumentazione di misura per obsolescenza tecnologica o per altri motivi; un caso da manuale è costituito dal passaggio dai tradizionali schermi antiradiazione per termometri e termografi (le cosiddette capannine meteorologiche) alle più moderne cupoline dei termometri elettronici da cui sono derivate discontinuità che è frequente osservare nelle serie.

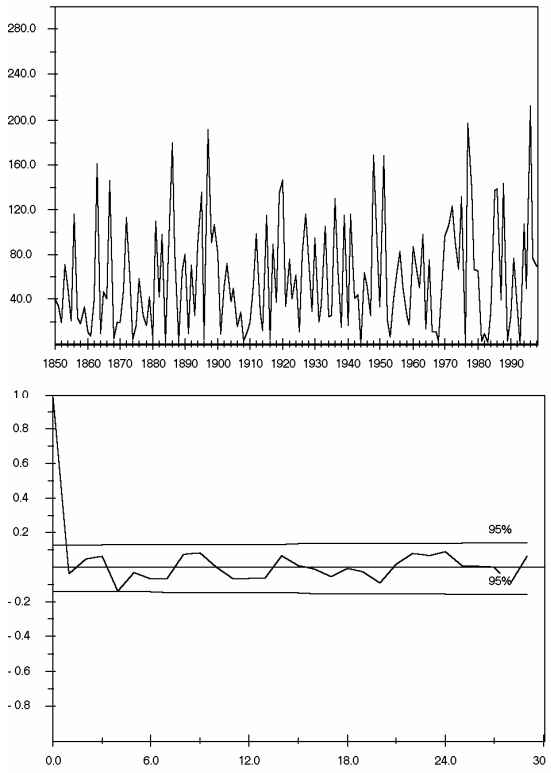

Fig. 5 –Analisi di autocorrelazione della serie delle precipitazioni mensili del mese di gennaio per l'osservatorio di Milano Brera. In alto si riporta la serie storica ed in basso il grafico dell'autocorrelazione per lag massimo di 29 anni.

Fig. 5 *– Autocorrelation analysis for monthly total precipitation of January at Milano Brera observatory. The upper figure show the time series; the lower one the autocorrelation diagram for a maximum lag of 29 years.*

Un esempio di non stazionarietà causata dalla non costanza della varianza è dato ad esempio da serie storiche che presentino medie costanti e deviazioni standard decrescenti nel tempo, fenomeno questo che potrebbe ad esempio essere causato da cambi di strumentazione (si pensi ad esempio al passaggio da strumentazione più sensibile a strumentazione meno sensibile ovvero da cambi di localizzazione delle stazioni meteorologiche).

Dovendo valutare la presenza di non stazionarietà in una serie si consiglia sempre di effettuare un'**ispezione visuale** preliminare e a seguito di questa di scegliere il metodo statistico più adatto ad analizzare quantitativamente il fenomeno.

L'individuazione delle discontinuità

L'individuazione della **discontinuità** nelle serie storiche è affidata ad appositi test statistici che nella bibliografia anglosassone sono indicati come *change point tests*. Test largamente diffusi per l'individuazione di discontinuità sono il test EP, recentemente proposto da Easterling e Peterson (1995), il test L (Siegel e Castellan, 1988) ed il test proposto da Bai (1994).

Qui di seguito si riportano alcuni esempi di analisi statistica di una serie storiche con individuazione di discontinuità.

La prima serie storica proposta è quella della precipita-

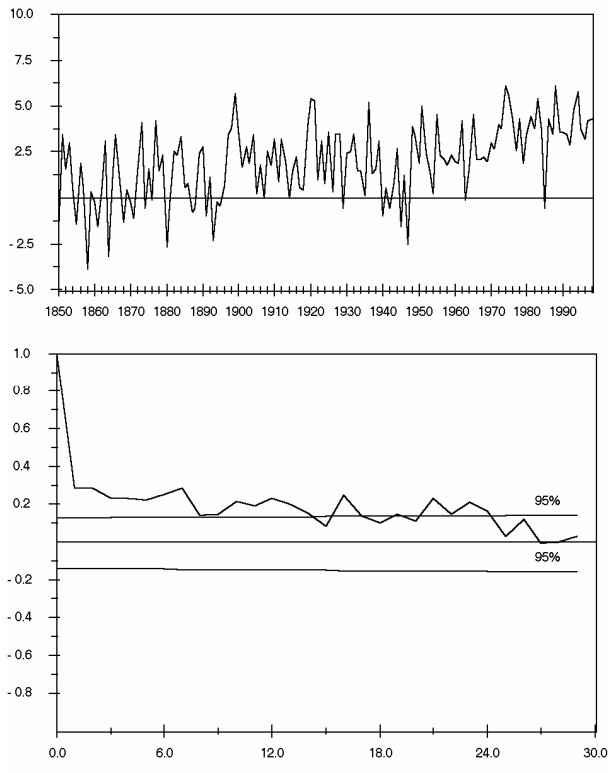

Fig. 6 – La figura presenta l'analisi di autocorrelazione della serie delle temperature mensili del mese di gennaio per l'osservatorio di Milano Brera. In alto si riporta la serie storica ed in basso il grafico dell'autocorrelazione per lag massimo di 29 anni.

Fig. 6 *– Autocorrelation analysis for monthly mean temperature of January at Milano Brera observatory. The upper figure show the time series; the lower one the autocorrelation diagram for a maximum lag of 29 years.*

zione media sul Mediterraneo Centro Occidentale per il periodo 1950-2002 [i dati per il periodo 1950-86 sono tratti da Colacino e Piervitali (1998) mentre i dati dal 1987 al 2002 sono stati ricavati dal dataset del Global Precipitation Climatology Project (http://wwwo.dwd.de/research/gpcc)].

L'analisi statistica è stata condotta con il software statistico di pubblico dominio R; l'idea di impiegare R deriva dal fatto che si tratta di un software disponibile gratuitamente in rete (per il download del software e della relativa documentazione si rinvia al sito ed il http://cran.r-project.org). R consente l'impiego di un vastissimo, ben documentato e costantemente aggiornato set di funzioni statistiche che permettono la risoluzione di una vastissima gamma di problemi.

L'individuazione dei change points è stata condotta impiegando l'algoritmo di analisi di change point presente nella libreria STRUCCHANGE di R e descritto in Zeileis *et al.* (2003). Tale algoritmo si fonda sul metodo di stima delle discontinuità in modelli di regressione su serie storiche descritto in Bai (1994) ed esteso all'individuazione simultanea di change points multipli da Bai (1997ab) e Bai & Perron (1998). La funzione di distribuzione utilizzata per gli intervalli di confidenza è descritta in Bai (1997b).

La sequenza di istruzioni R impiegata è descritta qui di seguito:

Sequenza operativa in R (i commenti appaiono dopo il cancelletto #):

1. attivare R
2. dal menu a tendina del programma caricare la libreria STRUCCHANGE (se la libreria non fosse disponibile è possibile scaricarla dal sito di R, applicando una semplice procedura guidata dalle opzioni del menu)
3. fornire i dati con la funzione di gestione delle serie storiche ts -> ricordare che il separatore accettato da ts è la virgola e che è cruciale la presenza di "c")

con un semplice "taglia/incolla" trasferire nell'ambiente R per l'esecuzione la seguente sequenza di istruzioni (eseguire dapprima PARTE1 e PARTE2 ed infine eseguire PARTE3 dopo aver parametrizzato correttamente il numero di breakpoint in conformità al numero individuato in PARTE2; ciò si ottiene agendo sull'istruzione "breaks=1")

```
#PARTE 1 - lettura dati
require(ts)
#precipitazioni Mediterraneo centro occidentale
(1951-2003)

ts1<-
ts(c(779,560,660,618,647,628,656,639,735,790,60
7,636,794,596,657,651,510,593,753,570,631,727,5
64,577,586,760,617,636,691,579,468,570,513,591,
527,594,615,639,568,577,552,815,653,545,613,563
,578,599,610,662,569,598),start=1951,frequency=
1)

# PARTE 2 - far prima girare questa parte per
vedere quanti breakpoints sono ipotizzati

if(! "package:stats" %in% search()) library(ts)
plot(ts1)
## F statistics indicate n breakpoint
fs.ts1 <- Fstats(ts1 ~ 1)
plot(fs.ts1)
breakpoints(fs.ts1)
lines(breakpoints(fs.ts1))

# PARTE 3 - far girare questa parte proponendo
il n° di breakpoints individuati nella parte 2
(per modificare tale numero rispetto a quello
qui indicato che è 1, agire sull'istrizione "
breaks = 1"

## fit null hypothesis model and model with n
breakpoint

#null hypothesis - unsegmented model
fm0 <- lm(ts1 ~ 1)
#segmented hypothesis -> model with n break-
points
fm1 <- lm(ts1 ~ breakfactor(bp.ts1, breaks =
1)) #marc1
plot(ts1)
lines(ts(fitted(fm0), start = 1951),col = 3)
#marc1
lines(ts(fitted(fm1), start = 1951), col = 4)
#marc1
lines(bp.ts1)
## confidence interval
ci.ts1 <- confint(bp.ts1,breaks=1,level=0.90)
#marc1
#breakdates(ci.ts1)
ci.ts1
#plot(ts1)
lines(ci.ts1)
```

Con tale sequenza di istruzioni è stato ottenuto il diagramma di figura 1, nel quale sono riportate anche le medie del periodo che precede (650 mm) e quelle del periodo che segue il cambiamento climatico (590 mm). E' intuibile che un'analisi di anomalia riferita ad una media piuttosto che all'altra conduce a risultati diversi. Con la stessa sequenza di istruzioni è stato anche analizzato l'andamento temporale dell'anomalia della pressione a 200 hPa ad Hong Kong per il periodo 1950-1996. Si tratta dello stessa serie storica analizzata da Lanzante (1996) con il test L e qui riproposta con l'ausilio di Strucchanghe, con risultati molto simili.

Si riporta infine l'analisi di discontinuità della serie fenologica di fioritura del sambuco per la località di De Bilt in Olanda pubblicata da van Vliet *et al.* (2002). Tale discontinuità è frutto del cambiamento climatico della seconda metà degli anni '80, in precedenza discusso e che per De Bilt, località posta proprio nella parte centrale dell'alveo delle grandi correnti occidentali, si è manifestato con un aumento delle temperature medie dei primi 150 giorni dell'anno (figura 4).

Di fronte a serie che presentino una discontinuità è consigliabile trattare separatamente il periodo a monte rispetto a quello a valle della discontinuità; è anche possibile eliminare la discontinuità introducendo un fattore di correzione delle serie a monte piuttosto che di quella a valle.

L'individuazione di trend

Anche l'individuazione di **trend** nelle serie storiche si basa su appositi test statistici. Una indicazione di massima circa l'esistenza di trend può venire anzitutto dallo studio dell'autocorrelazione per lag successivi. Data una serie $X_1..X_n$ per autocorrelazione si intende la correlazione esistente fra Xi e Xi+j ove j è il lag (j=0, 1, 2, 3..). E' ovvio che per j=0 avremo correlazione =1 e che se gli Xi sono numeri casuali tutti i coefficienti di correlazione per lag maggiori di 0 saranno pari a 0. Le serie storiche prive di trend presentano autocorrelazioni che si avvicinano a 0 al crescere del lag (lag che comunque non dovrebbe eccedere il 20-30% della lunghezza complessiva della serie) mentre al contrario le serie affette da trend presenteranno valori successivi nella serie in media superiori rispetto ai precedenti, il che si tradurrà in valori di autocorrelazione diversi da 0.

L'esempio di analisi di autocorrelazione di serie storiche qui proposto si riferisce alla serie storica 1850-1998 delle temperature medie e delle precipitazioni del mese di gennaio per l'osservatorio di Milano Brera (Chlistovsky *et al.*, 1997 e 1999) . In figura 5 si riporta il grafico del livello di autocorrelazione della precipitazione mensile ricavato per un lag massimo di 29 anni. Il fatto che al crescere del lag l'autocorrelazione rientri rapidamente all'interno del livello di confidenza del 95% fa propendere per la stazionarietà della serie. Al contrario nel caso delle temperature (figura 6) si osserva il persistere un livello di autocorrelazione superiore allo 0 e al di fuori dei limiti di confidenza del 95%, il che fa propendere per la presenza di un trend crescente legato all'incremento dell'isola di calore di Milano.

L'analisi è stata eseguita utilizzando il software freeware per l'analisi di serie storiche ANCLIM, realizzato da Petr Stepanek, climatologo presso lo Czech Hydrometeorological Institute e disponibile in rete al sito http://www.sci.muni.cz/~pest.

Per individuare trend monotonici in serie storiche climatologiche è spesso utilizzato il test di Mann – Kendall per i Trend (MKT) (Sneyers, 1998). Il test, suggerito da Mann (1945) è stato in seguito ampiamente utilizzato per analisi di serie storiche ambientali. L'algoritmo su cui si fonda il test è descritto ad esempio da Hipel and McLeod (2005).

Il test MKT è un'applicazione particolare del test di Kendall per la correlazione, anche noto come Kendall t test. Si tratta di un test non parametrico, in quanto non assume alcuna distribuzione a priori per i dati il che porta di solito ad una maggior robustezza rispetto a metodi parametrici.

In MKT l'ipotesi nulla (H0) è quella secondo cui i dati provengono da una popolazione in cui i dati sono indipendenti e identicamente distribuiti. L'ipotesi alternativa (H1) è invece che i dati seguano nel tempo un trend monotonico (positivo o negativo).

Nel caso del test MKT presente nella libreria Kendall di R il p-value indica il livello di rischio che si ha rifiutando l'ipotesi H0 per cui ad esempio un p-value di 0.03 indica una probabile presenza di trend. Infatti rifiutando H0 (l'assenza di trend) si ha una probabilità del 3% che si sia fatto erroneamente tale rifiuto. In altre parole H0 viene respinta (e quindi i dati presentano un trend) e il rischio di sbagliare respingendo H0 è solo del 3%. Per valutare il segno del trend si utilizza l'indice tau, per cui in presenza di un tau positivo il trend sarà positivo e viceversa nel caso di un tau negativo.

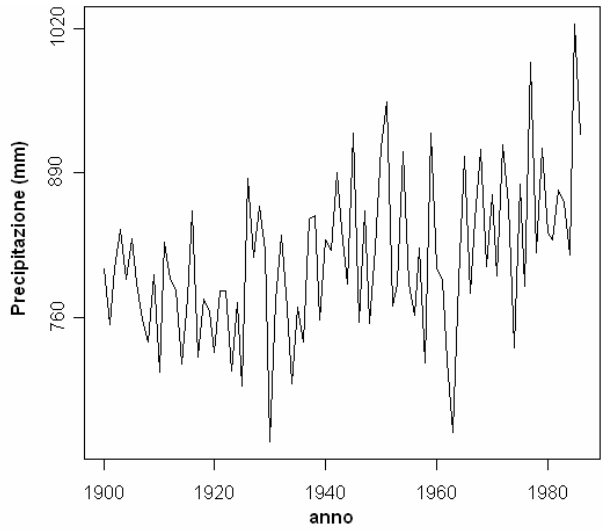

Fig. 7 – Serie delle precipitazioni annue dell'area dei grandi laghi americani per il periodo 1900-1986 (McLeod, 2005).
***Fig. 7** – Time series of yearly total precipitation for Greater Lakes of North America – period 1900-1986 (McLeod, 2005)*

Il listato di R qui di seguito riportato e per la cui esecuzione in R è necessario il caricamento della libreria Kendall[1], indica l'analisi della precipitazione media dell'area americana dei Grandi Laghi per il periodo 1900-1986 di cui al file PrecipGL allegato alla libreria Kendall stessa (McLeod, 2005) e riportata nel grafico in figura 7. Si noti che tale grafico viene prodotto dal software qui presentato applicando la tecnica LOWESS (Locally Weighted Scatterplot Smooting) sviluppata da W.S. Cleveland (1979) come estensione della tecnica LP (weighted Local Polynomial smooting), a sua volta estensione dello smooting a medie mobili.

```
# Annual precipitation entire Great
Lakes
# The time series plot with lowess
smooth suggests an upward trend
# The autocorrelation in this data
does not appear significant.
# The Mann-Kendall Trend MKT test con-
firms the upward trend.
data(PrecipGL)
plot(PrecipGL)
lines(lowess(time(PrecipGL),PrecipGL),
lwd=3, col=2)
# autocorrelation function
```

[1] *La libreria Kendall è richiamabile dal menu a tendina di R; qualora non fosse disponibile è possibile scaricarla dal sito di R, applicando una semplice procedura guidata dalle opzioni del menu)*

```
acf(PrecipGL)
#test MKT
MannKendall(PrecipGL)
```

Il risultato ottenuto è il seguente:

tau = 0.265
2-sided pvalue =0.00029206

Ove l'indice tau indica che il trend è positivo ed il p-value molto basso indica che l'ipotesi H0 va respinta e quindi siamo in presenza di trend (con un rischio di affermare erroneamente l'esistenza di trend ben inferiore allo 0.1%)
Come nel caso delle discontinuità il trend, una volta individuato, può essere rimosso effettuando un' interpolazione lineare o curvilinea della serie dei dati e sottraendo o addizionando ad ogni valore una quantità appropriata in modo da generare una nuova serie omogenea.

Le tecniche di regressione

Una volta individuata la presenza di discontinuità o trend può essere utile ricavare un modello a partire dai dati. Un modo assai usato per ricavare modelli è quello di utilizzare tecniche di regressione che, nel caso di serie temporali, saranno regressioni dei valori della serie rispetto al tempo; per una analisi delle tecniche di regressione si rinvia al capitolo 24 del testo di Hipel e McLeod (2005).
Le regressioni consentono di ottenere un modello dei dati in forma di equazione matematica da utilizzare anche per interpolazioni o estrapolazioni di dati. Tale equazione consente altresì di cogliere visivamente l'andamento temporale di serie storiche limitando l'effetto di disturbo dovuto a fluttuazioni di breve periodo.
Le regressioni sono tecniche parametriche di largo impiego e che tuttavia devono essere utilizzare a ragion veduta per evitare di far dire ai dati cose che di per sé non dicono. Si prenda ad esempio il caso della serie storica delle precipitazioni sul Mediterraneo di cui alla figura 1. La discontinuità mostra due fasi climatiche distinte, ognuna delle quali manifesta una sufficiente stazionarietà dei valori di precipitazione. In tal caso due regressioni lineari dovrebbero essere impiegate in riferimento alla due fasi distinte e non all'intera serie, pena la generazione di un modello che evidenzia precipitazioni monotonicamente decrescenti e che potrebbe portarci ad estrapolazioni irrealistiche per il futuro.
Una volta ricavato il modello di regressione è possibile valutarne le prestazioni ricavando il coefficiente di correlazione R^2 che rappresenta la percentuale di variabilità totale che il modello riesce a descrivere. In genere possiamo dire che tanto più R^2 è elevato e tanto più il modello descrive la variabilità i dati. Tuttavia anche un modello con un R^2 elevato potrebbe essere statisticamente inadeguato ed in particolare:

- la bontà del modello potrebbe non essere uguale su tutto il range delle osservazioni, il che si manifesta con una irregolare distribuzione dei residui;
- l'inferenza statistica eseguita su una regressione calcolata con i minimi quadrati richiede la normalità per produrre P values corretti;
- il modello potrebbe essere "guidato" dalla presenza di punti troppo influenti sul risultato.

Nel software R il metodo generalmente usato per ricavare un modello lineare è lm(), il cui compito è in sostanza quello di rispondere alla domanda su quale sia il miglior modello lineare in relazione alle osservazioni (Rossiter, 2006).
Può essere utile inoltre valutare la significatività della regressione il che nel software R e nel caso di regressioni lineari è possibile tramite la funzione cor.test().
Per meglio cogliere l'andamento temporale di serie storiche eliminando l'effetto di disturbo dovuto a fluttuazioni di breve periodo è possibile applicare altri metodi quali le medie mobili oppure le medie mobili pesate, in cui cioè il peso più elevato sarà assegnato ai valori più vicini nel tempo a quello da ricostruire.

Conclusioni

La presente nota tecnica, ben lungi dall'avere pretese di esaustività su un argomento tanto ampio, ha illustrato alcuni metodi di analisi di serie storiche finalizzati all'individuazione di discontinuità e trend.
L'applicazione di software di pubblico dominio rende l'uso di tali metodi relativamente immediato e ci si augura pertanto che tale approccio possa risultare interessante per i colleghi.

Ringraziamenti

L'autore ringrazia uno dei referee per le utilissime indicazioni che gli hanno consentito di dare coerenza ai concetti statistici espressi in questo lavoro.

Bibliografia

Bai J., 1994. Least Squares Estimation of a Shift in Linear Processes, Journal of Time Series Analysis, 15, 453-472.
Bai J., 1997a. Estimating Multiple Breaks One at a Time, Econometric Theory, 13, 315-352.
Bai J., 1997b. Estimation of a Change Point in Multiple Regression Models, Review of Economics and Statistics, 79, 551-563.
Bai J., Perron P., 1998. Estimating and Testing Linear Models With Multiple Structural Changes, Econometrica, 66, 47-78.
Bai J., Perron P., 2003. Computation and Analysis of Multiple Structural Change Models, Journal of Applied Econometrics, 18, 1-22.
Chlistovsky F., Buffoni L., Maugeri M., 1997. La temperatura a Milano Brera, CUSL, Milano, 176 pp.
Chlistovsky F., Buffoni L., Maugeri M., 1999. La precipitazione a Milano Brera, CUSL, Milano, 192 pp.
Cleveland W. S., Loader C. L. 1996 Smoothing by Local Regression: Principles and Methods. In W. Hardle and M. G. Schimek, editors, Statistical Theory and Computational Aspects of Smoothing*, pages 10–49. Springer, New York.*
Colacino M., Piervitali E., 1998. Segnali di cambiamento climatico nel Mediterraneo Centro Occidentale nel secondo dopoguerra, in Atti del Convegno Due secoli di osservazioni meteorologiche a Mantova, ERSAL, 47-58
Ferrari P., 1997. Quanto costa un dato? La direttiva per il libero accesso alle informzioni ambientali, Rivista Italiana di Agrometeorologia, anno 1, n.3, novembre 1997, pp. 3-4.
Easterling, D.R.,Peterson T.C., 1995. A new method for detecting undocumented discontinuities in climatological time series. Int. J. Climatol., 15, 369-377.
Fila G., Bellocchi G., Acutis M., Donatelli M., 2003. IRENE: a software to evaluate model performance. Eur. J. Agron., 18, 369-372.
Hipel K.W., McLeod A.I., 2005. Time Series Modelling of Water Resources and Environmental Systems. Electronic reprint of our book orginally published in 1994. http://www.stats.uwo.ca/faculty/aim/1994Book/

Lamb H.H., 1966. The changing climate, Methuen, London, 236 pp.

Lanzante J.R., 1996. Resistant, robust and non-parametric techniques for the analysis of climate data: theory and examples, including applications to historical radiosonde station data, International Journal fo Climatology, Vol. 16, 1197-1226.

Mariani L., 2002. Dispensa di Agrometeorologia, Clesav, Milano, 292 pp.

McLeod A.I, 2005. The Kendall Package, 10 pp., disponibile on line all'indirizzo cran.r-project.org/doc/packages/Kendall.pdf

Mennella, 1977. I climi d'Italia, F.lli Conte, Napoli, 3 vol.

Mitchell J.M., 1966. Climatic change, Tech. Note n. 79; World Meteorological Organisation, Geneva (op.cit.).

Munn R.E., 1970. Biometeorological methods, Academic Press, 336 pp.

Rana G., Rinaldi M., Introna M., 2004. Metodologie e algoritmi per il controllo di qualità di dati orari e giornalieri acquisiti da una rete agrometeorologica: applicazioni alle rete lucana SAL, Rivista Italiana di Agrometeorologia, Anno 9 - n.1- ottobre 2004, pp.14-23.

Rossiter D.G., 2006. An example of statistical data analysis using the R environment forstatistical computing, 143 pp. http://www.itc.nl/ personal/ rossiter.

Siegel S., Castellan N., 1988. Nonparametric statistics for the behavioural Sciences, McGraw-Hill, New York, 399 pp.

Sneyers R., 1998. Homogenizing time series of clima tological observations. The search and adjustment for inhomogeneities. Principle of methodology and example of results. Proceedings. Second Seminar for Homogenization of Surface Climatological Data. Hungarian Meteorological Service. Budapest, Hungary.

Van Vliet A.J.H., Overeem A., de Groot R.S., Jacobs A.F.G., Spieksma F.T.M., 2002. The influence of temperature and climate change on the timing of pollen release in the Netherlands, International Journal of Climatology, 22, pp. 1757–1767

Werner P. C., Gerstengarbe F.W., Fraedrich K, Oesterle K., 2000. Recent climate change in the North Atlantic/European sector, International Journal of Climatology, Vol. 20, Issue 5, 2000: 463-471.

Zeileis A., Kleiber C., Krämer W., Hornik K., 2003. Testing and Dating of Structural Changes in Practice, Computational Statistics and Data Analysis, forthcoming