UNIVERSITÀ DEGLI STUDI DI PADOVA
Facoltà di Scienze Statistiche

Corso di Laurea in Statistica e Tecnologie Informatiche

ANALISI DELLE MALATTIE INFETTIVE E LA LORO SORVEGLIANZA

Relatore:

Chiar.ma Prof.ssa Monica Chiogna

Tesi di Laurea di:

Sabrina PAVAN

Matricola n. 525230

ANNO ACCADEMICO 2009 / 2010

Indice

Introduzione

Per lungo tempo le malattie infettive sono state causa di malattia e di morte per l'uomo tant'è che la storia dell'uomo è caratterizzata dalla storia delle epidemie e della pandemia[1]: la peste, il vaiolo, il colera hanno decimato le popolazioni nell'era classica, nel Medioevo e nell'era moderna fino al ventesimo secolo, ma anche nel terzo millennio la questione si ripresenta con l'AIDS con il suo impatto demografico, sociale, sanitario ed economico per interi continenti: l'Africa sub-sahariana, il sub continente Indiano e il Sud Est asiatico.

Nel corso dei secoli sono stati raggiunti - dal punto di vista nutrizionale ed igienico - dei grossi risultati grazie soprattutto alle nuove scoperte scientifiche, alle migliorate condizioni di vita e alla crescente disponibilità di potenti antibiotici e vaccini efficaci che hanno cancellato o ridotto ai minimi termini malattie che da sempre avevano afflitto le popolazioni. Tuttavia, molte malattie infettive, come ad esempio la malaria, la sifilide, la lebbra, il colera, la peste che hanno determinato nel passato migrazioni di popolazioni e il declino di intere civiltà, ancora oggi non sono state eradicate ed arrestano lo sviluppo di interi continenti in un intreccio di condizioni di povertà e degrado igienico, sociale e sanitario.

Le malattie infettive non sono però esclusivo appannaggio dei paesi più poveri o in via di sviluppo proprio perché non sono sotto controllo e sono anzi in espansione. Anche nei paesi più industrializzati è cambiato l'atteggiamento di sicurezza e di facile dominio verso le malattie infettive e non solo per l'esplosione del fenomeno AIDS o la ricomparsa nei paesi più evoluti di alcune infezioni che si consideravano definitivamente debellate, ma per l'insorgere di situazioni ambientali del tutto nuove create dall'uomo alle quali c'è stato un adattamento o l'insorgere di nuovi organismi che caratterizzano la malattia.

Uno dei più importanti obiettivi dell'epidemiologia è l'individuazione delle cause di malattia e, ancora, prima la prevenzione o riduzione della frequenza di malattia in una popolazione. Si cerca cioè di estrapolare dai meccanismi di base di un'infezione, delle previsioni affidabili sullo

[1] Nel capitolo primo verrà descritta la distinzione tra pandemia ed epidemia.

sviluppo del contagio nel tempo per poter poi prevedere gli scenari generati dalle diverse contromisure e scegliere quindi la strategia di intervento ottimale. A questo scopo le organizzazioni sanitarie di tutto il mondo si servono ormai da decenni di modelli matematici e statistici, sia per le operazioni di routine (come le vaccinazioni), che per affrontare momenti di emergenza (come ad esempio la SARS o il nuovo ceppo A/H1N1 di influenza suina). Lo scopo di questo lavoro è quello di descrivere alcuni di questi modelli partendo da quelli deterministici sviluppati e poi raffinati a partire dal famoso lavoro del 1927 di Kermack e McKendrick[2], continuando con la descrizione di alcuni modelli probabilistici tra qui quello di Reed Frost, per concludere con l'applicazione di alcune tecniche statistiche per l'individuazione di focolai relativi alle malattie infettive. Prima però di procedere allo studio di questi modelli, è utile descrivere alcuni concetti di base utili per meglio comprendere quanto descritto nei successivi capitoli.

[2] Vedi capitolo 2, par. 2.2 e successivi.

Capitolo 1:
Le malattie infettive e il concetto di epidemia.

1.1 Il concetto di malattia epidemica.

Epidemia in greco vuol dire "sulla popolazione" e con tale termine si intende il verificarsi di una malattia che interessa un numero di individui (casi) nettamente superiore a quanto ci si sarebbe "atteso" - basandosi sulla recente esperienza[3] - in un determinato intervallo temporale, in una determinata popolazione e/o area geografica e per una causa comune sospetta o determinata. Quindi, perché si possa parlare di epidemia, si deve verificare un certo incremento, ossia è necessario che compaia un certo numero di casi. Questo numero dipende da numerose variabili, fra cui le più importanti sono: il tipo di agente, il tipo di popolazione, il periodo di tempo (es. stagione) considerato. Quando una epidemia è geograficamente molto estesa ed interessa molti individui della popolazione si parla di *pandemia*.

Tradizionalmente il termine epidemia viene usato quando l'infezione colpisce una popolazione: ciò si verifica spesso quando c'è un raggruppamento di persone (o anche animali, pesci, uccelli), in quanto questo fornisce le condizioni necessarie per permettere il moltiplicarsi e il diffondersi dei microrganismi che caratterizzano la malattia.

È però da sottolineare che, in base alla definizione esposta, non è indispensabile un numero rilevante di casi per dar luogo ad una epidemia, a volte anche un solo caso può essere considerato un evento epidemico. Ad esempio, il verificarsi di un solo caso di una rara e pericolosa malattia contagiosa mai avvenuta prima, o che è stata a lungo assente da una comunità, rappresenta un potenziale epidemia, così come anche un piccolo raggruppamento di casi di una malattia come il tifo in una comunità urbana in buone condizioni igienico sanitarie può essere considerata un'epidemia.

1.2 Identificazione di una epidemia.

Nel paragrafo precedente si è detto che si parla di epidemia quando è possibile evidenziare una frequenza inattesa di una certa malattia, a

[3] Il numero di nuovi casi nella popolazione durante un determinato periodo di tempo è chiamato il "tasso di incidenza".

condizione che siano noti i tassi attesi e sia presente una sorveglianza accurata. L'*incidenza attesa* si può calcolare in base ai dati esistenti, riguardanti la situazione endemica nel passato e può essere calcolata applicando alcune tecniche statistiche come ad esempio modelli di regressione ciclica[4]. Essa rappresenta la probabilità di contrarre una malattia e considera il numero dei nuovi casi di malattia in una popolazione per unità di tempo. Il numeratore è dunque costituito dal numero di nuovi eventi sorti nel periodo di tempo *t* (che può essere un anno, un mese o anche una settimana a seconda del tipo di malattia) e il denominatore dalle unità di tempo durante le quali i soggetti esaminati sono a rischio di contrarre la malattia.

Si calcola poi la *soglia epidemica* aggiungendo all'incidenza attesa (o numero atteso di casi) un valore pari ad (errore standard * 1.65[5]). Se il numero di casi rilevato supera la soglia epidemica così calcolata, allora si è in presenza di una epidemia[6]. Si parla anche di *herd immunity*, ossia la soglia del tasso di immunità collettiva al di sopra della quale la malattia non ha la possibilità di diffondersi in quanto è minima la probabilità che avvenga un contagio interumano: la popolazione immune agisce da barriera tra i soggetti contagiosi e quelli recettivi[7].

Il tasso di immunità collettiva necessario a raggiungere una *herd immunity* dipende dal grado di contagiosità dell'agente infettivo, dalla durata del periodo di contagiosità, dalla dimensione e dal comportamento sociale della comunità. Per raggiungere una *herd immunity* soddisfacente è necessario che la soglia richiesta sia uniformemente distribuita fra i gruppi della collettività (se ad esempio tra i bambini di una scuola si concentra una elevata quota di recettivi può verificarsi una epidemia, anche se nella collettività da cui i bambini provengono è stata raggiunta la soglia di eliminazione).

1.3 Quando è un'epidemia.

È molto facile individuare una epidemia quando una malattia infettiva è altamente contagiosa e penetra in una popolazione o in un territorio nel quale non è presente né la malattia infettiva né il relativo agente

[4] Ad esempio con modelli lineari generalizzati, inserendo nel predittore lineare una componente sinusoidale per cogliere l'aspetto stagionale, se i dati osservati presentano una caratteristica stagionale.
[5] 1.65 è il quantile di una variabile casuale normale standardizzata al 95% di probabilità.
[6] Sul concetto di soglia epidemica ed il relativo calcolo, si rinvia al capitolo successivo.
[7] Vedi par.2.7 cap. 2.

eziologico (cioè la malattia è stata "eradicata" da quel territorio), colpendo in breve tempo un numero molto elevato di soggetti.

Non sempre però le cose sono così evidenti, e talvolta può non essere facile stabilire se un certo incremento di casi di malattia costituisca o no una epidemia. Nella Figura 1.1 viene esemplificato l'andamento di una certa malattia in una popolazione da dicembre 2003 a dicembre 2005 .Per semplicità supponiamo che la popolazione sia "chiusa", ossia il numero di soggetti non abbia subito variazioni nel periodo considerato. Osserviamo il grafico in cui in ascissa abbiamo posto il tempo ed in ordinata il numero di nuovi casi.

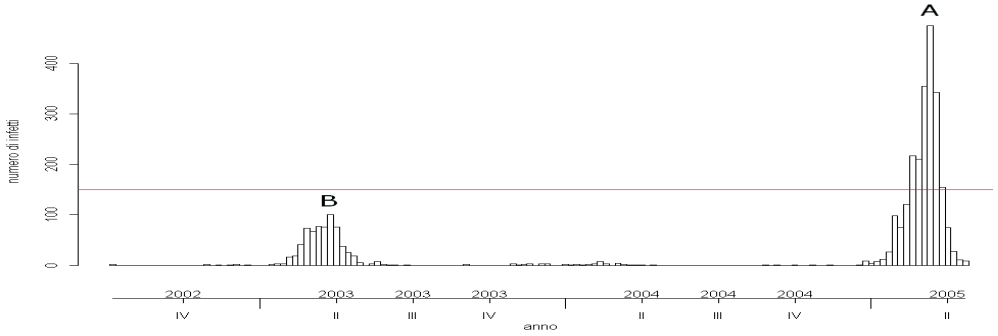

Fig. 1. 1: *andamento nel tempo di una malattia infettiva.*

Osservando il grafico della Figura 1.1 ci si può chiedere se il picco A rappresenta una epidemia. Il buon senso fa propendere per una risposta affermativa; ma allora che dire del picco B? Rappresenta anch'esso una epidemia? Fornire una risposta a problemi come questo può essere piuttosto complicato. Tuttavia, semplificando il problema, si può trovare almeno una linea-guida (derivante dalla definizione stessa di epidemia) che conduce ad una prima risposta ragionevole.

Si è già detto che una epidemia si ha in presenza di un numero imprevisto di casi, ossia superiore al *normale*; quindi, ci si può domandare se il picco A rappresenta una situazione *anormale.* Per rispondere ci si può basare sulla definizione di *normalità*, ossia supponendo che le frequenze dei nuovi casi rilevati nel tempo precedente il picco considerato abbiano una distribuzione approssimativamente normale allora, in questo caso si può adottare il criterio secondo cui *è anormale un valore al di fuori dell'intervallo media ± 2 volte la deviazione standard*. Tale intervallo è stato evidenziato nel grafico con la linea rossa. Il picco A si trova al di fuori del limite superiore della normalità, e ciò depone a favore dell'ipotesi che si tratti di una epidemia. Per avere un'idea riguardo al picco B, sarebbe necessario ripetere l'analisi statistica utilizzando però le frequenze di

malattia osservate nelle settimane antecedenti il secondo trimestre del 2003 . In questo caso però si giungerebbe a conclusioni poco affidabili, soprattutto a motivo del basso numero di osservazioni utilizzabili.
In realtà il tipo di analisi ora esposto ha molte limitazioni, non è applicabile in tutte le condizioni ed ha soltanto uno scopo orientativo. Nella pratica non ci si può basare soltanto sul calcolo di una semplice media e deviazione standard dei dati storici, ma è necessario ricorrere ad analisi statistiche più specifiche. Inoltre, bisogna anche tener conto di altri elementi, dipendenti sia dalla storia naturale della malattia che dalle caratteristiche popolazione in studio (es. distribuzione spaziale, eventuali relazione tra i diversi casi di malattia ecc.). Nel capitoli successivi, avremo modo di evidenziare e descrivere brevemente alcuni modelli per lo studio delle epidemie e definire i criteri per valutare se, da un punto di vista statistico, siamo in presenza o meno di epidemia.

1.4 Le curve epidemiche.

La rappresentazione attraverso un grafico del numero di nuovi casi di una malattia in funzione del tempo, con il numero di nuovi casi in ordinata ed il tempo in ascissa è una delle più comuni forme di visualizzazione dell'andamento di una malattia in una popolazione.
Il grafico che si ottiene dai dati raccolti durante una epidemia produce una *curva epidemica* (più correttamente rappresentata da un diagramma a barre). Una curva epidemica è caratterizzata nel tempo dalla crescita dell'incidenza da un livello trascurabile ad uno massimo, e poi dalla sua diminuzione fino ad un ritorno ai livelli pre-epidemici.
La forma della curva epidemica e la scala temporale dipendono da:

1. dal periodo di incubazione della malattia;
2. dalla forza dell'agente infettivo;
3. dalla proporzione degli individui *recettivi*[8] sull'intera popolazione;
4. dalla distanza tra gli individui (densità degli individui).

Così un agente fortemente infettivo con un breve periodo di incubazione che colpisce una popolazione con una grande proporzione di individui suscettibili e ad alta densità produce una curva con una pendenza ripida iniziale in un intervallo piccolo, dimostrando così una rapida diffusione dell'infezione tra la popolazione.

[8] La recettività è la potenzialità di un individuo ad ospitare un agente patogeno e a permetterne lo sviluppo o la moltiplicazione.

La curva epidemica, e la sua forma, fornisce infatti indicazioni preziose riguardo all'andamento di una epidemia, e può contribuire a rispondere ad importanti domande riguardanti il *tipo di esposizione*, la via di diffusione della malattia, quando si è verificata l'esposizione all'agente della malattia, o quale è stato il *periodo di incubazione*.

Per quanto riguarda ad esempio il tipo di esposizione, un'epidemia viene detta a *sorgente comune* quando tutti i casi (intendendo per *caso* il singolo soggetto ammalato) hanno origine da una stessa causa (es. un soggetto infetto o un alimento contenente un agente patogeno). Se il periodo di esposizione è breve, allora l'epidemia con sorgente comune è detta epidemia *puntiforme* o puntuale (in inglese: *point source*), e tutti i casi si verificano, all'incirca, entro un lasso di tempo corrispondente al periodo di incubazione. Tipiche epidemie a sorgente comune e con breve periodo di incubazione sono le intossicazioni alimentari che derivano dall'assunzione, da parte di una collettività, di alimenti contaminati da patogeni. Una *epidemia a propagazione* è, invece, quella causata da un agente che viene escreto inizialmente da uno o più casi primari[9], e quindi si propaga nel tempo ad individui recettivi che costituiscono casi secondari. Uno dei casi primari è spesso il *caso-indice*, cioè il primo che è stato notato dagli investigatori.

In Figura 1.2 vengono riportate esempi di curve epidemiche rispettivamente relative ad epidemie di tipo puntiforme ed epidemie a propagazione:

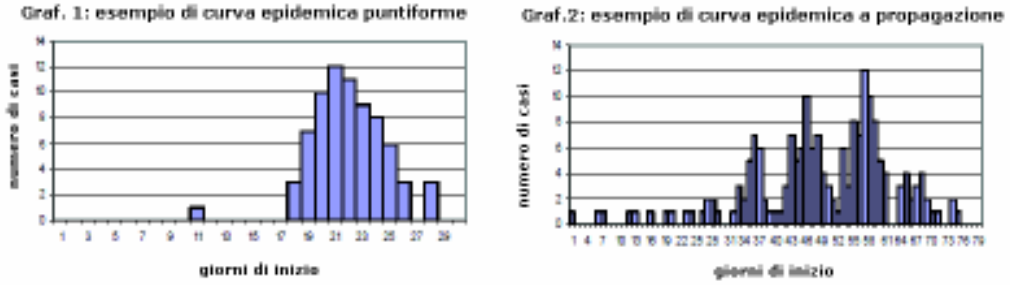

Fig. 1.2: *curve epidemiche puntiformi e a propagazione*.

L'intervallo di tempo fra picchi di successivi o *grappoli* (*cluster*)[10] temporali di casi, che separa i casi primari da quelli secondari, riflette il periodo di incubazione della malattia.

[9] I casi primari sono soggetti infettati inizialmente attraverso una sorgente comune, mentre i casi secondari sono soggetti infettati dai casi secondari. Di solito sono conviventi, familiari o persone nello stesso luogo di lavoro.

[10] Per *cluster* si intende un gruppo geograficamente circoscritto di eventi di dimensione e concentrazione tali da rendere improbabile che siano di natura casuale.

Di solito tutti i casi di epidemie point-source avvengono all'interno dell'unico periodo di incubazione dell'infezione. Però se la lunghezza del periodo tra i picchi successivi, nel caso di epidemia a propagazione, è inferiore al più comune periodo di incubazione allora è difficile differenziare tra epidemia a propagazione e una serie di epidemie point-source.

Ci sono poi curve epidemiche che presentano picchi irregolari che si alternano in periodi successivi: queste sono caratterizzate da malattie provocate da una comune fonte infettiva e gli individui sono sottoposti a periodi intermittenti ad uno stesso fattore di rischio. Infine le curve epidemiche caratterizzate da un graduale aumento del numero dei casi, seguito da una sua successiva graduale sua diminuzione, senza perciò presentare picchi evidenti, si riferiscono a malattie causate da una esposizione continuativa ad uno stesso fattore di rischio. In Figura 1.3 vengono riportate le forme tipiche delle due curve appena descritte:

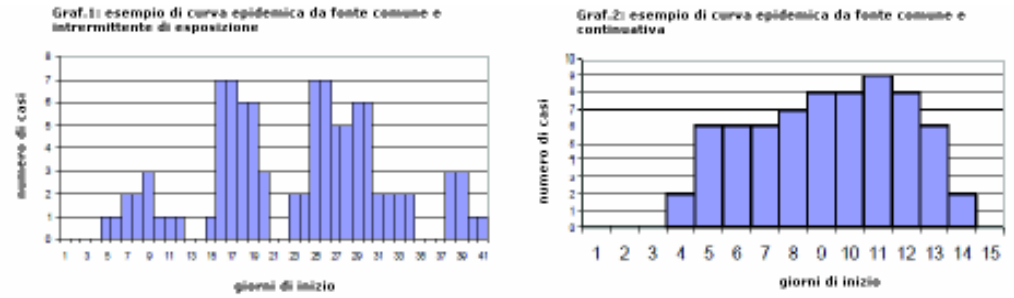

Fig. 1.3: *curve epidemiche da fonte comune intermittente e continuativa*.

1.5 Andamento delle epidemie.

Nel grafico precedente si è introdotto il concetto di curva epidemica e di come essa fornisca delle informazioni utili sulla malattia osservata. Illustriamo adesso un esempio di andamento della diffusione di una malattia in una popolazione: nella Figura 1.4 viene evidenziato l'andamento tipico di un'epidemia:

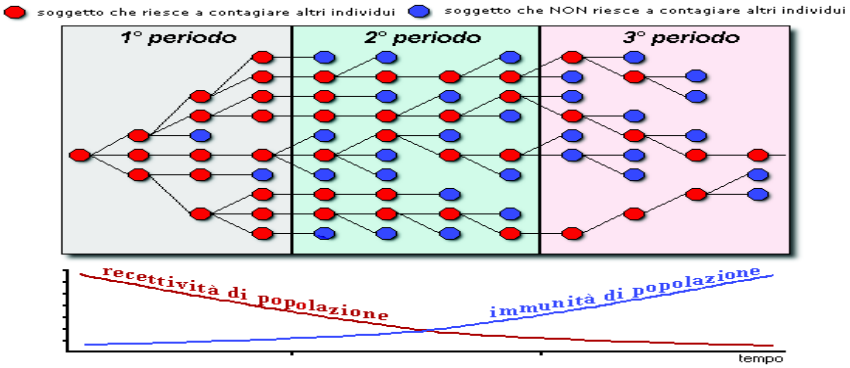

Fig. 1.4: *schema dell'andamento di una epidemia in una popolazione recettiva*.

Nello schema della Figura 1.4 è illustrato l'andamento di una tipica epidemia che si verifica quando un agente infetta una popolazione costituita da soggetti pienamente recettivi.

Ogni cerchio rappresenta un soggetto infetto; le linee nere indicano l'avvenuto trasferimento dell'infezione da un individuo all'altro. I cerchi rossi rappresentano le persone che riescono a trasmettere il contagio ad altri soggetti. I cerchi blu simboleggiano quelli che invece non sono riusciti ad infettarne altri.

Durante il primo periodo, la maggior parte della popolazione è suscettibile (curva rossa), e quindi la malattia ha modo di diffondersi facilmente negli individui della popolazione. Contemporaneamente, si assiste ad un lieve incremento dell'immunità di popolazione, dovuta ai soggetti che si sono infettati e successivamente si sono immunizzati; l'andamento dell'immunità di popolazione nel tempo è rappresentata dalla curva blu sull'asse cartesiano.

Durante il secondo periodo, il numero dei suscettibili diminuisce: ciò è la conseguenza del fatto che quelli ammalatisi durante il primo periodo sono morti oppure sono passati nella categoria degli "immuni". Pertanto, aumenta il numero di infetti che, non avendo sufficienti contatti con i soggetti recettivi, non riescono a trasmettere il contagio (cerchi blu); tuttavia, la malattia si manifesta ancora con discreta frequenza, in quanto i recettivi sono ancora relativamente numerosi. L'immunità di popolazione continua a crescere.

Nel terzo periodo l'immunità di popolazione raggiunge il massimo livello. Il numero di contagianti si fa via via più basso, e quindi l'epidemia si esaurisce spontaneamente.

Il modello appena descritto è molto semplice, in quanto presuppone una popolazione inizialmente del tutto recettiva e tiene conto - in sostanza - soltanto della variabile "immunità di popolazione". Tuttavia, è necessario ricordare che i fattori associati alla diffusione delle infezioni, come le caratteristiche dell'ospite o dell'agente infettante, sono numerosi ed interagiscono fra loro e con altre variabili ambientali, formando uno schema più complesso[11].

Attraverso invece l'analisi grafica delle serie storiche relative alle malattie infettive, ossia osservando l'andamento nel tempo delle malattie infettive, si possono notare alcuni andamenti tipici delle curve, in particolare i cambiamenti temporali e le fluttuazioni che a volte

[11] Nel capitolo successivo, verrà attuata una descrizione ben dettagliata del modello, noto anche come modello S.I.R.

caratterizzano tali curve. In particolare, l'osservazione nel tempo di una malattia infettiva, consente principalmente di distinguere dall'epidemia due tipi di malattie:

- Malattia endemica;
- Malattia sporadica.

La *malattia endemica* è una forma morbosa che è costantemente presente in una popolazione o in una determinata area geografica. Se la prevalenza[12] della malattia è bassa, allora si tratta di malattia *ipoendemica*; se, invece, la prevalenza è alta la malattia è *iperendemica*. Un esempio di una malattia endemica è la malaria in alcune parti dell'Africa.

Infine, si dice *sporadica* una malattia che si presenta irregolarmente ed imprevedibilmente nello spazio e nel tempo, generalmente con bassa frequenza. Nella Figura 1.5 vengono illustrati esempi di andamento nel tempo di una malattia epidemica, endemica e sporadica.

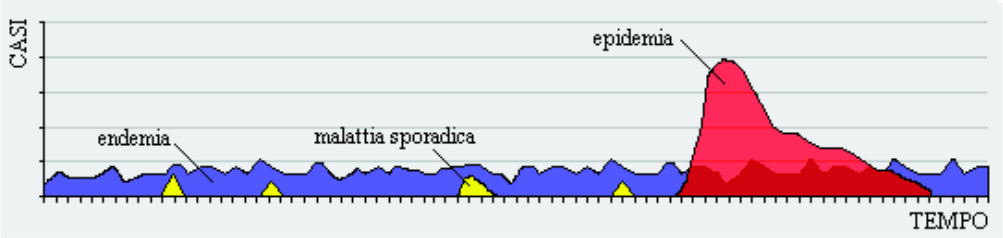

Fig. 1.5: *esempio di andamento nel tempo di una malattia endemica, sporadica, epidemica*

In alcuni casi, la frequenza della malattia ha un andamento temporale particolare, con fluttuazioni abbastanza prevedibili. In questi casi però, non si tratta di epidemie, in quanto nel concetto di epidemia è insita la imprevedibilità dell'evento. Le malattie che vengono trasmesse da punture di vettori (insetti e artropodi) hanno una forte predilezione per i mesi caldi (estate ed inizio autunno) in corrispondenza con la maggiore attività dei vettori. Anche malattie non trasmesse da insetti possono manifestare un andamento stagionale, come la leptospirosi nell'uomo in alcune aree geografiche: la frequenza della malattia subisce un aumento in corrispondenza delle stagioni calde (estate-autunno). Ciò dipende dalle caratteristiche dell'agente e da motivi legati al comportamento dell'uomo, che durante la buona stagione trascorre più tempo all'aperto

[12] La prevalenza misura la proporzione di individui di una popolazione che, in un dato momento, presentano la malattia. Essa è data dal rapporto tra i soggetti colpiti da malattia (ammalati) e gli stessi soggetti più tutta la popolazione a rischio, cioè quelli non ancora ammalati ma "suscettibili" di ammalarsi, ossia che possono contrarre la malattia in studio. La prevalenza rappresenta la probabilità di avere una malattia.

e ha maggiori probabilità di infettarsi per contatto con animali portatori. In questi casi si parla di trend-stagionali in quanto le fluttuazioni periodiche della malattia sono correlate ad alcune stagioni in particolare; un altro esempio tipico di malattia infettiva che presenta caratteristiche di stagionalità è l'influenza che colpisce soprattutto nei mesi invernali[13].
In altri casi, le fluttuazioni avvengono in tempi più lunghi. In tal caso si parla di *andamento secolare*. La registrazione della frequenza di casi di malattia in una popolazione per un lungo periodo (anni) è utile, oltre che per conoscere meglio la storia naturale della malattia, anche per prevederne la probabile incidenza futura e per pianificare i più appropriati programmi di controllo o prevenzione.

1.6 Analisi delle epidemie attraverso l'impiego dei modelli.

Le prime analisi condotte per lo studio delle epidemie sono state attuate principalmente attraverso l'impiego di modelli con i quali si cercava di rappresentare, in maniera semplificata, il fenomeno studiato attraverso un numero limitato di parametri da cui far dipendere tutto il comportamento. Principalmente i modelli destinati a studiare il comportamento delle malattie infettive, furono di tipo deterministico e solo successivamente si cominciarono ad applicare modelli di tipo probabilistico. La differenza tra i due tipi modelli sta nel fatto che nei modelli *deterministici*, si esaminano grandezze i cui valori sono noti in corrispondenza a tutti gli istanti successivi all'origini del processo. La loro espressione si ricava sempre da un'equazione differenziale che contiene delle costanti, le quali costituiscono poi i parametri del modello.
I modelli di tipo *probabilistico* (stocastico) che sono concettualmente più complicati dei precedenti, introducono, invece, la probabilità che in un certo istante il fenomeno abbia un dato valore. In altre parole, quest'ultimi modelli non mirano a permettere di calcolare con esattezza l'evoluzione di un processo, ma danno una indicazione del grado di probabilità che, ad un certo istante, il processo assuma un certo valore. Entrambi i tipi di modelli descritti sono di tipo temporale in quanto studiano l'evoluzione del sistema nel tempo.
Uno dei primi modelli deterministici di tipo più semplice è quello che si ricollega ai lavori di Hammer (1906), il quale divise la popolazione colpita

[13] Nel capitolo 3 verrà attuata un'analisi della serie storica relativa all'influenza B.

da epidemia in due gruppi a seconda che essi siano infetti o non infetti[14]. Successivamente vennero proposti studi matematici più elaborati dello stesso tipo attraverso l'introduzione di nuovi parametri: in particolare, i più noti sono gli studi di Kermack e McKendrick (1927) i quali introdussero un aspetto di estremo interesse nella teoria delle epidemie, e cioè l'esistenza di una *soglia epidemiologica*.[15] Questi però urtarono contro difficoltà analitiche molto rilevanti per la soluzione delle equazioni differenziali che governavano il modello da essi proposto e dovettero perciò ripiegare su una soluzione approssimata ottenuta mediante uno sviluppo in serie. Più recentemente il lavoro di Kermack e McKendrick è stato riesaminato da David Kendall (1956), il quale oltre ad aver trovato una soluzione esatta del modello, studiò sia la posizione del centro dell'epidemia, ossia l'istante in cui si verifica il massimo del processo infettivo, sia l'intensità dell'epidemia, espressa dal rapporto tra il numero totale degli individui che alla fine dell'epidemia hanno contatto l'infezione e il numero totale delle persone suscettibili. Da questa modellistica, di tipo generale, si sono staccati alcuni rami con lo scopo specifico di studiare il comportamento di alcune particolare malattie. L'esempio tipico di queste indagini specialistiche è offerto dai modelli che Sir Ronald Ross (1911) sviluppò per lo studio della malaria.

In questi ultimi anni, ai modelli deterministici citati, sono state aggiunte alcune generalizzazioni che non presentano particolare interesse. Una sola eccezione è costituita dal tentativo di studiare la distribuzione territoriale sia degli infetti che dei suscettibili, dovuta a Kendall (1956) e successivamente sviluppata da Bailey (1957).

Notevolmente più complicati sono i concetti che stanno alla base dei modelli probabilistici relativi all'epidemiologia. Questi si possono dividere in due grandi classi corrispondenti al caso discreto e a quello continuo. La prima classe, quella dei processi discontinui, comprende i modelli del tipo a catena binomiale introdotta negli Stati Uniti da Reed e Forst, nel 1928[16] (Abbey, 1952, Frost, 1976). Il modello di questi due autori, rimase per molto tempo ignorato alla maggior parte degli studiosi, in quanto esso non venne pubblicato. In Inghilterra invece, questi modelli vennero introdotti, nella loro forma più ortodossa, ad opera di Greenwood, nel 1931, il quale li impiegò prevalentemente per studiare lo sviluppo delle epidemie nell'ambito familiare. Nel capitolo successivo

[14] Nel capitolo successivo verranno descritti brevemente alcuni semplici modelli deterministici.
[15] Per una descrizione più dettagliata si rinvia al secondo capitolo .
[16] Vedi paragrafo 2.15 del capitolo 2.

verrà fatta una breve descrizione del modello probabilistico di Reed e Forst.

L'impostazione probabilistica è diversa ovviamente da quella deterministica e mentre quest'ultima considera come variabile fondamentale il numero dei suscettibili all'istante temporale t, che possiamo indicare con S(t), la teoria stocastica utilizza ancora S(t), ma la considera come una variabile casuale, la cui probabilità è $P_x(t)$. Ciò significa che nei modelli probabilistici si suppone che il verificarsi di un numero n_1 oppure n_2 di infetti in una certa collettività sia effetto del caso, e corrispondano, rispettivamente, le probabilità P_{n1} e P_{n2}.

Questa impostazione probabilistica dei modelli presenta, rispetto a quella deterministica, dei vantaggi e degli inconvenienti: in genere, però, gli autori che si sono interessati di queste questioni, propendono nel riconoscere una superiorità al modello stocastico. In modo speciale nella teoria dell'epidemiologia, dove il verificarsi di un caso infettivo nella Famiglia A, piuttosto che nella B, sembra assolutamente casuale, gli studiosi dei modelli sono d'accordo che la via da seguire è quella di tipo probabilistico. Purtroppo, come già detto, essa presenta delle difficoltà maggiori dell'altra e questo è uno dei motivi per cui questi modelli sono comparsi solo in epoca molto recente.

1.7 La sorveglianza delle malattie infettive.

Dopo aver introdotto nei paragrafi precedenti alcuni concetti di base relativi alle epidemie, concludiamo questo capitolo descrivendo brevemente il concetto di sorveglianza delle malattie infettive, rinviando per maggiori dettagli al capitolo 4 del presente lavoro.

La sorveglianza è un'attività routinaria e il suo scopo è quello di individuare variazioni nella distribuzione spazio-temporale della morbosità e della mortalità (focolai epidemici e *clusters*) e/o dei determinanti dello stato di salute che richiedono interventi di controllo a breve termine.

Il "Sistema di sorveglianza" è il sistema che realizza le attività di sorveglianza e che integra le tappe del flusso continuativo di informazioni necessarie alla Sanità pubblica per decidere sulle azioni di controllo dei problemi di salute della popolazione.

La presenza di fattori di rischio per la salute che possono generare patologie negli individui di una certa popolazione crea comprensibilmente uno stato di allarme e la necessità di ricevere

risposte adeguate coerenti sul tema dello stato di salute, della possibilità di azioni preventive, di controllo, come pure di suggerimenti di tipo comportamentale.

Per identificare situazioni di *allarme* e testare rapidamente ipotesi esplicative sullo stesso, occorre perciò dotare il sistema di sorveglianza di metodi di indagine-analisi basati su eventi sentinella, analisi periodiche, monitoraggio statistico.

Sul piano ambientale, gli eventi sentinella sono tutti quegli eventi anomali che non possono essere ritenuti casuali, come molte tipologie di incidenti; sul piano degli eventi di salute sono invece disponibili liste di malattie che possono costituire metodo e base informativa di riferimento per identificare associazioni con cause ambientali.

La valutazione dell'andamento nel tempo e nello spazio viene poi effettuata sia mediante analisi al termine di periodi prefissati (periodiche) sia in continuo, cioè via via che gli eventi si verificano. Per le *analisi periodiche*, le valutazioni si basano su test a posteriori, solitamente di tipo concettualmente semplice, di facile esecuzione e comprensione, quali ad esempio il test Chi-quadro per il trend o il rapporto tra casi osservati e casi attesi, in forma standardizzata o meno per altri parametri.

Per le *analisi in continuo* esistono metodi che tengono conto, in vario modo, degli eventi che si verificano. Questi metodi di analisi sequenziale, definiti anche "autorinforzanti" in quanto rafforzano la loro prestazione cumulando gli eventi precedenti, hanno bisogno di dati disaggregati nel tempo, e di tecniche più sofisticate, che a fronte di maggiore difficoltà concettuale ed esecutiva hanno il pregio di offrire migliori prestazioni.

I metodi più conosciuti e utilizzati sono le carte cumulative di controllo o CUSUM, mutuate dalle tecniche per il controllo della qualità dei processi produttivi, oppure tecniche di regressione e diversi altri più o meno sofisticati. La scelta del metodo è una fase piuttosto delicata perché in talune situazioni può comportare importanti conseguenze in termini di risultati raggiungibili.

Una volta che il sistema ha segnalato una uscita dalle condizioni prestabilite come di normale funzionamento, includenti anche livelli di fluttuazione accettabili, è opportuno attuare una serie di accertamenti per confermare che l'aumento del numero dei casi è reale. In particolare la sequenza delle domande a cui dare una risposta è la seguente:

- la diagnosi è corretta?
- I casi sono veri casi o sono falsi positivi?
- Vi sono altri casi potenziali?
- I casi noti sono tutti i casi insorti?
- L'aumento osservato dei casi è un artefatto? È cambiata la diagnosi, la sorveglianza, la popolazione a rischio, ma non la frequenza abituale dei casi insorti?
- I casi sono associati fra loro?
- I casi insorgono "per caso" allo stesso tempo?

Questa è una fase cruciale e delicata perché le sue conseguenze si ripercuotono sia sulla mobilitazione di risorse di sanità pubblica sia, a cerchi via via più ampi, sulle altre istituzioni pubbliche e sulla comunità. Successivamente verranno poi attuate indagini complete per l'identificazione dell'agente causale o della sorgente nonché della modalità di trasmissione della malattia, e attuate le principali misure di comunicazione e controllo della diffusione dell'infezione.

È molto importante dunque che il sistema di sorveglianza pubblica in materia di salute si avvalga di buoni strumenti adeguati a rispondere alle numerose domande che nascono sia da una preoccupazione lecita dei cittadini nei confronti dell'ambiente in cui vivono e dell'impatto sulla propria salute, sia dalla richiesta da parte dei decisori politici di conoscenze affidabili, con le quali operare e valutare scelte politiche e programmatiche basate su prove. Nella sorveglianza pubblica della salute, specie in aree circoscritte, su eventi rari, su esposizioni ridotte o su situazioni miste, esiste infatti un'elevata probabilità statistica che molti allarmi si verifichino per effetto del caso, pertanto nessuna segnalazione deve essere mai considerata come conclusiva, essendo sempre necessaria l'indagine mirata alla conferma statistico-probabilistica e alla ricerca della causa. Nessuna segnalazione deve essere trascurata perché, fino a prova contraria, è da considerarsi anomala.

Nel capitolo finale del presente lavoro, verrà descritto e applicato un metodo statistico per il controllo routinario delle malattie infettive proposto da Farrington, il quale sulla base di dei dati storici della serie osservata, fornisce un criterio per segnalare un allarme in caso vi sia presenza di epidemia.

Capitolo 2:
l'impiego dei modelli nello studio delle epidemie.

Nel capitolo primo del presente lavoro, si sono introdotte alcune nozioni fondamentali sulle malattie infettive ed epidemiche: in particolare si sono descritti alcuni criteri generici per cercare di individuare un epidemia introducendo il concetto di curva epidemica come strumento in grado di fornire utili indicazioni sulla sua identificazione. Si è poi cercato di descrivere brevemente il meccanismo di diffusione di un'epidemia sulla base di un semplice modello che tiene conto soltanto della variabile immunità di popolazione. Si è visto inoltre, che i modelli per lo studio delle malattie infettive e la loro diffusione nel tempo, hanno sempre rappresentato un buon strumento di analisi, che nel corso degli anni è stato sottoposto a molteplici studi portando all'elaborazione di strumenti anche molto complessi in grado di spiegare l'evoluzione delle epidemie fornendo così utili indicazioni ai fini di un loro controllo anche da parte delle pubbliche autorità. Si possono distinguere principalmente due tipi di modelli: quelli deterministici e quelli probabilistici. Si è visto che i primi sono i più semplici, e in essi, le variabili di input assumono valori fissi, mentre quelli probabilistici hanno una struttura più complessa di quelli deterministici poiché tengono in considerazione le variazioni delle variabili di input, dando risultati in termini di probabilità. In questo capitolo svilupperemo un modello semplice generale per il diffondersi di malattie infettive in comunità chiuse, considerando dapprima il modello base deterministico S.I.R. proposto da Kermack e McKendrick, introducendo successivamente per lo stesso modello la componente stocastica. Concluderemo il capitolo con una breve presentazione del modello di Reed e Frost.

2.1 Breve storia sull'utilizzo dei modelli per lo studio delle epidemie.

L'applicazione di modelli matematici allo studio della diffusione di malattie ha una lunga storia il cui inizio è stato segnato dal lavoro di Daniel Bernoulli sugli effetti del vaccino antivaiolo e la diffusione dello

stesso nel 1760[17]. Successivamente, ci furono ricerche molto interessanti condotte da William Farr nel 1840 e nel 1866 sulle curve rappresentanti l'andamento di certe malattie infettive, e da Evans nel 1875 relativi sull'epidemia inglese di vaiolo del 1871-72. Si tratta tuttavia di indagini di tipo descrittivo che tendono a desumere la forma della curva epidemica dai dati osservati per una certa epidemia, che non entrano nel vivo del meccanismo di trasmissione delle epidemie. Questo aspetto compare nel 1906, principalmente ad opera di Hammer e in parte anche da Brownlee, i quali introdussero i concetti di individuo suscettibile e di tasso di contatto tra infetti e suscettibili tramite cui è possibile giungere alla nozione di curva epidemica che esprime, in forma sintetica, l'andamento del processo in esame. Tuttavia, l'impiego di strumenti matematici più complessi attraverso l'implementazione di modelli di diffusione virale tali da studiare e prevedere gli effetti di un possibile scoppio epidemico avviene principalmente nel 1927 con gli studi di Kermack e McKendrick, che proposero un modello deterministico a tempo discreto strutturato per età, per lo studio dell'evoluzione nel tempo delle epidemie. Il modello da essi proposto è noto come modello S.I.R. (*Susceptible*, *Infected* e *Recovered*) che portò ad un risultato importante: la densità degli individui suscettibili deve eccedere un certo valore critico affinché possa verificarsi un'epidemia. Successivamente ci fu un nuovo impeto sugli studi delle epidemie negli anni 40; in particolare Bartlett (1949) e Bailey (1950) introdussero l'idea di processo stocastico ai modelli deterministici di Kermack e McKendrick, e svilupparono un sistema di equazioni differenziali di funzioni di probabilità a due variabili: il numero dei suscettibili e il numero degli infetti.

Negli ultimi decenni la modellizzazione matematica delle epidemie è diventata molto importante per riuscire a comprendere le dinamiche di molte malattie infettive, volte a confrontare le strategie di controllo e di prevenzione nonchè per studiare i possibili effetti dei vari fattori biologici e sociologici nell'espandersi della malattia all'interno di una popolazione. Le variabili fondamentali sono la modalità di trasmissione, i diversi tempi di incubazione, le caratteristiche del decorso, le dinamiche di diffusione tra diversi sottogruppi e i possibili effetti delle vaccinazioni.

[17] Egli sviluppò un modello matematico usando i dati delle tabelle di mortalità e, risolvendo un'equazione differenziale, concluse che una vaccinazione di massa avrebbe aumentato la speranza di vita di circa tre anni.

2.2 I modelli S.I.R.

Il modello base deterministico per descrivere nel tempo continuo lo sviluppo di una malattia infettiva o di una epidemia, è quello proposto dagli studi di Kermack e McKendrick contenuti in una sequenza di tre articoli pubblicati rispettivamente nel 1927, 1932 e 1933. Nel primo di questi articoli gli autori descrivono il modello base generico.

Nel modello S.I.R. la popolazione viene suddivisa in diverse classi, in funzione delle caratteristiche principali che sono rilevanti per l'infezione in esame. In particolare, il modello base della teoria epidemiologica prevede che la popolazione si suddivida in tre sottoclassi in base alle caratteristiche degli individui in relazione alla malattia.

Nei modelli S.I.R., non considerando la struttura d'età, la popolazione è suddivisa nei seguenti gruppi:

- i **Suscettibili**: cioè persone sane e sensibili al contagio;
- gli **Infetti**: cioè le persone che hanno contratto la malattia e quindi contagiose;
- i **Recuperati** o guariti; cioè le persone che dopo aver contratto la malattia in passato sono da essa guarite. In particolare, essi non sono più né infettivi né suscettibili o perché guariti, o perché isolati, o anche – nel caso di malattie gravi – perché morti.

Con riferimento ai Recuperati, si può assumere che una volta passato per la fase infettiva l'individuo sia rimosso in permanenza dalla dinamica dell'infezione e, che una volta guarito, mantenga una immunità perpetua all'infezione in questione.

Il numero di persone appartenenti a ognuno dei tre gruppi è indicato rispettivamente con le variabili S,I, e R (da cui il nome del modello).

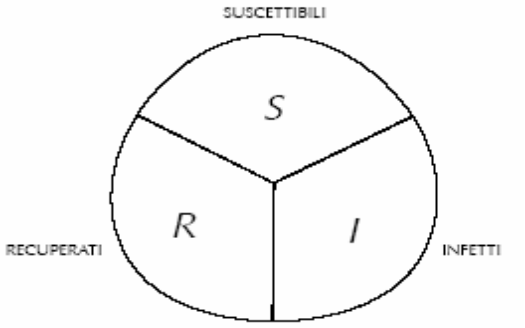

Fig. 2.1: *Rappresentazione insiemistica della popolazione in un modello S.I.R.*

Lo schema dell'andamento secondo il modello è:

$$S \rightarrow I \rightarrow R$$

Naturalmente possono esistere situazioni più complesse, ad esempio che l'immunità conferita sia solo temporanea per cui si ha $S \rightarrow I \rightarrow R \rightarrow S$: questo è il caso delle epidemie stagionali di influenza per le quali il virus muta di anno in anno, e l'acquisizione dell'immunità non dura tutto l'anno.[18]

Il modello in questione studia le tre variabili S, I, e R in funzione del tempo, in quanto esse assumono valori diversi in funzione del tempo. Lo scopo del modello S.I.R. è valutare l'andamento temporale dei valori delle tre variabili a partire dal tempo presente e proiettarsi nel futuro. Se il modello S.I.R. rappresenta bene la realtà, allora i valori da esso forniti saranno vicini a quelli che si avranno dall'evoluzione della malattia. Il modello si basa principalmente sulle seguente ipotesi:

1. non c'è tempo di incubazione; ossia ogni infetto è immediatamente infettivo;
2. il contagio avviene per contatto diretto;
3. la probabilità di incontro fra due possibili individui della popolazione è uguale;
4. ogni individuo malato ha una eguale probabilità di guarigione per unità di tempo costante.

A queste ipotesi, aggiungeremo l'ipotesi che la popolazione studiata sia molto grande, cosicché si possano ignorare le fluttuazioni casuali e usare un modello deterministico.

Si indica con N il numero totale di individui nella popolazione che si suppone costante (gli eventuali morti a seguito dell'infezione vengono conteggiati nella categoria R) e con $S(t)$, $I(t)$ ed $R(t)$ il numero di individui di ciascuna delle tre classi al tempo t. Pertanto si avrà:

$$S(t) + I(t) + R(t) = N \qquad (2.1)$$

Questo perché il periodo di durata della malattia è breve, pertanto la popolazione si considera costante. Il modello SIR evidenzia come gli individui suscettibili diventino parti della popolazione infetta, cercando di quantificare la velocità con la quale ciò avviene. Ciò dipende dalla numerosità dei due gruppi, quindi dai valori I e S e dal tipo di malattia.

[18] Esistono inoltre altre classi di modelli che ipotizzano situazioni più o meno complesse: come i modelli SI in cui dopo la fase infettiva non resta immunità alla malattia, oppure i modelli SEIR in cui lo stato di infetto non coincide con quello di infettivo, per cui si considera la classe degli *esposti* (E), coloro cioè che pur incubando la malattia non sono ancora in grado di trasmetterla.

2.3 Il numero medio di contatti giornalieri tra Suscettibili e Infetti.

Se I è il numero d'infetti, allora una sola persona suscettibile può al massimo entrare in contatto con I persone infette. Essendo le persone suscettibili in numero S, segue che il numero massimo di contatti possibili tra i due gruppi è S *volte* I, quindi SI. E` ipotizzabile però che nella realtà una persona suscettibile durante l'arco d'una giornata non entrerà in contatto con tutti gli infetti ma solo con un loro sottogruppo. Sebbene questo numero dipenda da molti fattori, come ad esempio il lavoro svolto da una persona, e quindi potrebbe essere diverso da un individuo suscettibile ad un altro, nel modello matematico che si sta sviluppando si decide di usare un valor medio[19], identico per tutta le persone suscettibili: si può ipotizzare che in media ogni persona suscettibile entri in contatto ad un certo *tasso di contatto*. Questo tasso di contatto dipende dal tipo di malattia e dal tempo t. Ci può essere una dipendenza esplicita (ad esempio stagionalità), oppure una implicita tramite altre variabili (ad esempio, potrebbe dipendere dalla densità della popolazione che potrebbe variare nel tempo). Per rendere più generale la discussione, si può usare un parametro c da usare nel modello: esso rappresenta la frazione delle persone infette che, in media, un suscettibile contatta nel periodo di una giornata. Prendendo per il momento il giorno come unità di tempo per lo studio delle variazioni delle tre grandezze considerate nel modello, si avrà che il numero di *contatti medi* tra la popolazione suscettibile e quella infetta durante l'arco di un giorno sarà:

$$S \cdot (I \cdot c) \tag{2.2}$$

Il parametro c modella il grado d'interazione sociale ed è legato al numero medio di contatti che la popolazione manifesta. Un contatto con una persona infetta non necessariamente implica il contagio, quindi non tutti i contatti, che si è detto essere in numero $S \cdot I \cdot c$, conducono ad un'infezione delle persone suscettibili coinvolte. Del numero di contatti medi se ne deve prendere solo una parte: ad esempio, se da riscontri pratici si stimasse che solo 4 su 100 contatti hanno come esito un contagio effettivo, allora analogamente a quanto detto precedentemente per i contatti effettivi, si potrebbe moltiplicare la quantità $S \cdot I \cdot c$ per

[19] Si ricorda l'ipotesi 3 del modello, che stabilisce ci sia la stessa probabilità di incontro per ciascun individuo.

4/100 ed ottenere il *numero medio atteso di nuove infezioni* giornaliere. Usando il parametro ρ per esprimere tale valore percentuale, è possibile scrivere:

$$S \cdot [(I \cdot c) \cdot \rho]. \tag{2.3}$$

Questo è il numero medio di persone che giornalmente si ammalano e ρ rappresenta la probabilità che un contatto risulti in una infezione se uno dei due individui è suscettibile e l'altro è infetto. Se, infine, si indica il prodotto $c \cdot \rho$ con un solo parametro a allora è possibile riscrivere la precedente espressione come:

$$aSI, \qquad \text{con} \qquad a = c \cdot \rho, \tag{2.4}$$

dove, fissato un istante del tempo, S è il numero di suscettibili in tale istante, I è il numero di infetti in tale istante e a è un parametro legato all'interazione tra individui e alla malattia e che si suppone costante.

2.4 L'equazione esprimente le variazioni del valore *S* nel tempo.

Ottenuta l'espressione per il calcolo del numero medio di persone che giornalmente si ammalano, sarà possibile scrivere un'equazione matematica esprimente la variazione giornaliera del valore di S. La quantità sopra calcolata può essere infatti utilizzata per determinare il valore S al trascorrere del tempo. Si è detto che S è il numero di persone suscettibili nell'istante di tempo d'interesse, quale l'*oggi*. Per descrivere la variazione di S nel tempo sarà necessario fare riferimento alla quantità S in diversi istanti di tempo, perciò per poter distinguere queste diverse istanze di S si userà a pedice un'indicazione temporale quale S_{oggi}.

Osservando nell'arco di 24 ore il gruppo suscettibili si vedrà che le persone ad esso appartenenti *domani* potranno o continuare ad esserlo o essersi ammalate e appartenere al gruppo infetti. Osservando pertanto il gruppo suscettibili, si vedrà quindi che un flusso giornaliero di persone si *sottrarrà* ad esso per andare ad *aggiungersi* al gruppo infetti. Inoltre, per il gruppo suscettibili non esiste un flusso entrante, in quanto si è ipotizzato che la popolazione totale fosse chiusa e che la malattia conferisse immunità. E' possibile scrivere che, secondo il modello proposto dove tutti i parametri sono considerati essere definiti valori medi su una giornata, domani si attende un valore S_{domani} tale che:

$$S_{domani} = S_{oggi} - a\, S_{oggi}\, I \qquad (2.5)$$
$$S_{domani} - S_{oggi} = - a\, S_{oggi}\, I$$

dove $a = c \cdot \rho$ *e* $S_{domani} - S_{oggi}$ è la variazione del valore di *S,* nel caso specifico nell'arco di 24 ore. Possiamo indicare tale variazione con *S'* e riscrivere la precedente equazione come:

$$S' = - aSI \qquad (2.6)$$

Essa costituisce *l'equazione della velocità* con la quale varia la grandezza *S* nel corso del tempo. L'equazione dice che tale variazione è legata, secondo il modello proposto, sia al numero *S* di suscettibili che al numero *I* di infetti e ad un parametro *a* che quantifica il grado di contagiosità della malattia e l'interazione sociale. Si osservi che la discussione sopra porta a rappresentare con un segno negativo i flussi uscenti da un comparto e con un segno positivo i flussi entranti.

2.5 Il modello del processo di recupero.

Nel modello S.I.R. qui descritto, il processo di recupero o di guarigione da un contagio viene fatto dipendere esclusivamente dal tipo di malattia considerato. Naturalmente quest'assunzione costituisce un'approssimazione della realtà, in quanto effettivamente il processo di guarigione potrà essere diverso da individuo a individuo e dipendere da diversi fattori quali ad esempio la struttura sociale di una popolazione.
Si prenda come caso la malattia del *morbillo*: studi statistici hanno osservato che in media da essa si guarisce dopo 14 giorni[20]. Tenendo conto di questo dato, se si immagina di osservare la popolazione infetti *oggi* e se si assume che la malattia sia già attiva da giorni, è possibile trovare in tale gruppo individui che si sono ammalati meno di un giorno fa, individui che si sono ammalati da meno di due giorni ma da più di uno e così di seguito fino ad arrivare agli individui ammalatisi 14 giorni fa. Avendo la malattia un tempo di guarigione di 14 giorni non vi potranno essere persone che si sono ammalate più di 15 giorni fa. Si arriverà pertanto a raggruppare le persone degli infetti in 14 gruppi distinti in base a quanto tempo addietro hanno contratto la malattia. Se *I* è il numero degli infetti nell'istante di tempo generico, si può ipotizzare che esso sia suddiviso in maniera uniforme nei 14 gruppi, ogni gruppo è pertanto composto di *I/14* persone. E' anche vero che nell'arco di

[20] http://it.wikipedia.org/wiki/Morbillo.

ventiquattro ore da *oggi* osserveremo la guarigione delle persone appartenenti al 14-esimo gruppo, pertanto domani vi saranno *I/14* persone in più nella popolazione recuperati e *I/14* persone in meno nella la popolazione infetti. Inoltre, essendo la malattia del morbillo tra quelle che conferiscono immunità alla stessa, allora non è possibile per una persona guarita tornare ad essere suscettibile e quindi la popolazione appartenente al gruppo dei recuperati è destinata esclusivamente ad aumentare. Nella Figura 2.2 è illustrato schematicamente questo fenomeno. Le frecce indicano il flusso di persone tra i due comparti.

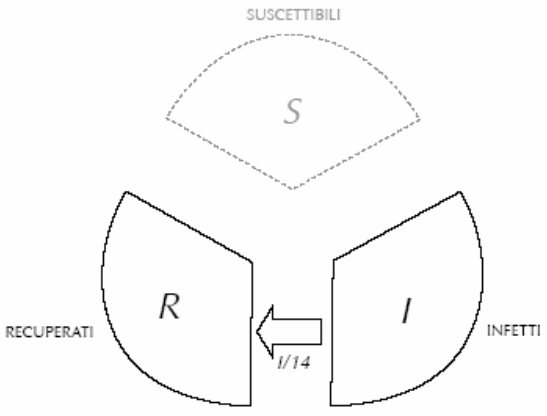

Fig. 2.2: *Rapporto tra i gruppi RECUPERATI e INFETTI per l'epidemia di morbillo.*

In base a quanto sopra detto, è possibile scrivere l'equazione che esprime in termini matematici la variazione giornaliera del valore di *R.* In essa, R_{domani} e R_{oggi} sono rispettivamente il valore della variabile *R* nei giorni di *domani* e *oggi*. Analogamente per I_{oggi}. Pertanto vale:

$$R_{domani} = R_{oggi} + I_{oggi} / 14, \tag{2.7}$$

da cui segue:

$$R_{domani} - R_{oggi} = I_{oggi} / 14. \tag{2.8}$$

La differenza $R_{domani} - R_{oggi}$ è la variazione del valore di *R,* nel caso specifico nell'arco di 24 ore. Indicandola con *R'*, è possibile riscrivere la precedente equazione come:

$$R' = I/14 \tag{2.9}$$

Inoltre, se si usa al posto del valore 1/14 - l'inverso del numero di giorni necessari alla guarigione il parametro *b*, è possibile riscrivere l'equazione in termini più generali come:

$$R' = b\ I. \tag{2.10}$$

L'equazione indica la velocità con cui varia *R* e descrive come calcolare ogni 24 ore il valore di *R* in base al numero di infetti *I* e al tipo di malattia. Si noti che precisamente *R'* è la variazione di *R* nell'unità di *un giorno*, cioè $R' = (R_{domani} - R_{oggi}) / 1$ e si misura pertanto in *persone/giorno* (*persone per giorno*). Da ciò deriva il termine velocità di variazione.

2.6 L' equazione che esprime le variazioni del valore *I* nel tempo.

Si osserva che la popolazione delle persone infette ha un flusso uscente, costituito dalle persone recuperate, che va a sottrarsi al valore di *I* ed il cui valore è *bI* secondo quanto calcolato nel paragrafo precedente, e un flusso entrante, costituito dalle persone suscettibili che si sono infettate nell'arco di una giornata e che è stato determinato sopra essere *aSI*. Mentre il flusso *aSI* si somma agli infetti, il flusso *bI* si sottrae. Quindi l'equazione che esprime la velocità con la quale varia la grandezza *I* nel tempo è:

$$I' = aSI - bI. \tag{2.11}$$

Si suppone che le guarigioni avvengano ad un tasso costante qui indicato con *b*, dando così ad un totale *bI* di guarigioni o recuperi. L'ipotesi che il tasso di guarigione sia costante è molto forte, in particolare si suppone che la probabilità di guarire sia indipendente da quanto a lungo un individuo sia stato infetto[21]. Questa ipotesi equivale al fatto che la durata del periodo di infettività sia esponenziale, ossia se T_I rappresenta la durata del periodo di infettività, allora:

$$P(T_I > t) = e^{-\beta t}. \tag{2.12}$$

Tale ipotesi contrasta con l'evidenza empirica, ma trascurarla comporterebbe problemi computazionali alla soluzione del modello. Il modo usuale di impiegare l'informazione sul periodo di infettività per la stima di β è quella di eguagliare la mediana empirica con la media di T_I: si ha che $E(T_I) = 1/\beta$, si può porre quindi β = 1/mediana empirica.

[21] si è infatti ipotizzato che sulla base del periodo di durata della malattia gli infetti si distribuiscano in maniera uniforme lungo tutto il periodo di malattia.

2.7 Il modello completo.

La Figura 2.3 riassume quanto finora calcolato mostrando i flussi di persone tra i due comparti con la relativa numerosità giornaliera. Le tre equazioni utilizzate per calcolare, ogni 24 ore nel nostro caso, le variazioni delle variabili S,I e R. sono date dal seguente sistema:

$$S' = - aSI, \tag{2.13}$$

$$I' = aSI - bI,$$

$$R' = b\ I.$$

Partendo da valori iniziali $S(0) = S_0$, $R(0) = R_0$ non necessariamente uguale a zero (ci saranno degli individui naturalmente immuni dall'infezione considerata) e $I(0) = I_0$ (un primo nucleo di individui infetti) e fissati dei valori per i parametri a e b, opportuni per una malattia e per il tipo di popolazione sotto esame, è possibile descrivere l'evoluzione della malattia usando il modello matematico.

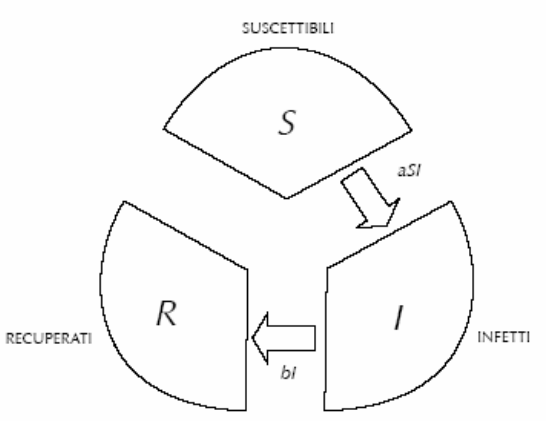

Fig. 2.3: *Modello S.I.R. ed equazioni dei tassi per il morbillo.*

Finora si sono considerate variazioni giornaliere delle tre variabili del modello SIR prendendo come unità di tempo t il giorno composto da 24 ore. Si può riscrivere il modello considerando il tempo una variabile continua e considerare variazioni infinitesime delle tre variabili S, I ed R, il sistema diventa perciò:

$$dS/dt = - \alpha SI, \tag{2.14.1}$$

$$dI/dt = \alpha SI - \beta I, \tag{2.14.2}$$

$$dR/dt = \beta I, \tag{2.14.3}$$

$$S_0>0, \quad I_0>0, \quad R_0\geq 0,$$

$$S_t + I_t + R_t = N.$$

I parametri α e β hanno lo stesso significato dei parametri *a* e *b* precedenti. Nel sistema (2.14) il modello mantiene fissa la popolazione poiché:

$$- \alpha SI + \alpha SI - \beta I + \beta I = 0.$$

α*SI* è una costante cinetica che si basa sull'idea che in unità di tempo il numero di incontri tra suscettibili ed infetti sia proporzionale al numero di ciascuno di questi. Si tratta di un termine non lineare il quale sta ad indicare che le infezioni avvengono ad un tasso alto solamente quando ci sono molti suscettibili e molti *infettivi*, cioè individui in grado di trasmettere l'infezione. I parametri α e β sono entrambi positivi.
Si può esprimere *R* in funzione di *N*, *S* e *I*, ponendo $R = N\text{-}S\text{-}I$[22], per cui la terza equazione diventa ridondante e il modello SIR si riduce a due equazioni:

$$\begin{aligned} dS/dt &= - \alpha SI, \\ dI/dt &= (\alpha S - \beta)I. \end{aligned} \qquad (2.15)$$

La derivata dei suscettibili e infetti è proporzionale al tasso di trasmissione α (probabilità di un infetto di contagiare un individuo suscettibile) e alla concentrazione d'infetti e suscettibili, β tiene conto del fatto che gli individui infetti sono ricoverati o rimossi dalla popolazione; in particolare, il numero degli infetti aumenta allo stesso tasso con cui decrescono i suscettibili e simultaneamente diminuisce secondo il tasso di rimozione o guarigione pari a β. Infine il numero dei rimossi aumenta esattamente allo stesso tasso con cui decrescono gli infetti.
Al fine di generalizzare lo studio, si possono esprimere tutti i parametri in termini di concentrazioni relative dividendo le quantità *S*, *R* e *I* per *N* ottenendo così le quantità *r*, *s* e *i*. È chiaro così che le dinamiche di una popolazione, ad esempio, di 100.000 abitanti e un numero iniziale d'infetti $I(t_0) = 100$ sono identiche a quelle di un'altra popolazione dieci volte minore e con $I(t_0) = 10$.
Dalla (2.14.1) e (2.14.3) dividendo l'equazione dei suscettibili per quella dei rimossi si ottiene:

$$\frac{dS/dt}{dI/dt} = -\frac{\alpha}{\beta}\cdot S \rightarrow \frac{dS}{dI} = -\frac{\alpha}{\beta}\cdot S$$

[22] Dividendo tutte le grandezze per N riducendo il sistema di equazioni in termini di frazioni si ottiene:

$$r(t) = 1 - s(t) - i(t).$$

Integrando questa equazione differenziale, usando il valore iniziale S_0 e ponendo $R_0 = 0$, si ottiene:

$$S(R) = S_0 e^{-\frac{\beta}{\alpha}R}, \tag{2.16}$$

che esprime la dinamica dei suscettibili in funzione di quella del numero dei rimossi.

Si può inoltre ottenere un secondo integrale risolvendo esplicitamente in forma parametrica le equazioni della (2.14): si può infatti scrivere I in funzione di S dividendo la (2.14.2) per la (2.14.1) ottenendo:

$$\frac{dI}{dS} = \frac{-\alpha SI + \beta I}{\alpha SI} = -1 + \frac{\beta}{\alpha S} \tag{2.17}$$

questa è un'equazione separabile che possiamo riscrivere così:

$$dI = [-1 + \frac{\beta}{(\alpha S)}]dS$$

ed integrando si ottiene:

$$I(S) = c_o - S + \beta/a \ln S_0. \tag{2.18}$$

La costante c_0 la si può esprimere in termini dei valori iniziali:

$$c_o = I_0 + S_0 - \beta/a \ln S_0,$$

la variazione dei suscettibili nel tempo è negativa: $dS/dt<0$.

Il fenomeno più importante dal punto di vista qualitativo è il cosiddetto *fenomeno di soglia*: se infatti:

$$S_0 < \frac{\beta}{\alpha}, \tag{2.19}$$

risulta:

$$S_t < \frac{\beta}{\alpha} \quad \text{per t > 0} \tag{2.20}$$

e quindi

$$dI/dt < 0 \text{ per } t > 0,$$

dove la frazione $\rho = \alpha/\beta$ è detto *valore di soglia* che rappresenta anche il numero di contatti. Questo significa che se sin dall'inizio il numero dei suscettibili è inferiore al livello di soglia, l'epidemia non si innesca e il numero degli infetti si estingue decrescendo. Nel caso invece di

$$S_0 > \frac{\beta}{\alpha},$$

risulta inizialmente

$$dI/dt > 0,$$

l'epidemia si propaga, cresce cioè il numero degli infetti fino a raggiungere un suo massimo, ossia finché *S* non scende al di sotto del valore di soglia: ρ= α/β, e poi decresce fino ad estinguersi. Infatti il momento più favorevole allo sviluppo dell'infezione è proprio quello iniziale caratterizzato dalla massima disponibilità di soggetti suscettibili. Il valore di soglia in particolare è dato dal rapporto tra il tasso di rimozione ed il tasso di infezione.
Kermack e McKendrick svilupparono formalmente tutto questo enunciando il teorema della soglia.

Teorema della soglia: *in generale un'epidemia si sviluppa sulla base delle equazioni differenziali date dalla (2.14.1-2.14.3) a partire da valori iniziali (S_0, I_0, 0) dove $S_0 + I_0 = N$. siano S_t ed I_t le soluzioni della (2.15), se (S_0 α/β) ≤ 1 allora I_t decresce a zero per $t \to \infty$. Se (S_0 α/β) >1, allora I_t prima aumenta fino ad un valore massimo I_m pari a $1-R_0-\beta/\alpha-[\ln(S_0\ \alpha/\beta)]/(\beta/\alpha)$ per poi decrescere per $t \to \infty$. La frazione dei suscettibili S_t è una funzione decrescente e il suo valore limite S(∞) per $t \to \infty$ ha unica soluzione nell'intervallo (0, β/α) dell'equazione:*

$$1 - R_0 - S(\infty) + [\ln(S(\infty)/S_0]/(\beta/\alpha) = 0. \qquad (2.21)$$

La prima equazione del sistema 2.14 può essere anche interpretata come l'*incidenza*, ossia il numero di nuovi casi di infetti in una determinata unità di tempo *t*, mentre il secondo addendo della 2.14.2, (β*I*), rappresenta la *prevalenza*, ossia il numero di individui infetti in una certa unità di tempo *t*. Sulla base di queste osservazioni, il teorema della soglia di Kermack e McKendrick stabilisce che se il numero medio di individui infettati da un infetto all'inizio dell'epidemia in una popolazione ossia: $R_0 = \frac{\alpha}{\beta} S_0$ è maggiore di 1, allora l'epidemia si propaga finché la prevalenza (la frazione degli infetti) aumenta fino a raggiungere un valore massimo per poi decrescere a zero. Altrimenti l'epidemia non si innesca quando la prevalenza continua a decrescere fino a zero. Il propagarsi dell'epidemia ad un certo punto si ferma poiché il numero medio di individui suscettibili durante l'epidemia, R_e[23], sarà via via inferiore al numero iniziale S_0. Tuttavia la classe finale dei suscettibili S(∞) non va a zero. Da qui il secondo teorema.

Secondo teorema della soglia: *se S_0 eccede la soglia ρ=β/α di una piccola quantità ε, e se il numero iniziale degli infetti I_0 è piccolo rispetto*

[23] Nel successivo paragrafo verranno descritti più dettagliatamente i parametri del modello.

a ε, allora il numero dei suscettibili presenti nella popolazione alla fine dell'epidemia, è approssimativamente pari a β/α – ε ed $R_\infty \approx 2\,\varepsilon$.

Il principale significato di questi teoremi enunciati e dimostrati da Kermack e McKendrick, è la dimostrazione matematica che anche nelle epidemie più importanti, non necessariamente tutti i suscettibili passano alla condizioni di infetti. Anche se l'epidemia effettivamente dovesse scoppiare, essa infatti non si esaurirà per mancanza di soggetti suscettibili, in quanto la presenza del gruppo *R* impedisce il contagio dell'intera popolazione.
Inoltre se l'epidemia colpisce una popolazione omogenea e non vengono somministrate vaccinazioni, è possibile stimare, a partire dai dati relativi all'epidemia, il numero dei contatti dato da $\rho^{-1} = \alpha/\beta$: infatti quando l'epidemia si trova nella sua fase iniziale il numero degli infetti è talmente piccolo da diventare insignificante per cui - considerando le concentrazioni relative – si può scrivere:

$$S_0 = 1 - R_0$$

e la (2.21) può essere risolta per ρ^{-1} :

$$\rho^{-1} = \frac{\ln(S_0 / S(\infty))}{S_0 - S(\infty)}\,. \qquad (2.22)$$

Alcuni studi sierologici, che si occupano di ricercare anticorpi diretti verso la maggior parte degli agenti patogeni, batterici, virali e parassiti, consentono di poter stimare la frazione dei suscettibili prima e dopo l'epidemia permettendo così di calcolare il tasso di contatto usando proprio la (2.22). Ad esempio, alcune pubblicazioni di indagini sierologiche condotte sugli studenti del primo anno dell'Università di Yale, stimavano la frazione dei suscettibili alla rosolia all'inizio e alla fine del loro anno di immatricolazione, rispettivamente pari a 0.25 e 0.0965; sulla base di tali stime, uno studio di Evans (1982), calcolava secondo la (2.22) il tasso di contatto pari a $\rho^{-1} = 6.2$. Lo stesso studio forniva la frazione iniziale e finale degli studenti suscettibili all'influenza pari a 0.911 e 0.5138 con una stima del tasso di contatto pari a $\rho^{-1} = 1.44$.

2.8 Parametri epidemiologici.

Nel paragrafo precedente, si è visto che, secondo il modello classico S.I.R. di Kermack e McKendrick, l'epidemia si innesca se il numero iniziale di suscettibili S_0 supera il livello di soglia $\rho = \beta/\alpha$. Si è anche

visto che la conoscenza di tutti questi parametri (*N, I*(t_0)*, S*(t_0)*, R*(t_0)*,* α e *ß*) determina l'evoluzione del sistema che, mantenendo costante la popolazione, descrive la "riallocazione" degli individui coinvolti tra i vari comparti. Indagando più a fondo sul significato dei vari parametri coinvolti nel processo epidemico è possibile trarre ulteriori utili informazioni, in particolare per i due parametri fondamentali che sono α e β.

Si consideri una situazione ipotetica ideale in cui si hanno I_0 soggetti infettivi tutti contagiosi allo stesso istante, limitandoci a considerare il solo processo di rimozione, questi individui, si è visto, abbandoneranno la classe infetta al tasso β, ossia secondo l'equazione:

$$\frac{dI}{dt} = -\beta I(t),$$

La cui soluzione è la funzione esponenziale di tasso (-β):

$$I(t) = I_0 e^{-\beta t} .$$

Si può dare un'interpretazione in termini probabilistici: si suppone cioè, che ad ogni singolo individuo sia associato un "rischio di abbandono per rimozione" della classe infettiva pari a β che sia costante nel tempo. Si indica a tal proposito con p(t) = Pr(T >t) la probabilità di essere ancora infetto al tempo t per un soggetto infettato al tempo zero. La definizione di rischio di morte, che nella fattispecie è un rischio di uscita dalla classe infetta, tra l'età t e t + h è:

$$\Pr(t < T < t+h / T > t) = \frac{\Pr(t < T < t+h)}{\Pr(T > t)} = \frac{\Pr(T > t) - \Pr(t+h)}{\Pr(T > t)} = \frac{p(t) - p(t+h)}{p(t)}$$

Con alcune manipolazioni della condizione di soglia (2.16) si può scrivere:

$$\frac{1}{\beta} > \frac{1}{\alpha s_0} \qquad (2.23)$$

che segnala il fatto per cui la malattia infettiva considerata è efficace nel contagio solo se la scala temporale del processo di guarigione è lunga rispetto a quelle del processo di infezione e soprattutto se:

$$R_0 = \frac{\alpha}{\beta} S_0 > 1 .$$

L'interpretazione epidemiologica di quest'ultima relazione è particolarmente importante. Il numero R_0 si può interpretare come il numero medio di individui infettati da un infetto durante il suo intero periodo infettivo (di durata media 1/β) all'inizio dell'epidemia in una

popolazione di $N=S_0$ individui tutti suscettibili; αN - ovvero αS_0[24] - è il tasso a cui un infetto infetta nuovi suscettibili, se il loro numero è fissato a N (o a S_0); $1/\beta$ è la durata media del periodo di infettività; di conseguenza R_0, il prodotto di queste due quantità, è il numero di individui contagiati da un infetto durante il suo periodo di infettività.

Se $R_0 < 1$, ogni infetto produrrà in media meno di un altro infetto; coloro che sono stati contagiati dai primi ne produrranno in media meno di un altro a testa, quindi un totale di R_2 e; andando avanti, si vede che l'epidemia si spegnerà velocemente. Se invece $R_0 > 1$ ogni infetto produrrà in media più di un infetto; questi ne produrranno anch'essi più di un altro, dando luogo ad una reazione a catena con crescita esponenziale del numero di infetti. Tale crescita dovrà però rallentare perché le prime infezioni faranno diminuire il numero di suscettibili, e quindi i successivi infetti non avranno più a disposizione S_0 suscettibili, ma un numero minore, e quindi produrranno meno di R_0 nuovi infetti; quando i suscettibili saranno scesi sotto il valore β/α, ogni nuovo infetto produrrà in media meno di un nuovo infetto e quindi l'epidemia si fermerà.

Il parametro $\rho = \beta/\alpha$ (il valore di soglia) è detto *tasso relativo di rimozione*, poiché misura quanto la rimozione (guarigione o morte) è più veloce dell'infezione, il suo inverso $\rho^{-1} = \alpha/\beta$ è detto *tasso di contatto*. Si è visto che R_0 - detto anche tasso di crescita iniziale dell'epidemia - è il numero medio di individui infettati da un infetto all'inizio dell'epidemia, mentre il numero $R_t = (\alpha/\beta)S_t$ è anche detto *tasso riproduttivo* dell'infezione o tasso di infezione nella popolazione durante tutto il periodo di infettività: esso misura quante infezioni secondarie sono prodotte da ogni infezione primaria quando l'infezione è introdotta nella popolazione. Si osserva che quando l'intera popolazione è composta da soli suscettibili, le tre quantità R_0 ρ^{-1} ed R_t sono uguali all'inizio dell'espandersi della malattia infettiva. Inoltre R_0 è una quantità definita solo all'inizio dell'epidemia, mentre ρ^{-1} ed R_t sono definiti in ogni momento. Tuttavia poiché nel modello S.I.R. il tasso di contatto rimane costante, R_t e ρ^{-1} sono uguali e possono essere interscambiate nell'applicazione del teorema enunciato nel paragrafo precedente. Inoltre, il tasso riproduttivo R_t è sempre inferiore a R_0, il tasso iniziale riproduttivo dell'infezione, in quanto la parte di individui suscettibili che

[24] si suppone infatti che il numero iniziale dei suscettibili sia pari all'intera popolazione: $S_0 = N$.

sono colpiti durante l'epidemia è inferiore al numero dei suscettibili all'inizio della malattia. Da questo risulta che:

$$R_0 \geq R_t,$$

nel caso in cui le tre quantità siano uguali all'inizio dell'epidemia[25]. Infine, supposta soddisfatta la condizione di soglia, l'impatto finale dell'epidemia dipenderà dalla dimensione dei due parametri α e β.

2.9 Proprietà del modello S.I.R.

Ai fini di una analisi qualitativa della dinamica del modello S.I.R. sono importanti gli stati di equilibrio, cioè gli stati in corrispondenza dei quali la componente infettiva è nulla, e quindi il processo infettivo non può aver luogo. Il modello relativo ad una popolazione di *N* individui avente un dato coefficiente di infettività e un dato coefficiente di recupero può essere decritto da un *vettore di stato*:

$$x(t)= [s(t), i(t), r(t)]^T$$

dove:

$$1\geq s \geq 0, \quad 1\geq i \geq 0, \quad 1 \geq r \geq 0, \quad \text{e} \quad s + i + r = 0.$$

Si può verificare che gli stati di equilibrio hanno la forma:

$$[(\beta/\alpha), 0, 1- (\beta/(\alpha N))].$$

Gli stati di equilibrio sono semplicemente stabili, nel senso che qualunque movimento del processo risulta sempre limitato, in particolare: il sistema S.I.R. soddisfa alle seguenti proprietà:

- il sistema è positivo;
- il vettore $x =[1, 0, 0]^T$ è uno stato di equilibrio;
- il vettore $[(\beta/\alpha), 0, 1- (\beta/(\alpha N))]$ con $1> s >0$ è stato di equilibrio se e solo se $R_0 = \alpha N/\beta > 1$.

Da ciò consegue che, dato un sistema S.I.R. con $R_0 = \alpha N/\beta > 1$ e stato iniziale:

$$x(0)=[1 - i(0), i(0), 0],$$

dove $i(0) > 0$ rappresenta la frazione iniziale di infetti supposta nota, sulla base del teorema enunciato nel paragrafo precedente, è possibile calcolare le seguenti grandezze:

[25] In molti modelli $R_0 = \rho^{-1}$ e $\rho^{-1} > R$ in tutti i modelli dopo l'espandersi della malattia.

- valore finale di suscettibili $s^{\infty} = \lim_{t\to\infty} s(t) \geq 0$ e ristabiliti $r^{\infty} = \lim_{t\to\infty} r(t) \geq 0$;
- il tempo di picco t_p e valore di picco degli infetti $i(t_p) = \max_{t\in(0,+\infty)} i(t)$;

i cui valori sono quelli calcolati secondo il teorema della soglia di Kermack e McKendrick.

Si è visto inoltre, che affinché il processo epidemico abbia inizio, è necessario che inizialmente il numero dei suscettibili sia maggiore del livello di soglia ρ = β/α: in tal caso il numero degli infetti continuerà ad aumentare (in particolare se il numero iniziale degli infetti è piccolo, allora la dinamica di I sarà esponenziale al tasso ($\alpha S_0 - \beta$)). La crescita degli infetti continuerà, anche se a ritmi sempre più lenti, finché non verrà raggiunto il picco dell'epidemia che si trova proprio in corrispondenza del valore di soglia ρ = β/α. Poi il numero degli infetti comincerà a ridursi fino ad esaurimento dell'epidemia per mancanza di materiale infetto. Dunque il massimo di I si raggiunge per S = ρ.

In particolare notiamo:

$$\frac{d^2 I}{dS^2} = \frac{d}{dS}\left[\frac{dI}{dS}\right] = \frac{d}{dS}\left[-1 + \frac{\beta}{\alpha S}\right] = -\frac{\beta}{\alpha S^2} \leq 0 \qquad (2.24)$$

che giustifica la forma concava delle soluzioni.

Per meglio capire la dinamica temporale dei tre gruppi S, I ed R del modello, si consideri un esempio numerico in cui al tempo t_0 si suppone che ci sia uno stato iniziale dato dal seguente vettore:

$$X_0 = (N = S(t_0),\ I(t_0),\ R(t_0),\ \alpha,\ \beta) = (10000,\ 1,\ 0\ ,0.00035,\ 0.1)$$

e, a partire da tali condizioni iniziali, attraverso la soluzione del sistema di equazioni differenziali dato dalla (2.14.1 - 2.14.3) si ricavino le dinamiche nel tempo delle tre variabili *suscettibili*, *infetti* e *rimossi*, ottenendo così il grafico in Figura 2.1:

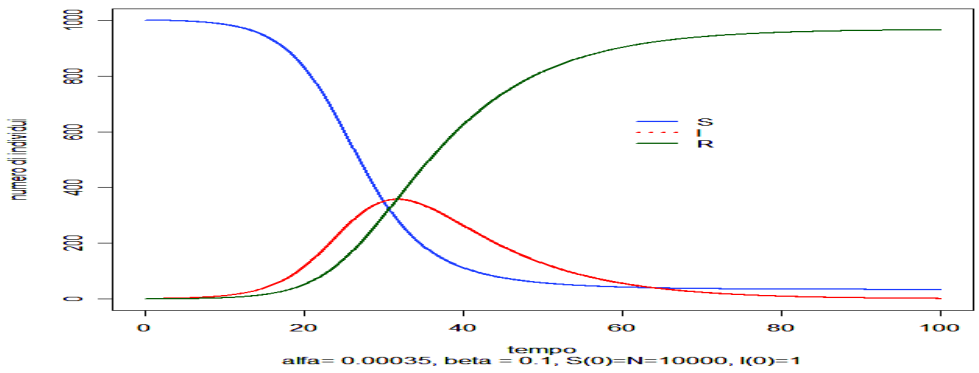

Fig. 2.4: *dinamica temporale del modello KMK.*

Come detto in precedenza, l'epidemia si conclude con l'esaurimento degli infetti e senza che l'intero gruppo dei suscettibili sia stato contagiato: si è infatti visto che $S(\infty) \geq 0$ (in quanto $S(t)$ è funzione non crescente e non negativa) e che $R(\infty) \leq N$ in quanto $R(t)$ è funzione crescente ma limitata superiormente da N[26]. Si è inoltre visto che dalla (2.16) si otteneva l'equazione dei suscettibili in funzione dei rimossi. Poiché in ogni caso vale $R \leq N$ avremo:

$$S(R) = S(R(t)) = S_0 e^{-\frac{\alpha}{\beta}R} \geq S_0 e^{-\frac{\alpha}{\beta}N} > 0 \qquad \forall t \tag{2.25}$$

e quindi in particolare

$$S(t = \infty) > 0,$$

strettamente.

Nel paragrafo 2.7, si è visto che si possono esprimere gli infetti in funzione dei suscettibili: $I = f(S)$, in maniera tale che ad ogni istante t si abbia $I(t) = f(S(t))$, ottenendo così il valore massimo degli infetti in base ai suscettibili. Infatti dalla (2.17) si è visto che:

$$\frac{dI}{dS} = \frac{-\alpha SI + \beta I}{\alpha SI} = -1 + \frac{\beta}{\alpha S}. \tag{2.26}$$

Questa equazione la si può riscrivere anche in questo modo:

$$dI = \left[-1 + \frac{\beta}{\alpha S}\right] dS \tag{2.27}$$

e integrando si ottiene:

$$I(S) = c_0 - S + (\beta/\alpha)\ln(S), \tag{2.28}$$

dove

$$c_0 = I_0 + S_0 - (\beta/\alpha)\ln(S_0). \tag{2.29}$$

Questo consente di costruire il grafico delle fasi S ed I. Nell'esempio considerato, ponendo a zero la derivata $I'(t)$, otteniamo il corrispondente valore di $S^* = \alpha/\beta$ tale per cui l'equazione degli infettivi è a inclinazione nulla:

$$I'(t) = \frac{\beta}{\alpha S} - 1 = \frac{0.1}{0.00035(S)} - 1 = 0$$

$$S = 286 \Rightarrow \frac{dI}{dS} = 0$$

[26] Si è infatti visto dal teorema che i due i due limiti esistono e sono finiti.

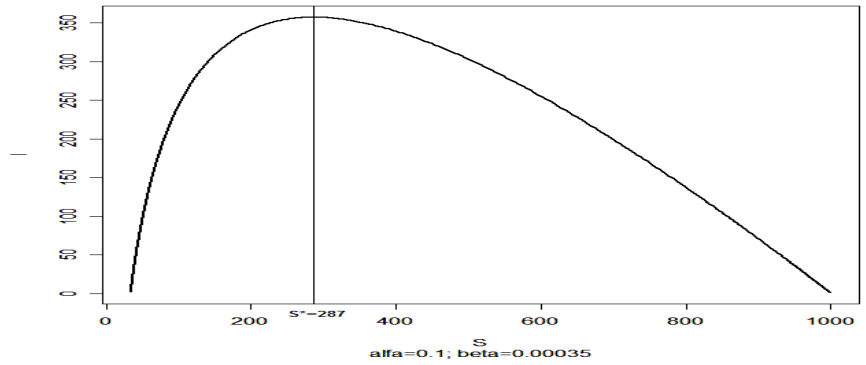

Fig. 2.5: *piano delle fasi (S,I) del modello KMK.*

La Figura 2.5 va letta da destra verso sinistra: al diminuire di *S* aumenta *I*. *S*(*t*) è decrescente, infatti *S*'(*t*) = - α*I*(*t*) < 0. Se *S* ≤ 286, *I*(*t*) è decrescente; infatti:

$$t_1 < t_2 \Rightarrow S(t_1) > S(t_2) \Rightarrow I(S(t_1)) > I(S(t_2)) \Rightarrow I(t_1) > I(t_2)$$

e quindi l'epidemia viene stroncata sul nascere.

Nel grafico della Figura 2.6 si riportano infine differenti possibili dinamiche corrispondenti a differenti condizioni iniziali mantenendo però identici i parametri α e *ß*:

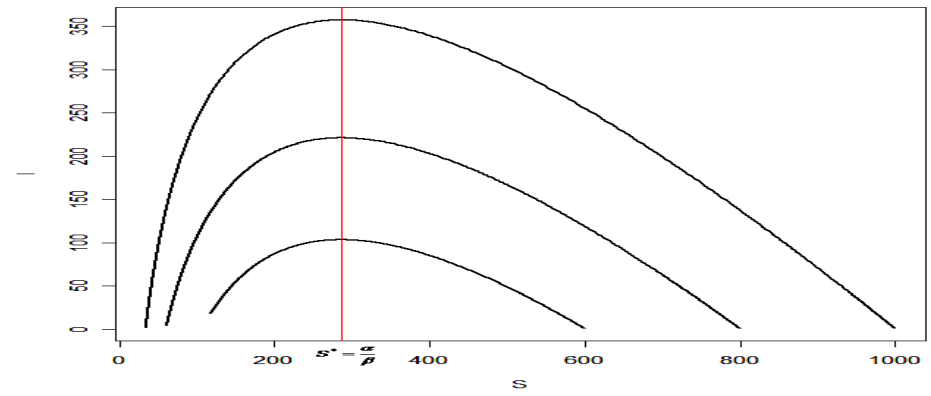

Fig. 2.6: *piano delle fasi (S,I) del modello KMK con differenti condizioni iniziali.*

Si nota come differenti condizioni iniziali diano luogo a differenti equilibri finali, una caratteristica tipica dei sistemi non lineari.

2.10 Esempio numerico del modello S.I.R. DI Kermack e McKendrick.

Nei paragrafi precedenti si è fornita la descrizione del modello deterministico S.I.R. di Kermack e McKendrick, che consiste nella soluzione matematica di due equazioni differenziali che descrivono la

variazione dei suscettibili e la variazione degli infetti. Si è anche visto che affinché l'epidemia si propaghi, deve essere soddisfatta la condizione di soglia così come enunciato dal teorema di Kermack e McKendrick (paragrafo 2.7). Si considera ora un esempio numerico.

Supponiamo che dopo aver passato le vacanze natalizie, dei 502 studenti di un collegio, 2 presentino sintomi influenzali, seguiti il giorno dopo da altri 3. Si suppone che L'influenza duri in media 2 giorni e il tempo di incubazione 0. L'unità di misura del tempo è data da 1 giorno. In tal caso abbiamo:

$$S(0) = 500,\ I(0) = 2,\ R(0) = 0$$
$$S(1) = 497,\ I(1) = 2 + 3 = 5,\ R(1) = 0$$
$$S(2) = \text{incognito},\ I(2) = \text{incognito},\ R(2) = 2$$

Il sistema (2.14.1) del paragrafo 2.8 diventa:

$$S'(t) = -a\, I(t)\, S(t) \qquad (2.30)$$
$$I'(t) = a\, I(t)\, S(t) - b\, I(t) \qquad (2.31)$$
$$R'(t) = b\, I(t) \qquad (2.32)$$

In base ai dati dell'esempio, possiamo ricavare una stima approssimata dei parametri *a* e *b*:

$$a \approx -\frac{S(1)-S(0)}{I(0)\cdot S(0)} = \frac{497-500}{2\cdot 500} = 0.003\,, \qquad (2.33)$$

$$b = \frac{1}{2} = 0.5\,,. \qquad (2.34)$$

la (2.33) è stata calcolata sfruttando il fatto che S'(0)= S(1)-S(0) = - $a\cdot$ I(0) $\cdot$ S(0), per cui nell'esempio numerico, sapendo che l'unità di misura del tempo è di un giorno0, si ha: (497 -500)/1 = - 3 = - $a \cdot 2 \cdot 500$ = 1000, da cui sui ricava che a = 3/1000 = 0.003.

Il parametro *b* della (2.34) è invece dato dall''inverso del tempo medio di durata della malattia. Si è detto che l'epidemia si propaga solo se $R_0 \geq 1$, nell'esempio:

$$R_0 = \frac{\alpha}{\beta} S_0 = \frac{0.003}{0.5} \cdot 500 = 3 > 1\,.$$

In base dunque ai parametri iniziali, si simuli[27] il modello deterministico S.I.R. di Kermack e McKendrick risolvendo il sistema di equazioni

[27] **Codice R**

```
library(odesolve)
parms <- c(alfa=0.003, beta=0.55)
inits <- c(S=500, I=2, R=0)
dt    <- seq(0,100,1)
N=500
R_0 <- with(as.list(parms),{alfa*N/beta})
print(paste("R_0 =",R_0),quote=FALSE)
SIR <- function(t, x, parms){
with(as.list(c(parms,x)),{
```

differenziali del sistema (2.14)[28]; i risultati che si ottengono sono i seguenti:

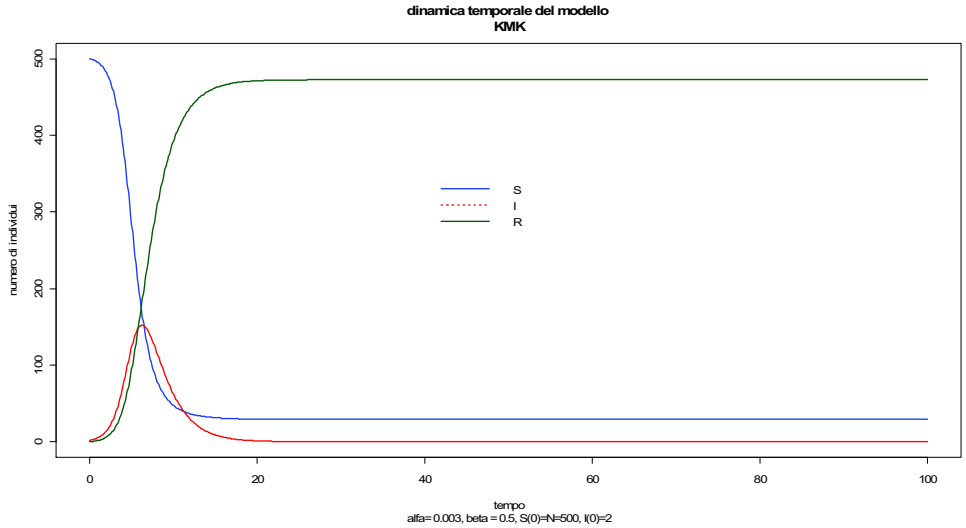

Fig. 2.7: *soluzione numerica delle (.30) e (31) per i parametri iniziali α = 0.003, β = 0.5, I = 2 ed S=N=500. Vengono rappresentate le funzioni S(t), I(t) ed R(t), notare che per* $t \to \infty$ $I(t) \to 0$, *ma* $S(t) \to S_1 \neq 0$, *restano dei suscettibili non infetti.*

La Figura 2.7 evidenzia l'andamento classico del modello S.I.R.: in particolare la curva degli infetti raggiunge un massimo in corrispondenza del livello di soglia dato da:

$$\rho = \frac{\beta}{\alpha} = \frac{0.5}{0.003} \approx 167 .$$

Il numero massimo di suscettibili oltre il quale gli infetti non possono più aumentare. Per determinare il numero massimo di infetti raggiunto, sfrutto la (2.28) e la (2.29) che permette di esprimere gli infetti in funzione dei suscettibili, ottenendo cosi:

$$I(S) = 167 \ln(S) - S - 536 \qquad (2.35)$$

dove -536 è la costante c_0 ricavata dalla (2.29):

$$c_0 = I_0 + S_0 - (\beta/\alpha)\ln(S_0) = 2 + 500 - (0.5/0.003)\ln(500) = -536$$

```
dS <- - alfa*S*I
dI <- + alfa*S*I - beta*I
dR <- beta*I
der <- c(dS, dI,dR)
list(der)
})
}
sim1 <- as.data.frame(lsoda(inits, dt, SIR, parms=parms))
attach(sim1)
plot(time, S, type="l", col="blue", lwd =2,xlim=c(0,30), ylim=c(0,N), xlab="tempo",
ylab="numero di individui", main="dinamica temporale del modello KMK", sub="alfa=
0.003, beta = 0.5, S(0)=N=500, I(0)=2")
lines(time, I, type="l", lwd =2,col="red")
lines(time, R, type="l", lwd =2,col="darkgreen")
legend(locator(1),legend=c("S","I","R"),col=c("blue","red","darkgreen"),lty=c(1,3),lwd=
c(2,2,2),bty="n")
```

[28] Dopo aver impostato I parametri iniziali, ho utilizzato la libreria del programma R: *odesolve*, impostando come unità di misura il giorno, e fissando 100 step per il calcolo del differenziale.

A questo punto si può creare il grafico[29] delle varie fasi *S*, *I* e calcolare il numero massimo di individui infetti semplicemente sostituendo alla (2.35) il valore di soglia pari a 167, ottenendo un $I_{max} \approx 152$, come mostrato in Figura 2.8.

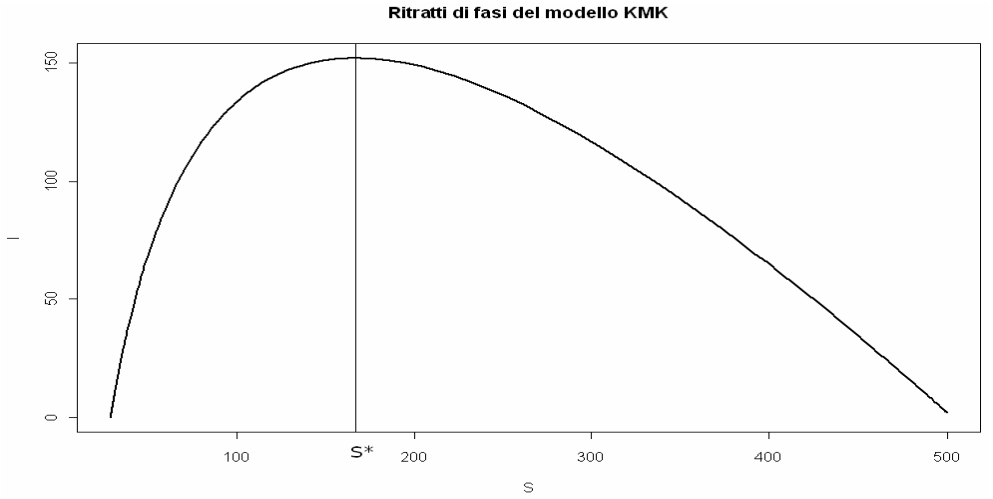

Fig. 2.8: *grafico della (2.35) per i parametri iniziali α = 0.003, β = 0.5, I = 2 ed S=N=500. La curva va letta da destra a sinistra al variare di t come segue da S'(t) ≤ 0. Si nota che I(t) si annulla per un tempo t_1 finito, per $t \to \infty$ $I(t) \to 0$. D'altra parte $S(t) \to S_1 \neq 0$ per $t \to \infty$.*

2.11 Aspetti stocastici del modello.

Nel paragrafo precedente si è descritto uno dei modelli più semplici capace di spiegare i meccanismi intrinseci di trasmissione di un'epidemia; in particolare si è visto che il modello classico S.I.R. proposto da Kermack e McKendrick, espone la dinamica di una popolazione nel tempo considerato come variabile continua ed è descritto tramite una coppia di equazioni differenziali di primo ordine. Tuttavia, per simulare un'epidemia, il modello prevede un approccio puramente deterministico che, mantenendo inalterati i parametri e le

[29]
```
library(odesolve)
parms <- c(alfa=0.003, beta=0.5)
inits <- c(S=500, I=2, R=0)
dt    <- seq(0,100,0.1)
N=1000
R_0 <- with(as.list(parms),{alfa*N/beta})
print(paste("R_0 =",R_0),quote=FALSE)
SIR <- function(t, x, parms){
with(as.list(c(parms,x)),{
dS <- - alfa*S*I
dI <- + alfa*S*I - beta*I
dR <- beta*I
der <- c(dS, dI,dR)
list(der)
})
}
simulation <- as.data.frame(lsoda(inits, dt, SIR, parms=parms))
attach(simulation)
plot(simulation$S,simulation$I,type="l",xlab="S",ylab="I", lwd=2,main="Ritratti di fasi
del modello KMK")
abline(v=0.5/0.003)
```

condizioni iniziali, produce sempre gli stessi risultati riuscendo a descrivere la dinamica di un'epidemia tanto più precisamente quanto più grande è il numero di soggetti che risulta infetto. Nel modello S.I.R. infatti tutti infettano alla stessa velocità e altrettanto è fisso il tempo in cui i soggetti malati vengono rimossi dalla popolazione circolante.
Nella realtà tuttavia ogni epidemia presenta una variabilità intrinseca legata a numerosi aspetti non considerati nel modello, che rendono difficile ottenere stime attendibili, soprattutto in presenza di uno scarso numero di casi, come avviene nelle fasi iniziali e terminali di diffusione di una infezione nella popolazione. Tale condizione può essere valutata introducendo nel modello una componente stocastica.
Il modo per introdurre la casualità in un modello matematico, è quello di formulare i tassi in termini di probabilità con cui accadono gli eventi, per cui il tasso di contatto diventa la *probabilità* che in individuo suscettibile diventi un infetto, mentre il tasso di rimozione diventa la *probabilità* che un soggetto ammalato guarisca. In particolare, indicando con a_i la probabilità del generico evento i, si può esprimere, ad esempio, con a_i, la probabilità che al tempo t un nuovo suscettibile diventi un infetto:

$$a_i = \alpha SI. \tag{2.36}$$

Se ciò avviene, il valore di S diminuisce di uno, mentre il valore di I aumenta di 1.
In altre parole, indicando sempre con S_t il numero dei suscettibili di una popolazione al tempo t, ed N il numero totale al tempo t=0, si ipotizza che S_t sia una variabile casuale e che la probabilità che s suscettibili al tempo t sia data da $p_s(t)$.

2.12 Simulazione stocastica del modello S.I.R.: algoritmo di Gillespie.

Sebbene il modello stocastico S.I.R. sia concettualmente più complesso di quello deterministico, non è tuttavia difficile da simulare; questo perché nel modello probabilistico, le variabili continue del modello matematico, sono sostituite con variabili discrete che vengono più facilmente elaborate dal calcolatore. Per simulare la stocasticità del modello è però necessario un generatore di numeri casuali.
Un primo modo per simulare il modello stocastico S.I.R., potrebbe essere quello di fissare per la simulazione, dei "passi" temporali talmente piccoli l'uno dall'altro, in modo tale che ad ogni *step* possa

accadere solamente un evento, evitando così il fatto che due o più eventi possano accadere nello stesso passo temporale. Il problema è che con questo modo di procedere, nella maggior parte degli *step* non succede alcun evento (ad esempio non si hanno nuovi infetti, o non si ha alcun guarito) rendendo così non efficiente la simulazione. Per ovviare a tale inconveniente, esistono alcuni algoritmi tra cui anche quello di *Gillespie*, il quale oltre ad essere molto efficiente, è anche un buon strumento per simulare modelli stocastici.

Negli anni '70, Daniel T. Gillespie (1977, 1976) sviluppò un metodo di esatta simulazione stocastica (SSA) applicato in ambito chimico, per studiare la velocità con cui avviene una reazione chimica e tutti i fattori in grado di influenzarla. Benché il metodo SSA e tutte le sue diverse implementazioni siano state principalmente concepite per modelli utilizzati in campo chimico e biologico, la procedura può essere ugualmente applicata a qualsiasi sistema temporale continuo che può essere descritto da coppie di equazioni differenziali di primo ordine e quindi anche per descrivere l'andamento nel tempo di popolazioni finite.

Nell'ambito del presente studio, l'algoritmo di simulazione stocastica di Gillespie (SSA), è quindi una procedura che studia nel tempo, trattato come variabile continua, la dinamica di popolazioni finite, generando per esse "traiettorie" statisticamente corrette. Benché l'impostazione originaria dell'algoritmo di Gillespie fosse numericamente esatta, presentava però alcune difficoltà a livello computazionale per la maggior parte delle applicazioni pratiche. Per ovviare a questo problema, nel corso degli anni sono stati quindi sviluppati numerosi metodi approssimativi alla procedura SSA.

I metodi SSA descritti nel presente studio, sono quelli implementati usando il pacchetto di R *GillespieSSA,* disponibili nel sito ufficiale del software. Il pacchetto è stato concepito per un suo utilizzo semplice ed intuitivo, allo scopo di facilitare l'elaborazione, lo sviluppo e la stima di modelli stocastici temporali, presentando allo stesso tempo una buona flessibilità che consente la predisposizione di un gran numero di modelli anche più complessi.

2.13 Descrizione dell'algoritmo di Gillespie.

Nel paragrafo precedente si è detto che l'algoritmo di simulazione stocastica SSA costruisce traiettorie simulate di popolazioni omogenee finite nel tempo continuo: indicando con $X_t(t)$ il numero degli individui di

una popolazione con *i*: 1,……N, l'SSA assume che la popolazione sia composta da un finito numero di individui distribuito su un insieme finito di "stati". Gli stati rappresentano le condizioni in cui si trovano gli individui in un determinato istante *t*: nel modello SIR qui trattato, gli stati o condizioni corrispondono ad individuo *suscettibile* (che si trova cioè nello stato o condizione di *suscettibile*), individuo *infetto* e individuo "recuperato" o *guarito*. Gli stati sono rappresentati da un vettore di stato: $X(t) \equiv (X_1(t)\ldots\ldots\ldots X_N(t))$ dove $X_i(t)$ è l'ampiezza della popolazione allo stato *i* al tempo *t* ed *N* è il numero degli stati. Supponendo che il sistema iniziale al tempo t_0 si trovi nello stato iniziale $X(t_0) = x_0$, l'algoritmo genera l'evoluzione temporale del vettore di stato $X(t) \equiv (X_1(t)\ldots\ldots\ldots X_N(t))$. All'interno della popolazione gli individui interagiscono tra loro attraverso $M \geq 1$ reazioni rappresentate dal vettore $\mathbf{R} = (R_1\ldots\ldots R_M)$, dove R_j rappresenta la *j*-esima reazione: una reazione è definita come un evento che varia istantaneamente lo stato di almeno un singolo individuo (ad esempio, nascita, morte, infezione) modificando così l'ampiezza di almeno uno stato in cui trova la popolazione. Ogni reazione R_j è caratterizzata da due quantità ad essa associate:

- il vettore di variazione dello stato: $\nu \equiv (\nu_{1j}\ldots\ldots\ldots\nu_{Nj})$ dove ν_{ij} è la variazione della popolazione che si trova nello stato *i* causata dall'unica reazione R_j;
- la *property function* indicata con $a_j(x)$ definita come la probabilità che una reazione R_j avvenga nella popolazione nel successivo infinitesimale di tempo $[t, t + dt]$[30];

In altre parole, se il sistema si trova nello stato x, assumendo $X(t) = x$ e avviene la reazione R_j, allora il sistema passa istantaneamente allo stato $x + \nu_j$.

Esistono diverse applicazioni dell'algoritmo di Gillespie, quella più semplice è denominata *Direct Method*, considerato uno standard tra gli algoritmi di selezione stocastica. Il metodo diretto dell'algoritmo di Gillespie calcola esplicitamente quale reazione R_j dovrà accadere e a quale periodo di tempo τ.

Con tale metodo, l'algoritmo genera due numeri casuali r_1 ed r_2 da una distribuzione uniforme nell'intervallo (0,1), con i quali si determina il passo temporale in cui avverrà la prossima reazione; in particolare esso sarà determinato da:

[30] Che indicherò con $[t, t + \tau]$.

$$\tau = \frac{1}{a_0(x)} \ln(1/r_1), \tag{2.37}$$

dove:

$$a_0(x) = \sum a_j(x).$$

$a_0(x)$ è la sommatoria delle singole funzioni di propensione associate a ciascuna delle j reazioni R_j. In base all'algoritmo, l'indice che farà scattare la reazione R_j sarà dato dall'intero j più piccolo che soddisfa la seguente diseguaglianza:

$$j = \sum_{i=1}^{j} a_i(x) > r_2 a_0(x). \tag{2.38}$$

A questo punto, la reazione avrà così luogo sostituendo l'instante t con $t + \tau$ e x con $x + \nu_j$.

Supponendo che il tempo corrente sia t, secondo l'algoritmo l'evoluzione del sistema è quindi simulata in relazione alle due seguenti funzioni di densità, la prima è:

$$P(\tau, j \mid x) = a_j(x) \cdot e^{-a_0(x)\cdot\tau}, \tag{2.39}$$

dove $a_0(x) = \sum a_j(x)$. La (2.39) è detta *Funzione di densità di probabilità della reazione* R_j e rappresenta la probabilità congiunta che la prossima reazione sia data da R_j e che avvenga nell'intervallo infinitesimale $[t, t + \tau]$.

La seconda è:

$$P(\tau \mid x,t) = a_0(x) e^{-a_0(x)\cdot\tau}, \qquad \tau \geq 0, \tag{2.39}$$

ossia il tempo del prossimo evento, $\tau = t + dt$, è dunque un numero casuale scaturito da una distribuzione esponenziale regolata dalla somma di tutte le probabilità di ciascun accadimento e con media $\frac{1}{a_0(x)}$.

La (2.39) consente di derivare correttamente le singole distribuzioni di probabilità per R_j e τ: per primo si integra la (2.39) da 0 a ∞: $P(\tau \mid x,t) = a_0(x) e^{-a_0(x)\cdot\tau}$:

$$P(j) = \int_0^{\infty} a_j(x) \cdot e^{-a_0(x)\cdot\tau} d\tau = \frac{a_j}{a_0}, \tag{2.40}$$

mentre $P(\tau \mid x,t)$ è dato dalla somma per j della probabilità congiunta $P(\tau, j \mid x,t)$ che appunto danno la (2.39).

Riassumendo, i passi dell'algoritmo sono dunque i seguenti:

1. imposta lo stato iniziale (il tempo, il numero di individui, ecc.),

2. calcola il valore di propensione a_j per ogni reazione,
3. sceglie l'azione *j* in accordo con la (2.40),
4. sceglie τ in accordo con la (2.39),
5. aggiorna il vettore di stato in accordo con la reazione *j* selezionata e aumenta *t* in base a τ,
6. torna al passo 2.

2.14 Implementazione dell'algoritmo di Gillespie al modello S.I.R. di Kermack e McKendrick.

L'algoritmo di simulazione stocastica (SSA) precedentemente descritto, viene qui implementato usando il pacchetto di R denominato *GillespieSSA*, facilmente scaricabile dal sito *Comprehensive R Archive Network*[31].

La principale funzione di interfaccia della libreria *GillespieSSA* è *ssa*(·), che è il comando principale per lanciare la simulazione. Prima di fare questo però, il modello stocastico ha bisogno delle quattro componenti descritte nel paragrafo precedente: il vettore di stato iniziale $X(t_0) = x_0$, il vettore di variazione dello stato: $\nu \equiv (\nu_{1j}.....\nu_{Nj})$[32], il vettore delle funzioni di propensione $a_j(x)$ e il tempo finale della simulazione *tf*.

Si è visto che il modello S.I.R. di Kermack e McKendrick consiste principalmente di tre equazioni differenziali:

$$dS/dt = - \alpha SI,$$
$$dI/dt = \alpha SI - \beta I,$$
$$dR/dt = \beta I,$$

dove α e β sono rispettivamente i tassi di contagio della malattia e il tasso di guarigione. Il vettore dello stato iniziale $X(t_0)$ sarà:

$$X(t_0) = (S_0, I_0, R_0).$$

Riprendendo l'esempio numerico del paragrafo 2.10, si può scrivere:

$$X(t_0) = (S_0, I_0, R_0) = (500, 2, 0)$$

In base all'algoritmo, questo vettore definisce poi l'ampiezza di ciascun gruppo in cui è stata suddivisa la popolazione. Gli elementi del vettore devono essere nominati con la stessa notazione utilizzata per le variabili di stato nel definire le funzioni di propensione.

---

[31] Al seguente link: http://CRAN.R-project.org/package=GillespieSSA.

[32] In questo caso si tratta di una matrice in quanto si tratta di un sistema di due equazioni relative a due diversi stati.

Il sistema consiste nelle seguenti due funzioni di reazione:

$$R_1 : S + I \xrightarrow{\alpha} 2I$$
$$R_2 : I \xrightarrow{\beta} R$$

La prima riflette il fatto che un incontro casuale tra un individuo infetto e un suscettibile da due infetti con probabilità α, mentre la seconda indica la rimozione degli infetti ad un tasso β.
Le corrispondenti funzioni di propensione sono date da:

$$a_1 = \alpha SI$$
$$a_2 = \beta I$$

Il vettore *a* delle *property function* definisce la probabilità che una particolare reazione avvenga in una variazione di tempo infinitesimale [*t*, *t* + *dt*]; nello specifico: *a* = [α*SI*, β*I*].
Infine la matrice di variazione degli stati è data da:

$$v = \begin{bmatrix} -1 & 0 \\ +1 & -1 \\ 0 & +1 \end{bmatrix},$$

dove le righe sono gli stati *i* e le colonne sono le reazioni *j*.
Riprendendo l'esempio numerico del paragrafo (2.10), i quattro elementi sono dunque:

- X(t_0) = (500, 2, 0)
- a = [α*SI*, β*I*]
- $v = \begin{bmatrix} -1 & 0 \\ +1 & -1 \\ 0 & +1 \end{bmatrix}$
- *tf*=100

dove i parametri α e β sono rispettivamente pari a 0.003 e 0.5.
In base dunque a questi valori iniziali si avvia la simulazione utilizzando il comando principale *ssa* della libreria di R scegliendo il metodo diretto.
Il grafico ottenuto è riportato in figura 2.9[33]

---

[33] Codice R:

```
library(GillespieSSA)
par <- c(alfa=.003, beta=.5)
#imposto il sistema
x0 <- c(S=500, I=2, R=0)
tf=100                                          # vettore di stato iniziale
nu <- matrix(c(-1,0,1,-1,0,1),nrow=3,byrow=T)   # matrice di variazione dello stato
a  <- c("alfa*S*I", "beta*I")                   #vettore delle funzioni di
                                                propensione: Propensity functiontf
                                                <- 100  tempo finale o numero di
                                                passi

nomeSim <- "Kermack-McKendrick SIR"
```

- ***Avvio della simulazione***

```
# Direct method
```

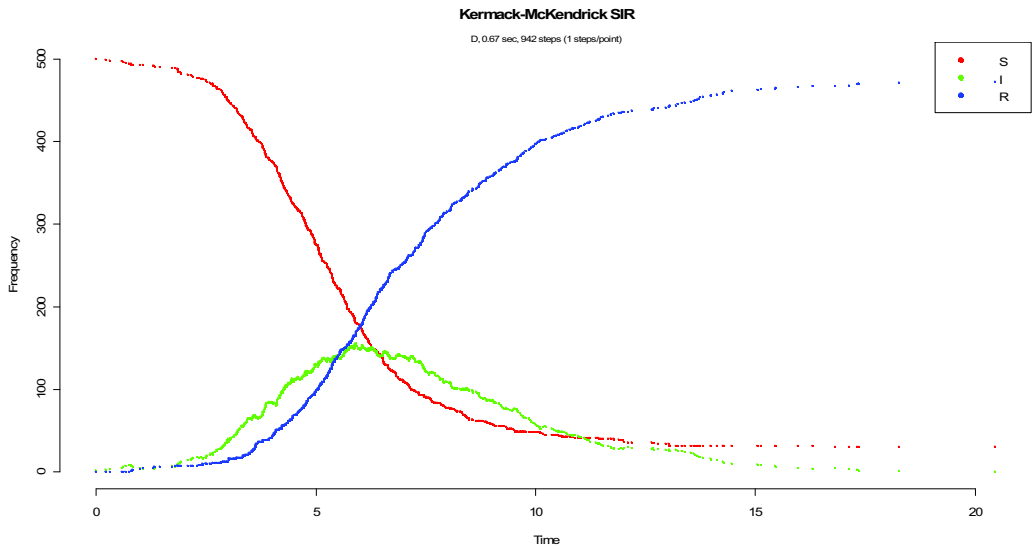

Fig. 2.9: *simulazione stocastica delle (.30) e (31) mediante l'applicazione dell'algoritmo di Gillespie per parametri iniziali a = 0.003, β = 0.5, I = 2 ed S=N=500. Vengono rappresentate le funzioni S(t), I(t) ed R(t).*

Il grafico della Figura 2.9 evidenzia l'andamento del modello S.I.R. ottenuto mediante simulazione stocastica: partendo dunque dal modello deterministico evidenziato dalle(2.30) - (2.31) del paragrafo 2.10, si è derivato il modello stocastico per una popolazione finita. Confrontando i dati di entrambe le simulazioni non si notano grandi differenze: in particolare si ottiene che il numero massimo di infetti scaturito dalla simulazione stocastica, è pari a 150 contro i 152 del modello

```
set.seed(2)

out <- ssa(x0,a,nu,par,tf,method="D", nomeSim,verbose=TRUE,consoleInterval=1)
Running D method with console output every 1 time step
Start wall time: 2009-12-26 17:58:13...
t=0 : 500,2,0
(0.02s) t=1.072276 : 493,4,5
(0.03s) t=2.084362 : 481,14,7
(0.06s) t=3.008038 : 449,40,13
(0.11s) t=4.002736 : 376,83,43
(0.2s) t=5.000259 : 275,129,98
(0.3s) t=6.01724 : 176,150,176
(0.38s) t=7.010942 : 110,139,253
(0.44s) t=8.027782 : 78,108,316
(0.47s) t=9.021583 : 57,87,358
(0.5s) t=10.02733 : 48,56,398
(0.52s) t=11.01434 : 41,42,419
(0.61s) t=12.00043 : 37,29,436
(0.63s) t=13.01873 : 34,26,442
(0.64s) t=14.12750 : 31,14,457
(0.64s) t=15.04642 : 31,9,462
(0.66s) t=16.30211 : 31,4,467
(0.66s) t=17.13028 : 30,4,468
(0.66s) t=18.26802 : 30,1,471
t=20.45879 : 30,0,472
tf: 20.45879
TerminationStatus: zeroProp
Duration: 0.67 seconds
Method: D
Nr of steps: 942
Mean step size: 0.02171846+/-0.08930695
End wall time: 2009-12-26 17:58:14
--------------------
ssa.plot(out)
```

deterministico. Il corrispondente numero di suscettibili in base al quale il numero di infetti comincia a diminuire è pari a 176 secondo il modello stocastico, contro i 167 del modello matematico.
La principale differenza tra il modello stocastico e quello deterministico è che nel primo caso vengono considerate le fluttuazioni casuali della popolazione traducendo così in termini di probabilità la variazione di ciascun stato in cui versa l'individuo, per il tempo successivo. Il modello stocastico è dunque una rappresentazione probabilistica del modello S.I.R. deterministico di Kermack e McKendrick, e come tale può essere utilmente utilizzato per convalidare le previsioni del modello deterministico.

2.15 Il modello di Reed e Frost.

Concludiamo questo capitolo sull'uso dei modelli per descrivere l'andamento nel tempo delle epidemie, descrivendo brevemente uno dei modelli base più conosciuti proposto da Reed Frost nel 1920 diventato famoso come il modello epidemico di Reed - Frost.
Si tratta di un modello probabilistico che oggi assume ormai un'importanza soltanto storica, via via che vengono messi a punto modelli più sofisticati, ma ha il pregio di essere facilmente comprensibile. Il modello di Reed e Frost è stato soggetto ad ampie revisioni e modifiche. Nella sua forma originaria più semplice, esso prevede una suddivisione della popolazione in 3 gruppi:

1. soggetti infetti (i casi);
2. recettivi,
3. gli immuni.

È previsto che ogni individuo infettato vada incontro a malattia (cioè diventi un caso) e quindi guarisca e diventi immune. Perciò questo modello rientra fra i modelli detti "SIR" (*Susceptible, Infected, Resistant*).
Le ipotesi di base del modello sono:

- il tempo è discreto;
- i soggetti hanno la stessa probabilità di infettarsi;
- l'infezione dura una unità di tempo;
- all'inizio dell'epidemia possono esserci soggetti rimossi;
- l'infezione si propaga per contatto;
- il contatto tra un rimosso e un infetto non causa l'infezione.

Secondo il modello, l'andamento della curva epidemica ed il pattern di immunità nella popolazione dipendono dal numero complessivo di individui che costituiscono la popolazione, dal loro stato (contagianti, recettivi ed immuni) e dalla facilità con cui la malattia si trasmette da un individuo all'altro. Supponendo che il periodo di contagiosità[34] degli infetti sia breve, e supponendo costante il periodo di latenza e di incubazione, allora, partendo con un caso singolo (o con più casi con infezione contemporanea), i nuovi casi si svilupperanno in una serie di stadi. In ogni stadio, ciascun caso di malattia può essere stimato mediante una distribuzione binomiale, in funzione del numero di individui ricettivi ed infetti nel precedente stadio. Si può dunque stimare una catena di una distribuzione binomiale attraverso il modello della catena binomiale. Il modello assume che tutti gli individui che sviluppano la malattia, si infettino al prossimo stadio e successivamente diventino immuni. Il comportamento del modello è dunque determinato dal numero di soggetti infetti presento al tempo t = 0 e poi dalla probabilità di transazione dallo stato suscettibile allo stato infetto.
Il modello si basa sulla seguente formula:

$$C_{t+1} = S_t(1-q^{Ct}),$$

dove *t* è l'unità di tempo: usualmente corrisponde al periodo di incubazione o periodo latente dell'agente patogeno;
C_{t+1} è il numero di casi infetti nel periodo $t + 1$;
S_t è il numero di recettivi al periodo *t*;
q è la probabilità di un individuo di non avere nessun contatto effettivo.
Il valore q è pari a (1-p) dove p è la probabilità di un individuo di venire effettivamente a contatto con uno infetto se questo fosse ricettivo e gli altri fossero infetti. Così la quantità $(1-q^{Ct})$ aumenta in quanto essa rappresenta la probabilità che almeno uno dei C_t casi di ammalati venga effettivamente a contatto. La grandezza di p dipende da vari fattori ed è stimata empiricamente dalle reali epidemie.
Se al tempo *t* (inizio dell'epidemia) ci sono ad esempio 100 individui recettivi, nessuno immune ed un solo caso di malattia, allora $S_t = 100$ e $C_t = 1$. Supponendo che $p = 0.06$, allora $q = (1 - 0.06) = 0.94$. Al tempo $t + 1$ avremo:

$$C_{t+1} = 100\ (1 - 0.94^1) = 6$$

e

[34] La contagiosità è la propensione di una malattia o di un agente a diffondere all'interno di una popolazione recettiva attraverso contatto (diretto o indiretto) tra soggetti infetti.

$$S_{t+1} = 100 - 6 = 94.$$

Al tempo $t + 2$

$$C_{t+2} = 94\ (1 - 0.94^{6}) = 29 \text{ e } S_{t+2} = 94 - 29 = 65.$$

Al tempo $t + 3$

$$C_{t+3} = 65\ (1 - 0.94^{29}) = 54 \text{ e } S_{t+3} = 65 - 54 = 11.$$

E così via.

Il totale degli individui immuni in ciascun periodo è pari alla somma cumulata degli individui infetti durante tutti i periodi precedenti. Così al periodo $t + 1$ il numero degli immuni è $I_{t+1} = 1$ (il primo caso al tempo t = 0). Al tempo $t + 2$, $I_{t+2} = 6 + 1 = 7$, al tempo $t + 3$, $I_{t+3} = 7 + 29 = 36$ e così via.

Nella Tabella 2.1 vengono illustrati i risultati del modello Reed-Frost, usando i parametri dell'esempio. I risultati vengono anche rappresentati graficamente nella Figura 2.7. Si osserva che un'epidemia avviene solo quando $p \times S_t > 1$ e declina (ovvero non può iniziare) quando $p \times S_t < 1$. La possibilità del verificarsi di evento epidemico, e la pendenza di una curva epidemica, sono tuttavia funzione della probabilità di un contatto effettivo e del numero di individui recettivi.

tempo	casi Ct	recettivi St	Immuni It	Totali	p	pS
0	1	100	0	101	0.06	6.00
1	6	94	1	101	0.06	5.64
2	29	65	7	101	0.06	3.90
3	54	11	36	101	0.06	0.66
4	11	0	90	101	0.06	0.00
5	0	0	101	101	0.06	0.00

Tabella 2.1: *Simulazione di una epidemia usando il modello di Reed-Frost.*

Dall'esame del modello si evince che la probabilità che si verifichi un'epidemia e l'aspetto della curva epidemica sono funzioni del contatto efficiente e del numero di soggetti recettivi. La proporzione di una popolazione che risulta recettiva viene usata spesso come guida generale della probabilità di diffusione di un'infezione. Si ritiene che - in genere e con larga approssimazione - almeno il 20-30% della popolazione debba essere recettiva perché abbia luogo una epidemia. Ne consegue che l'infezione non diffonderà se il 70-80% della popolazione è immune. Questo è utile a prevenire epidemie di grandi dimensioni, tuttavia, è da notare che l'infezione potrà diffondere ugualmente, anche

in presenza di una elevata immunità di popolazione, se il numero degli individui recettivi è tale che (p * St) ›1

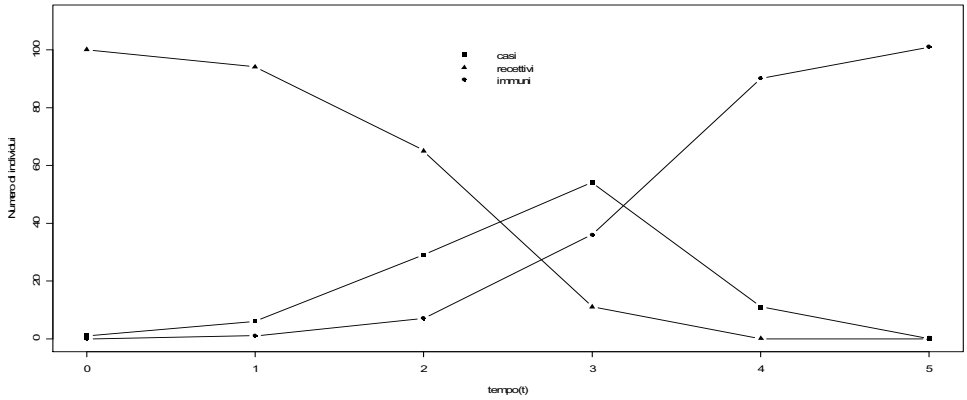

Fig. 2.10: *esempio di curva epidemica, numero di individui recettivi e numero di individui immuni secondo il modello di Reed-Frost*

Il grafico della Figura 2.10[35] descrive diverse curve epidemiche simulate usando diversi valori di individui immuni, recettivi e parametri *p*. Si vede che allo stato iniziale di una epidemia, un certo numero di individui immuni può diminuire l'estensione dell'epidemia e ritardarne il suo picco così come un cambiamento nell'effettivo contatto può altresì alterarne l'ampiezza. *(devo creare la figura con R)* Il modello base di Reed-Frost può essere anche modificato aggiungendo alcune variabili di controllo come ad esempio i vaccini con periodo variabile di immunità e variabile periodo di infezione.

[35] Codice R:

```
tempo=c(1,2,3,4,5,6)
casi=c(1,6,29,54,11,0)
recettivi=c(100,94,65,11,0,0)
immuni=c(0,1,7,36,90,101)
totali=c(101,101,101,101,101,101)
p=c(0.6,0.6,0.6,0.6,0.6,0.06)
pS=c(6.00,5.64,3.90,0.66,0.00,0.00)
plot(tempo,casi,"b",pch=15,ylim=c(0, 111), ylab="Numero di individui",xlab="tempo(t)")
lines(tempo,recettivi,"b",pch=17)
lines(tempo,immuni,"b",pch=19)
legend(locator(1),legend=c("casi","recettivi","immuni"),pch=c(15,17,19),bty="n")
```

Capitolo 3:
analisi statistica di alcune malattie infettive.

3.1 Introduzione.

Prima di procedere alla descrizione dell'algoritmo di Farrington e la sua implementazione in R, si analizzerà brevemente alcune serie temporali relative a diverse malattie infettive. Le serie storiche sono state ricavate dai dati raccolti una struttura di livello nazionale competente per il controllo e la prevenzione delle malattie, nonché responsabile del sistema di sorveglianza nazionale delle malattie infettive[36]. In particolare, l'istituto gestisce un database di casi di denuncia delle malattie attuando il monitoraggio in accordo con l'*Infectous Diseases Act* del 2001, che obbliga i medici generici ed i laboratori a notificare alle autorità sanitarie locali i casi di particolari malattie infettive. Dopo aver verificato la segnalazione, i dati sono resi anonimi ed elettronicamente trasmessi prima alle autorità sanitarie statali e poi all' Istituto Robert Koch. Il monitoraggio è basato su SurvNet@RKI, un database elettronico integrato, amministrato statalmente e sviluppato dall'RKI. I dati sono resi pubblici sul sito internet *http://www3.rki.de/survstat*, e aggiornati settimanalmente.

SurvStat@RKI è un'applicazione che consente il recupero di questi dati e la creazione di tabelle, grafici e mappe in base alle esigenze dell'utente.

I dati accessibili sono sincronizzati con il Bollettino Epidemiologico che rappresenta il report ufficiale dei dati settimanalmente trasmessi. L'interfaccia che si presenta all'utente, consente di ottenere i dati attraverso l'interrogazione combinata delle seguenti variabili:

- la malattia infettiva;
- l'individuazione del caso, ossia il tipo di malattia infettiva in base all'agente infettante;
- la zona geografica: scelta fra stato, regione e distretto;
- la frequenza della rilevazione: settimana, mese, trimestre e anno;
- gruppo di età (vengono offerte diverse stratificazioni);

[36] www.rki.de

- sesso;
- agenti patogeni (ad esempio, sierotipo, tipo di fago, ecc).

Una volta scelto i criteri per eseguire le analisi nei confronti di un determinato sottoinsieme di dati, è possibile attuare la selezione attraverso una query. Il raggruppamento può essere attuato solo per una singola variabile anche se nell'ultimo pannello *Display*, può essere selezionata una seconda variabile eseguendo così una query a campi incrociati. È inoltre possibile selezionare il tipo di risultato che si vuole ottenere; se il numero assoluto dei casi o i tassi di incidenza ogni 100.000.
Per le analisi qui eseguite, sono stati estrapolati i dai relativi alle seguenti malattie infettive:

- Rotavirus
- Salmonellosi
- Epatite A
- Epatite B
- Legionellosi
- Tifo
- E. Coli Enterite
- Influenza B

Di ciascuna si sono considerati i dati mensili globali per gli anni che vanno dal 2001 ai primi mesi del 2010. La maggior parte delle serie storiche è caratterizzata da una componente stagionale e da assenza di trend. Le analisi sono state eseguite utilizzando la libreria *ast* di R, disponibile nel sito: *http://sirio.stat.unipd.it.*
Le tecniche di analisi effettuate, utilizzano modelli dinamici stocastici basati sul lisciamento esponenziale. Si tratta di modelli utilizzati principalmente per le previsioni nel breve periodo, che cercano di descrivere la legge con cui un certo processo stocastico si evolve nel tempo. Il lisciamento esponenziale consiste nell'applicazione alla serie dei dati, di una media mobile *ponderata esponenzialmente*. In questo modo ciascun valore della serie lisciata dipende da tutti i valori osservati precedenti. Inoltre, nel calcolo dei valori della serie livellata, i pesi assegnati a ciascun valore osservato in precedenza non sono costanti, ma decrescono passando dai più recenti a quelli più lontani nel tempo.

Così, ad esempio, nel calcolo del livellamento esponenziale per il periodo *i* verrà assegnato il peso maggiore al valore osservato nel periodo *i* – 1, un peso inferiore al valore osservato nel periodo *i* – 2, e pesi via via decrescenti fino ad arrivare al primo valore osservato della serie, al quale è assegnato il peso minore. I pesi sono definiti dalla *successione esponenziale*:

$$c_j = \alpha(1 - \alpha)^{j-1} \; ; \qquad j \geq 1$$

Le formule per il livellamento esponenziale di una serie storica si basano su tre soli termini: il valore corrente della serie Y_i, il valore della serie smussata calcolato per il periodo precedente, l_{t-1}, e un peso, o fattore di lisciamento, assegnato soggettivamente, α. Per ogni periodo *i* si ha quindi la seguente formula per la determinazione della serie smussata:

$$y_t = l_{t-1} + u_t,$$
$$l\text{t} = (1 - \alpha)\, l_{t-1} + \alpha\, y_t \,,$$

l_{t-1} è il valore della serie lisciata esponenzialmente relativo al periodo *i* – 1 (il *livello* delle osservazioni al tempo t è già noto al tempo t – 1), mentre u_t costituisce la deviazione del valore corrente dal livello determinato precedentemente. Quindi l_t è una media pesata di $y_t, \ldots, y_1, l_0$, dove la somma dei pesi vale 1. Se $0 < \alpha < 1$, i pesi assegnati alle osservazioni passate decrescono geometricamente; sono quindi posti su una curva di tipo esponenziale, da cui il nome *lisciamento esponenziale*. Le $\{u_t\}$ sono una successione di variabili casuali indipendenti ed identicamente distribuite di media nulla e varianza σ^2.
In tutte le analisi qui condotte, alcuni concetti saranno dati per scontati, rinviando per questo al materiale didattico che si trova nel sopra citato sito.

3.2 Prime analisi esplorative.

Prima di procedere all'analisi di ogni singola serie storica estrapolata, si visualizza nella Figura 3.1 il grafico dei dati raccolti relativi alle malattie infettive che verranno poi studiate nei successivi paragrafi:

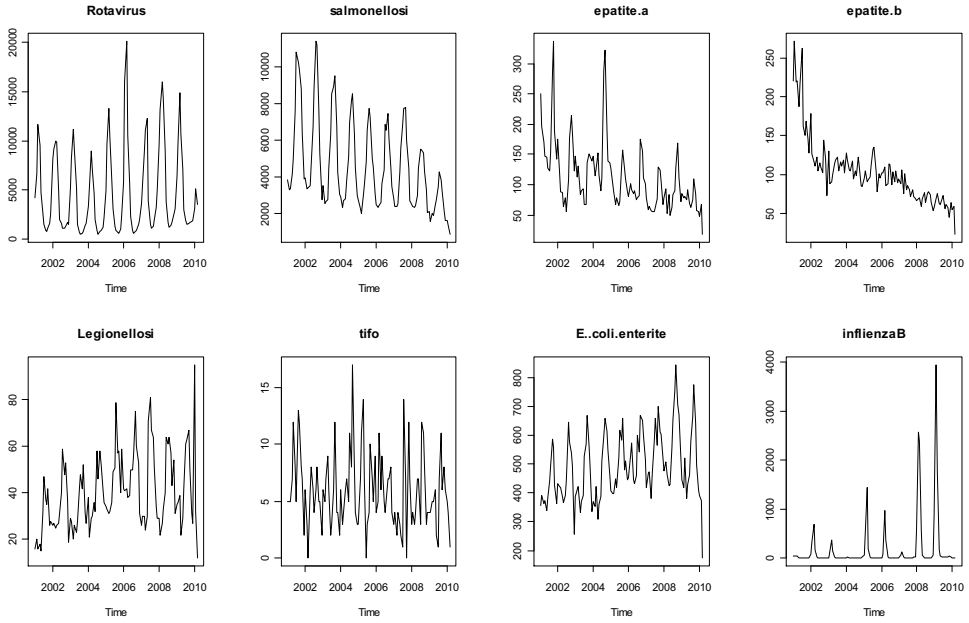

Fig. 3.1: *serie storiche relative ad alcune malattie infettive*.

Da una prima analisi esplorativa dei dati, si nota che la maggior parte delle serie storiche considerate presenta una componente ciclica di tipo *stagionale*; ossia delle fluttuazioni periodiche più o meno regolari che si ripetono annualmente, mentre quasi tutte sono prive di *trend* o questo si presenta in maniera molto debole. Il *trend* è quella componente sistematica di lungo termine della serie storica che si manifesta in un aumento o diminuzione dei valori della serie stessa.

Ciascuna serie è stata scomposta nella componente di *trend*, là dove ci sia, e nella componente *stagionale*, adottando per esse il modello a lisciamento esponenziale che meglio si adatta alla serie sottoposta ad analisi, ottenendo così, in via residuale rispetto alle due componenti di trend e stagionalità, la *componente irregolare* o *casuale*, ossia quella parte di informazioni che il modello adattato non riesce a cogliere. Se il modello si adatta bene ai dati, allora la componente irregolare si presenta con assenza di autocorrelazione dei residui, ovvero come un processo white-noise.

Le analisi partono dal presupposto che si tratti di processi *stazionari* nel tempo tali per cui, per ogni h (il numero di ritardi considerati nella serie), t e $t + h$ vale:

$$E(Y_t) = \eta, \tag{3.1}$$

$$var(Y_t) = \sigma^2, \tag{3.2}$$

$$cov(Y_{t+h}, Y_t) = \gamma(h), \tag{3.3}$$

$$corr(Y_{t+h}, Y_t) = \rho(h), \tag{3.4}$$

ovvero, se un processo è stazionario, la media e la varianza non variano con il tempo, mentre la covarianza, e quindi la *funzione di autocorrelazione parziale*, ossia la funzione che descrive la dipendenza temporale lineare della serie, è solo funzione della distanza nel tempo tra le due variabili casuali coinvolte.

3.3 Rotavirus.

La gastroenterite da Rotavirus è un virus intestinale diffuso in tutto il mondo che colpisce principalmente i neonati. In Europa e nel resto delle zone temperate del pianeta, il virus si presenta con picchi di incidenza stagionale che, alle nostre latitudini, si verificano nel periodo invernale tra novembre e marzo. Nei Paesi tropicali si possono verificare picchi di incidenza, ma il virus è presente sostanzialmente tutto l'anno. Nei Paesi occidentali, la gastroenterite da rotavirus non è una malattia letale, ma può dare complicanze anche molto gravi nelle persone anziane e in quelle immuno-compromesse. Nei Paesi del Sud del mondo, al contrario, causa la morte di almeno 600 mila bambini ogni anno per diarrea, secondo le stime dell'Organizzazione mondiale della sanità che considera la malattia una vera e propria emergenza sanitaria[37].
I dati sono mensili e vanno da gennaio 2001 a marzo 2010:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001	4204	6650	11642	9575	4715	2568	1489	851	751	1131	1551	2522
2002	8147	9170	10009	9846	4978	2033	1562	1071	1079	1303	1697	1480
2003	5175	8285	11196	9225	5246	1481	679	474	545	805	1399	1629
2004	3326	5086	8998	7257	4590	2038	916	476	805	920	1178	2221
2005	4907	9641	13251	10452	6107	3092	1265	902	702	586	1033	2361
2006	5782	16022	20124	10518	5637	2316	927	573	819	875	1449	1985
2007	3618	5446	8990	11069	12262	3959	1491	1079	1314	1968	3255	4939
2008	9696	13180	16037	14256	9002	3761	1715	1235	1372	1665	2487	3102
2009	6435	9794	14904	10640	7184	2965	1894	1503	1582	1647	1802	1858
2010	3101	5099	3545									

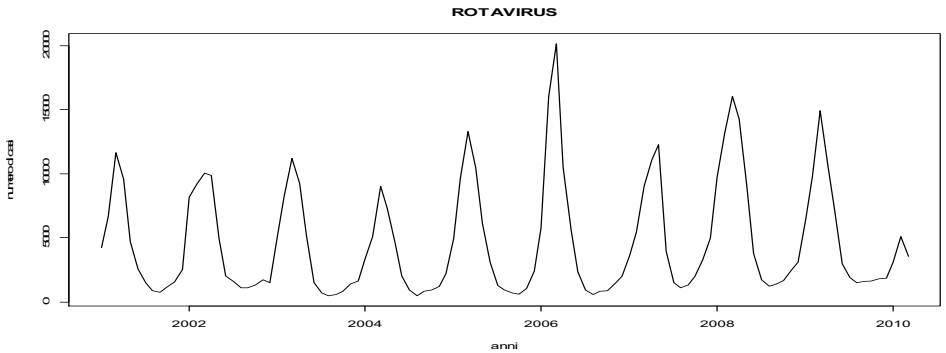

Fig. 3.2: *serie storica relativa all'infezione da Rotavirus*.

[37] http://www.epicentro.iss.it/problemi/rotavirus/rotavirus.asp.

Da una prima analisi grafica, la serie storica presenta una forte componente stagionale ed un trend molto debole se non inesistente (si veda la Figura 3.2). Visualizziamo la componente stagionale della serie considerata in Figura 3.3:

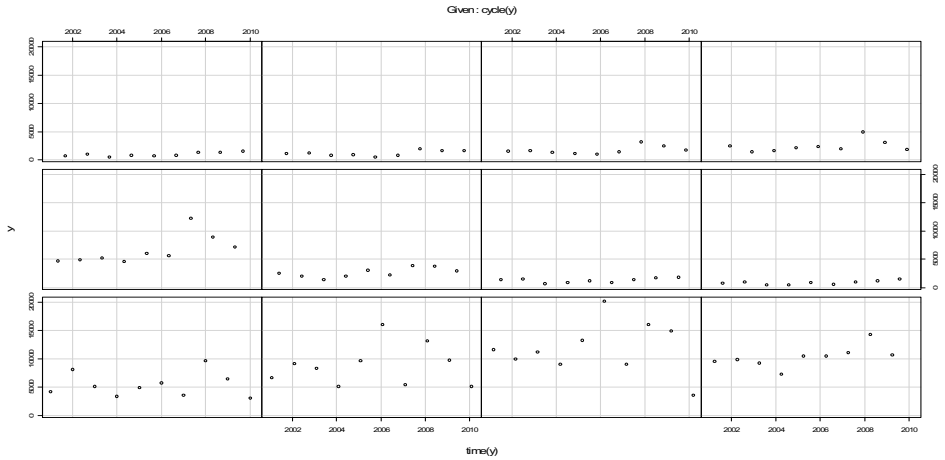

Fig. 3.3: *componente stagionale* della *serie storica relativa all'infezione da Rotavirus.*

La Fig. 3.3 mostra le 12 sottoserie mensili (tutte le osservazioni di gennaio, febbraio ecc.) utilizzando un campo di variazione comune per l'asse delle ordinate. Il grafico in basso a sinistra mostra i valori osservati nei vari mesi di gennaio, quello alla sua destra i valori osservati nei vari mesi di febbraio; e così via; l'ordinamento è da sinistra verso destra e dal basso in alto (ovvero il grafico sulla seconda riga, terza colonna riporta i valori osservati nei vari anni durante il mese di luglio). Si osserva in particolare che, durante tutti gli anni considerati, il virus colpisce soprattutto nei mesi invernali a partire dal mese di gennaio, per raggiungere il picco nel mese di marzo, solo per l'anno 2007 il massimo numero di casi registrati è avvenuto nel mese di maggio per un totale di 12.262 di infetti.

La Figura 3.4 evidenzia in maniera diversa la componente stagionale della serie storica: il primo mostra le sottoserie mensili per ciascun anno considerato: in particolare ogni curva mostra l'andamento del *Rotavirus* in un particolare mese (ovvero una curva mostra i valori osservati di gennaio durante tutti gli anni di osservazione, un'altra quelli osservati a febbraio e così via). Il secondo grafico invece, rappresenta il *profilo stagionale* della serie ed è stato ottenuto sottraendo ad ogni osservazione, la media delle misurazioni dello stesso anno: quindi, ad esempio, all'osservazione di aprile 2001, è stata sottratta la media delle osservazioni di gennaio 2001, febbraio 2001, ... dicembre 2001. Le

singole curve mostrano così l'andamento delle singole osservazioni ricentrate per i vari anni. Ad esempio una curva è riferita al 2003, e mostra le osservazioni precedentemente ricentrate riferite al gennaio 2003, febbraio 2003,....dicembre 2003.

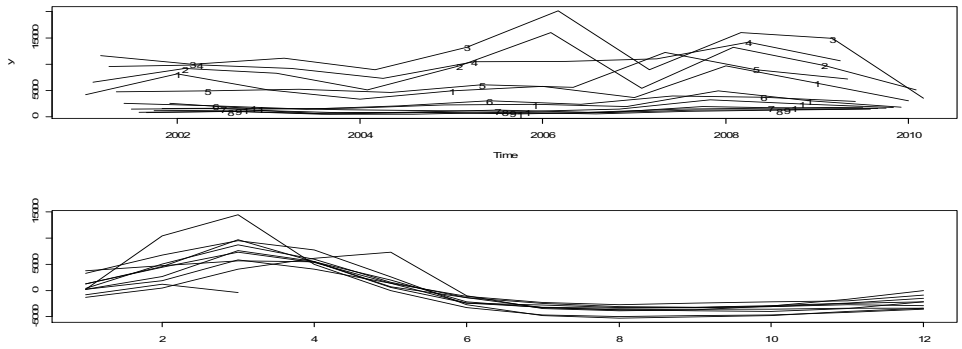

Fig. 3.4: *sottoserie mensili e profilo stagionale della serie storica relativa all'infezione da Rotavirus.*

I due grafici della Figura 3.4, evidenziano che il *profilo stagionale* non è rimasto molto costante nei nove anni considerati: se così fosse, le linee del primo grafico della Figura 3.4, sarebbero parallele, cosa che non accade specie per i mesi che vanno da gennaio a maggio. Nel grafico in basso si osserva quanto detto sopra: il numero degli infetti comincia crescere a partire dal mese di gennaio, raggiunge il suo picco nel mese di marzo, per poi diminuire e raggiungere il minimo nei mesi estivi, tuttavia l'andamento non è sempre uguale per tutti gli anni. Inoltre si nota una forte variabilità dei picchi raggiunti lungo i diversi anni.

La *funzione di autocorrelazione*, definita nel paragrafo precedente, evidenzia il grado di dipendenza tra le osservazioni ai vari ritardi. Si è detto che questa è solamente funzione della distanza nel tempo tra le due variabili casuali coinvolte, e per h=0 si ha $cov(Y_{t+h},Y_t) = var(Y_t) = \sigma^2$

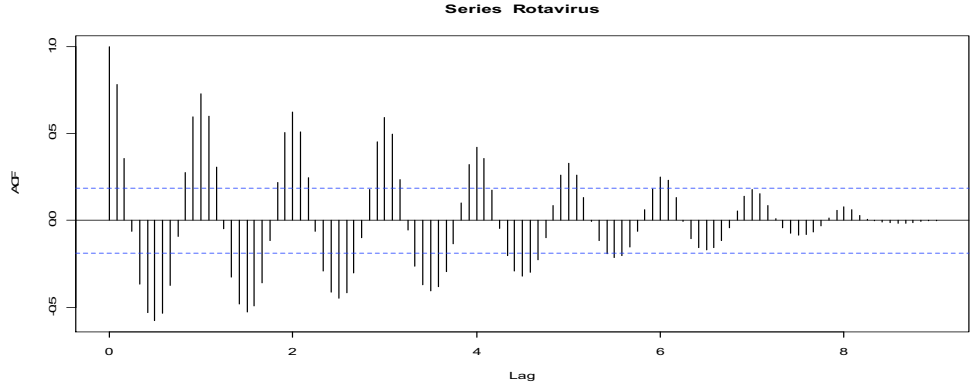

Fig. 3.5: *Infezione da Rotavirus: stima della funzione di autocorrelazione. La funzione stimata etichetta l'asse delle ascisse utilizzando il tempo espresso in anni.*

Il correlogramma, mostrato in Figura 3.5, è quello tipico in casi di serie con una forte componente stagionale e poco trend, ovvero mostra un andamento "sinusoidale" che si smorza lentamente. L'andamento sinusoidale può essere spiegato osservando che, ad esempio, osservazioni dello stesso mese in anni differenti tendono a stare dalla stessa parte rispetto alla media di tutte le osservazioni e quindi che quasi tutti gli addendi che entrano nel calcolo della autocovarianza ai vari ritardi hanno un "segno prevalente" facilmente determinabile e anche prevedibile.

Questo andamento tipico è in buona parte collegato al fatto che lo stimatore usuale della funzione di autocovarianza si basa sull'assunzione di stazionarietà e quindi ipotizza che la media del processo in ogni istante di tempo sia la stessa. Viceversa, in questo caso, è evidente che la media della serie varia al variare del mese[38]. Il processo non è stazionario e questo è legato soprattutto all'ampiezza delle oscillazioni stagionali.

Il primo coefficiente disegnato al ritardo zero vale 1 perchè rappresenta la correlazione di Y_t con se stesso, il secondo coefficiente di correlazione è disegnato in corrispondenza di una ascissa uguale ad 7; infatti rappresenta la correlazione tra due osservazioni consecutive ma che sappiamo essere distanti di un mese. Si osservi come le onde nel periodogramma "si smorzino" lentamente. A sei anni di distanza c'è ancora della dipendenza.

Dopo una prima analisi esplorativa, si è passato a stimare il modello a lisciamento esponenziale più adatto alla serie, utilizzando il comando *esId*(·) della libreria *ast* di R. La funzione *esId*(·) stima, utilizzando il metodo della massima verosimiglianza, tutti i modelli basati sul lisciamento esponenziale compatibili con la serie osservata (in questo caso tutti i modelli stagionali) e mostra i modelli ordinati utilizzando il criterio BIC. I modelli proposti sono i seguenti:

```
   drift sea inn np nlog.lik      BIC      AIC rankAIC
1    c/m c/m   m 15 1973.934 2044.577 2003.934       1
2      n c/m   m 14 1984.320 2050.253 2012.320       2
3      a   a   m 17 1997.865 2077.927 2031.865       3
4      n c/a   m 14 2030.738 2096.672 2058.738       4
5      n   m   m 15 2051.338 2121.981 2081.338       5
6      d   m   m 18 2054.939 2139.711 2090.939       6
7      d c/m   m 17 2061.382 2141.444 2095.382       7
```

[38] Un modo per rendere stazionario il processo e quello di attuare una trasformazione logaritmica, e se questo non basta, considerando il logaritmo della serie differenziata una o più volte, fino ad ottenere una stazionarietà accettabile. Per approfondimenti si rinvia a: Masarotto G., "*Diario delle lezioni in laboratorio informatico di "analisi delle serie temporali*", Facoltà di Scienze Statistiche Università di Padova.

```
8    c/m c/a   m 15 2087.590 2158.233 2117.590       8
9      d   a   m 18 2110.279 2195.050 2146.279       9
10   c/a c/m   a 15 2126.748 2197.391 2156.748      11
11     a c/m   a 16 2123.516 2198.868 2155.516      10
12     n c/m   a 14 2133.485 2199.419 2161.485      13
13   c/m c/m   a 15 2128.963 2199.606 2158.963      12
```

I differenti modelli sono identificati nell'output con le "iniziali" del tipo di *deriva* (*drift*), stagionalità (*sea*) e innovazione (*inn*). L'output del comando mostra anche il numero di parametri del modello (*np*), la log-verosimiglianza (cambiata di segno), i valori di BIC e AIC e il numero d'ordine del modello se si utilizzasse AIC per ordinare i vari modelli (rankAIC).

In questo caso, i due criteri sono in accordo per i primi nove modelli, e suggeriscono un modello con deriva e stagionalità costante moltiplicativa, e una innovazione di tipo moltiplicativo.

La deriva tiene conto del trend della serie storica e modifica i valori lisciati secondo grandezze variabili nel tempo; le variazioni temporali della deriva evidenziano i cambiamenti del trend che possono essere di diverso tipo. Una deriva costante/moltiplicativa, sta a significare che questa aumenta in modo esponenziale ma ad un tasso di crescita costante, mentre la stagionalità costante/moltiplicativa significa che la stagionalità segue il trend secondo un modello moltiplicativo[39], anche se i coefficienti stagionali stimati dal modello non cambiano nel tempo.

L'ultima componente del modello riguarda l'i*nnovazione*, ovvero tutta quella parte non predicibile dal modello e che viene indicata con la variabile casuale u_t. In generale un modello di lisciamento esponenziale può essere scritto in questo modo:

$$y_t = g_t + u_t,$$

dove g_t è la parte di y_t predicibile sulla base del passato. Le u_t (l'innovazione) invece, sono appunto quella parte residuale non prevista dal modello. Per alcune serie temporali la varianza di u_t, ovvero dell'innovazione, non sembra dipendere dal livello della serie (ovvero da g_t). Supponendo che la varianza sia anche costante nel tempo possiamo allora scrivere

$$\text{var}(u_t) = \sigma^2,$$

[39] Le componenti di una serie storica possono essere legate tra loro in modo additivo:

$$Y_t = T_t + S_t + u_t$$

oppure in modo moltiplicativo:

$$Y_t = T_t * S_t * u_t.$$

dove σ^2 è una costante appropriata. Si parla, in questi casi, di innovazione *addittiva*.

Il primo modello del comando *esId*(·), invece, propone un'innovazione *moltiplicativa*: si ha quando la variabilità di u_t sembra dipendere da g_t ed in particolare, lo scarto quadratico medio di u_t sembra essere proporzionale a g_t. E' quindi usuale considerare anche la possibilità che:

$$\text{var}(u_t) = \sigma^2 g_t^2.$$

In questi casi, si parla appunto di innovazione *moltiplicativa*, ossia la dimensione dell'errore è proporzionale al livello della serie.

Detto questo, si è stimato il primo modello proposto (i due criteri sono concordi), ossia quel modello che tra tutti i parametri possibili, assegna la massima probabilità ai dati osservati.

Il modello adattatato ha dato le stime seguenti:

```
Call: esFit(y = rot, drift = "c/m", seasonality = "c/m", innovation = "m")
drift=c/multiplicative, seasonality=c/multiplicative,
innovation=multiplicative
       alpha      l.start      d.start         s[1]         s[2]         s[3]
   1.0000000 3476.0559694    1.0659692    1.1252763    1.8625429    2.5120617
        s[4]         s[5]         s[6]         s[7]         s[8]         s[9]
   2.0725496    1.3418631    0.5415277    0.2713481    0.1735789    0.1959804
       s[10]        s[11]        s[12]        sigma
   0.2318775    0.3296002    0.4855710    0.2170993
-2log(likelihood)= 1973.901  AIC= 2003.901   BIC= 2044.544
```

La funzione di stima restituisce le stime dei parametri: in particolare le costanti di lisciamento non vincolate, ovvero i vari α e β, le condizioni iniziali per le equazioni alle differenze che definiscono il modello (ovvero,α, l_0 , d_0 e s_0, s_{-1}, . . .) e il parametro di dispersione dell'innovazione σ. La costante di lisciamento α è pari a 1. Il valore di α determina di quanto l'osservazione corrente influenza il valore futuro. Quanto più prossimo a zero è il valore di α, tanto meno il valore corrente influenza la previsione (cioè la nuova previsione sarà molto simile alla vecchia), cioè per α tendente a uno la nuova previsione sarà molto vicina all'ultimo valore della serie, e viceversa. In questo caso α è esattamente pari a 1: quindi l'informazione dell'ultima osservazione influenza moltissimo la successiva.

Si è verificata la capacità del modello di cogliere l'autocorrelazione dei residui: per questo si è utilizzato il comando *tsdiag*(·) che, applicato all'oggetto calcolato da *esFit*(·), visualizza tre grafici: quello dell'innovazione, ossia il grafico dei residui standardizzati, il grafico della loro autocorrelazione campionaria, e dei livelli di significatività

osservati del test di Ljung-Box[40] calcolato sul primo coefficiente di autocorrelazione, sui primi due, e così via. Il grafico è mostrato nella Figura 3.6:

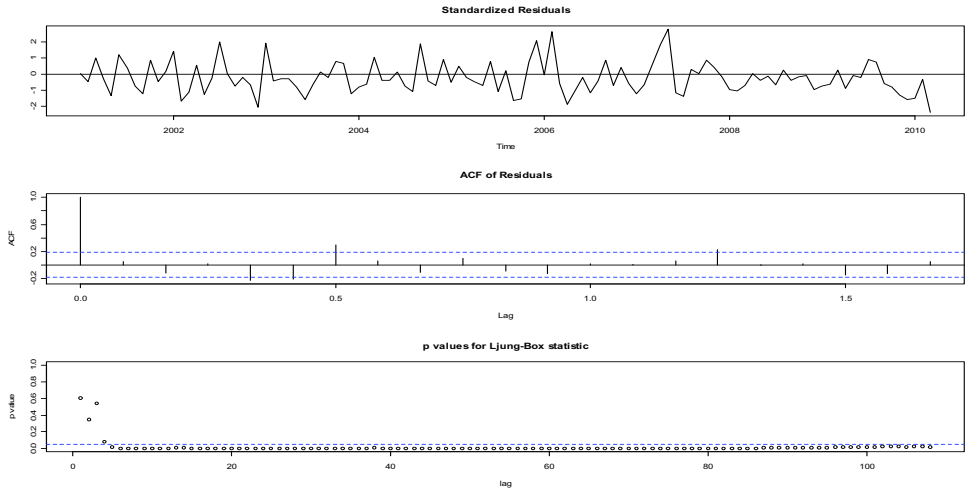

Fig. 3.6: *Alcuni grafici diagnostici* (*deriva e stagionalità costante/moltiplicativa, innovazione moltiplicativa)*.

La linea orizzontale tratteggiata nel terzo pannello (quello riferito al test di Ljung-Box) è disegnata ad una altezza pari a 0,05. Valori sotto la linea del *p-value* indicano quindi un risultato significativo al 5%.

Dagli ultimi due grafici della figura 3.6, si osserva che il modello proposto non si adatta bene alla serie storica osservata in quanto rimane ancora della correlazione seriale tra i residui. In particolare, l'ultimo grafico evidenzia i coefficienti relativi al test di Ljung-Box, e si vede che, tranne i primi tre coefficienti, tutti gli altri stanno al di sotto della barra del 5%. Non siamo dunque di fronte ad un processo white-noise. Se il modello fosse buono, tutti i coefficienti del terzo grafico dovrebbero stare sopra la linea del 5%[41].

Anche adattando i modelli successivi proposti dal comando *esFit*(·), le cose non cambiano di molto; rimane sempre della correlazione seriale tra i residui. Molto probabilmente questo dipende dal fatto che la serie storica del Rotavirus non è stazionaria. Un modo per ottenere un modello che meglio si adatti alla serie potrebbe essere quello di applicare i modelli ARIMA, passando prima a differenziare la serie tante

[40]Il test di *Ljung-Box* verifica l'ipotesi che le prime *m* autocorrelazioni siano nulle, valutando la significatività "complessiva" dei coefficienti di autocorrelazione ai vari ritardi che vengono scelti a discrezione dell'analista. Per cui se il test di Ljung-Box presenta un p-value inferiore a 0.05, allora non esiste autocorrelazione e siamo in presenza di quello che viene chiamato processo *whyte-noise*, e quindi i livelli di significatività rimangono molto bassi.

[41] Si Veda Masarotto G. (2004) "*Diario delle lezioni in laboratorio informatico di "analisi delle serie temporali*", - Facoltà di Scienze Statistiche Università di Padova.

volte quanto necessario per renderla stazionaria. Tuttavia, non è competenza di questo lavoro esporre i diversi approcci teorici per studiare il comportamento temporale di una serie storica.

Come ultima cosa, per ottenere un quadro informativo finale sulle componenti stimate, si possono evidenziare le componenti della serie Rotavirus associate con il modello stimato; come mostrato in Figura 3.7.

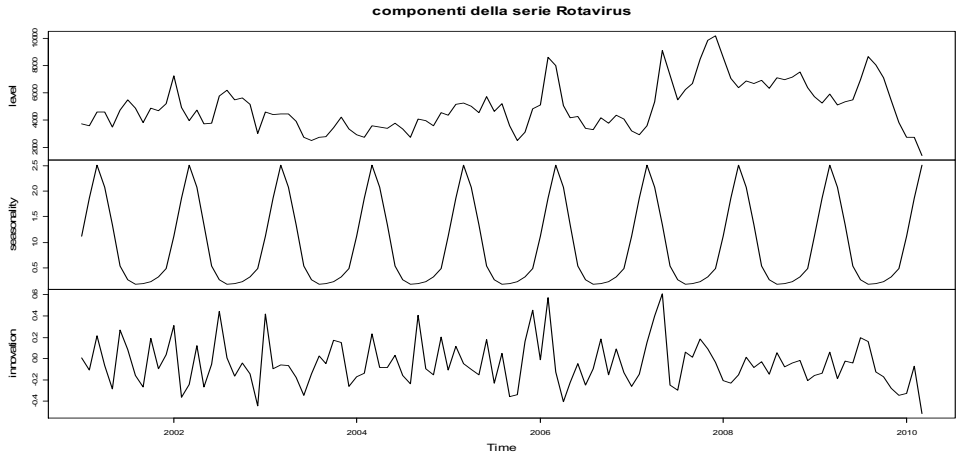

Fig. 3.7: *scomposizione della serie Rotavirus*.

3.4 Salmonellosi.

La Salmonellosi, È una malattia che colpisce soprattutto l'intestino, ed è causata da un battere, la Salmonella, del quale si conoscono differenti gruppi, a loro volta suddivisi in sierotipi. Qui sono stati considerati tutti i gruppi e tutti i sierotipi. La malattia si trasmette attraverso il contatto diretto o indiretto con le feci di persone malate o portatrici, ma soprattutto attraverso il consumo di cibi contaminati. Gli alimenti che più frequentemente agiscono da veicolo dell'infezione sono uova, latte e derivati, carne e pollame e loro derivati. Anche l'uomo può trasmettere la Salmonella, specialmente quando è presente diarrea; il rischio di trasmissione è maggiore in presenza di bambini e adulti con incontinenza fecale.

Si riportano qui di seguito, i dati mensili e la rispettiva rappresentazione grafica in Figura 3.8:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001	3804	3272	3330	4156	4987	7566	10783	10266	9896	8766	6404	3885
2002	3949	3341	3390	3505	4621	6725	8824	11370	11148	7613	5235	2736
2003	3487	2559	2618	2724	4471	6295	8516	8896	9519	6767	4232	3007
2004	2889	2308	2672	2778	3847	5053	6953	8139	8539	6367	4465	2980
2005	2775	2267	2014	2940	3481	4775	6535	7691	7283	4949	4536	3036
2006	2486	2324	2431	2596	3614	4349	6826	6515	7449	6117	4461	3441
2007	2795	2359	2384	2555	4247	5576	6890	7699	7754	5988	4453	2709
2008	2470	2375	2336	2439	2947	4485	5495	5440	5311	4291	3266	2065
2009	2082	1568	2010	1875	2488	2750	3451	4263	3838	3242	2204	1631
2010	1622	1359	855									

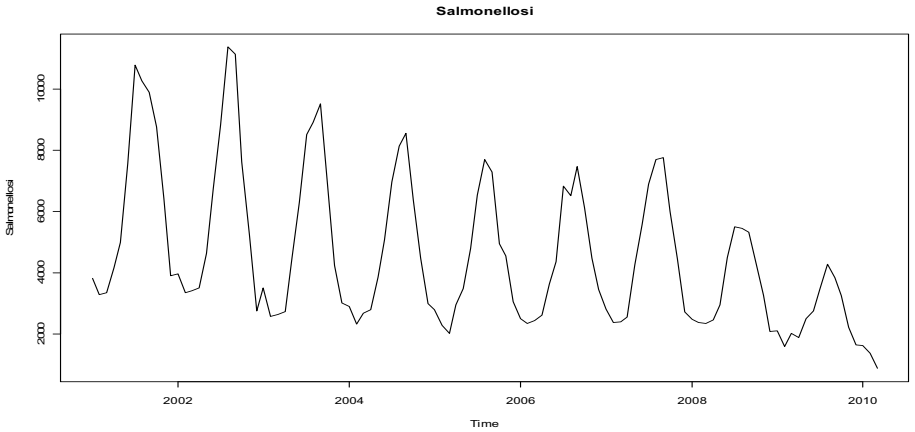

Fig. 3.8: *serie storica relativa all'infezione da Salmonella.*

Il grafico presenta una forte stagionalità che sembrerebbe di tipo moltiplicativo, si nota inoltre un leggero di trend decrescente. Analizziamo più da vicino la componente stagionale:

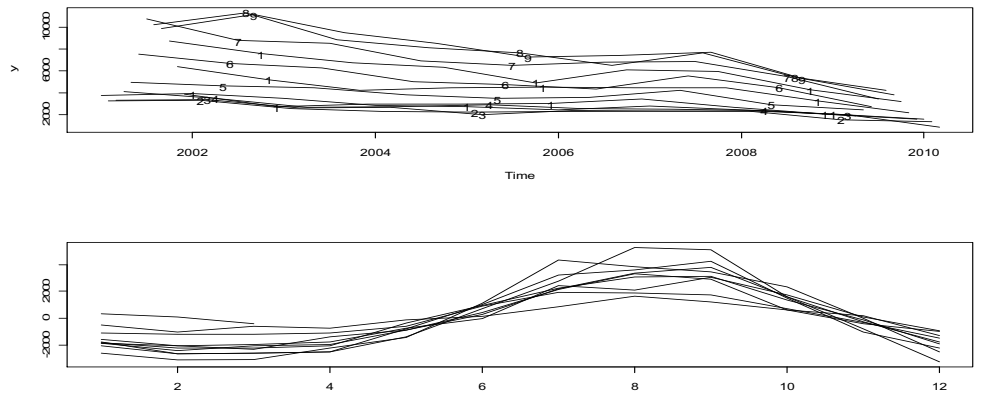

Fig. 3.9: *sottoserie mensili e profilo stagionale della serie storica relativa all'infezione da Salmonella.*

Come si nota dalla Figura 3.9, il virus si manifesta nei mesi estivi con un picco tra il mese di agosto e settembre per poi raggiungere il numero minimo di casi nei mesi invernali tra gennaio e febbraio. Durante tutto il decennio di osservazione, l'andamento del profilo stagionale nei vari mesi, è rimasto più o meno costante, diminuendo gradualmente con il passare degli anni.

Si passi ora a stimare il modello a lisciamento esponenziale che meglio si adatta ai dati osservati. I modelli proposti sono i seguenti:

	drift	sea	inn	np	nlog.lik	BIC	AIC	rankAIC
1	n	c/m	m	14	1847.325	1913.258	1875.325	2
2	c/m	c/m	m	15	1843.095	1913.738	1873.095	1
3	n	m	m	15	1848.773	1919.416	1878.773	4
4	m	c/m	m	16	1844.124	1919.477	1876.124	3
5	c/m	m	m	16	1855.475	1930.828	1887.475	5
6	n	c/m	a	14	1870.678	1936.611	1898.678	8
7	c/a	c/m	a	15	1867.271	1937.914	1897.271	6

```
8     c/m c/m   a 15 1868.598 1939.241 1898.598       7
9       n   m   a 15 1871.775 1942.418 1901.775      11
10      a c/m   a 16 1868.968 1944.321 1900.968       9
11    c/m   m   a 16 1868.973 1944.326 1900.973      10
12    c/a   m   a 16 1872.106 1947.459 1904.106      13
13      a   m   a 17 1869.488 1949.550 1903.488      12
```

Si osserva innanzitutto che i due criteri sono in disaccordo tra di loro: infatti il secondo modello per AIC è il primo per BIC, così come il quarto modello per AIC è i terzo per BIC. Lo stesso vale per tutte le restanti posizioni. Si inizi col stimare il primo modello secondo il criterio BIC, che propone nessuna deriva, una stagionalità costante moltiplicativa ed una innovazione di tipo moltiplicativo.

```
Call: esFit(y = Salmonellosi, drift = "n", seasonality = "c/m", innovation = "m")
drift=none, seasonality=c/multiplicative, innovation=multiplicative
       alpha      l.start         s[1]         s[2]         s[3]         s[4]
   0.4762143 7178.3795653    0.5424476    0.4558623    0.4687964    0.5258551
        s[5]         s[6]         s[7]         s[8]         s[9]        s[10]
   0.7221206    0.9842112    1.3314648    1.4751519    1.4870864    1.1518543
       s[11]        s[12]        sigma
   0.8472895    0.5666369    0.0952938
-2log(likelihood)= 1847.333  AIC= 1875.333   BIC= 1913.267
```

Si passi a stimare anche il secondo modello proposto, che corrisponde al primo secondo il criterio AIC:

```
Call: esFit(y = Salmonellosi, drift = "c/m", seasonality = "c/m", innovation = "m")
drift=c/multiplicative, seasonality=c/multiplicative, innovation=multiplicative
       alpha      l.start      d.start         s[1]         s[2]         s[3]
4.096396e-01 7.321943e+03 9.944754e-01 5.406963e-01 4.550203e-01 4.679129e-01
        s[4]         s[5]         s[6]         s[7]         s[8]         s[9]
5.247265e-01 7.205203e-01 9.826855e-01 1.330180e+00 1.473164e+00 1.486595e+00
       s[10]        s[11]        s[12]        sigma
1.150686e+00 8.466347e-01 5.666369e-01 9.446515e-02
-2log(likelihood)= 1843.186  AIC= 1873.186   BIC= 1913.829
```

Il secondo modello è lo stesso del primo solo che propone una deriva di tipo costante moltiplicativo. Il valore dei parametri rimane più o meno lo stesso per entrambi i modelli, anche se il secondo ha un parametro in meno rispetto al primo modello del criterio AIC.

A questo punto si faccia un confronto grafico tra i due modelli:

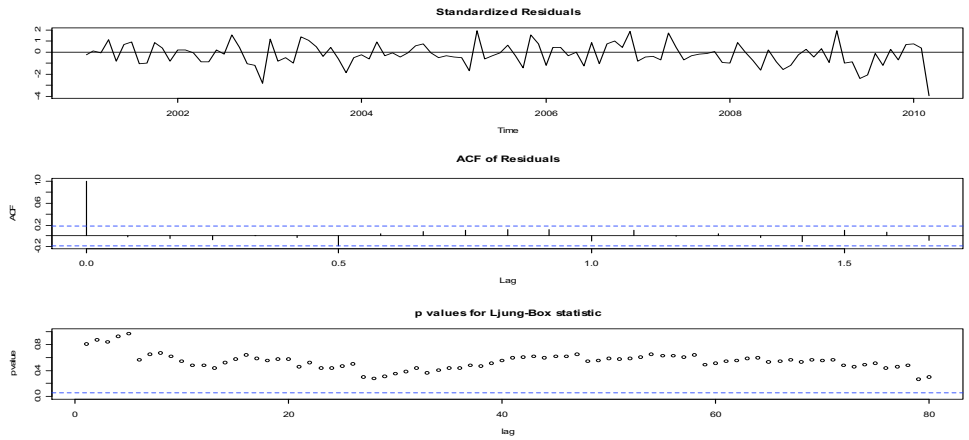

Fig. 3.10: *grafici diagnostici relativi al primo modello secondo il criterio BIC: con nessuna deriva, stagionalità costante/moltiplicativa ed innovazione moltiplicativa .*

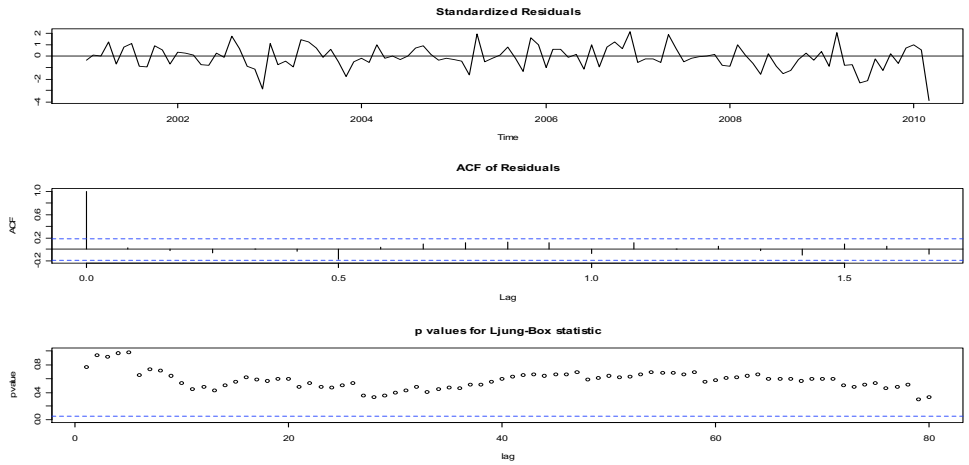

Fig. 3.11: *grafici diagnostici relativi al primo modello secondo il criterio AIC: con deriva costante/moltiplicativa, stagionalità costante/moltiplicativa ed innovazione moltiplicativa .*

I due grafici delle Figure 3.10 e 3.11 non evidenziano significative differenze e sembrano cogliere bene la correlazione tra i residui: infatti nella terza figura di entrambi i grafici, tutti i coefficienti di autocorrelazione stanno al di sopra della banda del 5%, pertanto i residui di entrambi i modelli si presentano come processi whyte-noise, anche se il secondo modello stimato, che propone una deriva costante moltiplicativa, sembra migliore del primo secondo il criterio BIC, sia per numero di parametri stimati, che sono 14 contro i 15 del secondo modello, sia come *p-value* restituito dal test di Ljung-Box. Questo test, che restituisce la significatività "complessiva" dei coefficienti di autocorrelazione, è infatti pari a 0.156, mentre quello relativo al modello senza deriva, è pari a 0.1225. Quindi tra i due, è meglio scegliere il modello proposto secondo il criterio AIC.

Vediamo infine la scomposizione della serie secondo le componenti stimati dal modello del criterio BIC, riprodotta in Figura 3.12

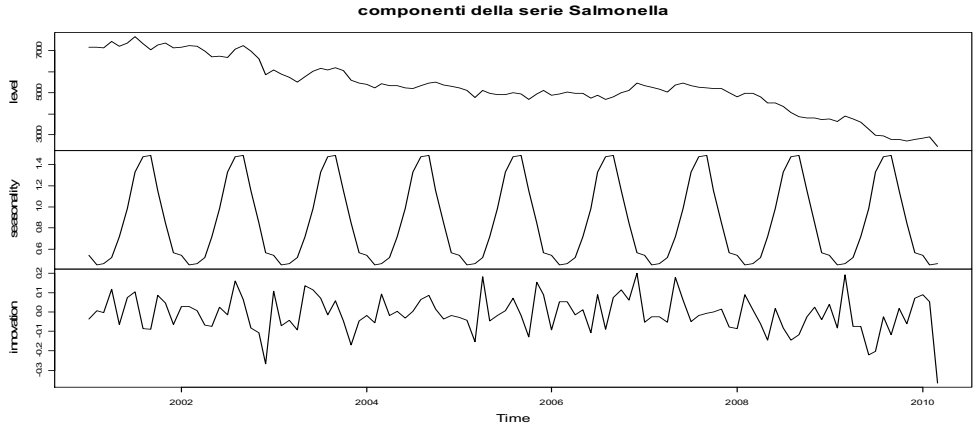

Fig. 3.12: *scomposizione della serie Salmonella*.

3.5 Epatite A.

Le epatiti virali sono infiammazioni che, nel caso delle forme dovute ai virus di tipo B, C e D, possono degenerare in croniche.

L'epatite A è una malattia altamente contagiosa che interessa il fegato; ne è responsabile un piccolo virus chiamato HAV (o virus dell'epatite A), che si trasmette attraverso il consumo di alimenti e bevande contaminate o tramite il contatto diretto con persone infette. Il virus dell'epatite A si trasmette generalmente attraverso il consumo di acqua ed alimenti contaminati da feci infette. E' quindi sufficiente che una persona portatrice del virus manipoli del cibo, senza essersi accuratamente lavata le mani dopo un soggiorno alla toilette, per trasformarlo in un pericoloso veicolo di infezione.

Di seguito si elencano i dati e successivamente si visualizza la loro rappresentazione grafica in Figura 3.13, sempre riferiti allo stesso periodo dei dati precedentemente analizzati:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001	251	196	170	147	146	128	123	154	289	337	189	142
2002	175	119	88	87	64	78	56	109	177	214	186	123
2003	147	113	131	83	90	94	68	68	136	151	147	140
2004	148	115	126	153	117	90	112	288	323	191	140	136
2005	128	102	88	67	78	66	72	133	158	127	110	89
2006	80	102	92	85	90	76	79	82	176	155	111	101
2007	79	59	64	58	56	55	62	77	130	125	106	68
2008	78	93	53	84	50	63	83	90	128	168	111	72
2009	86	78	81	75	92	69	63	76	110	85	57	56
2010	47	67	18									

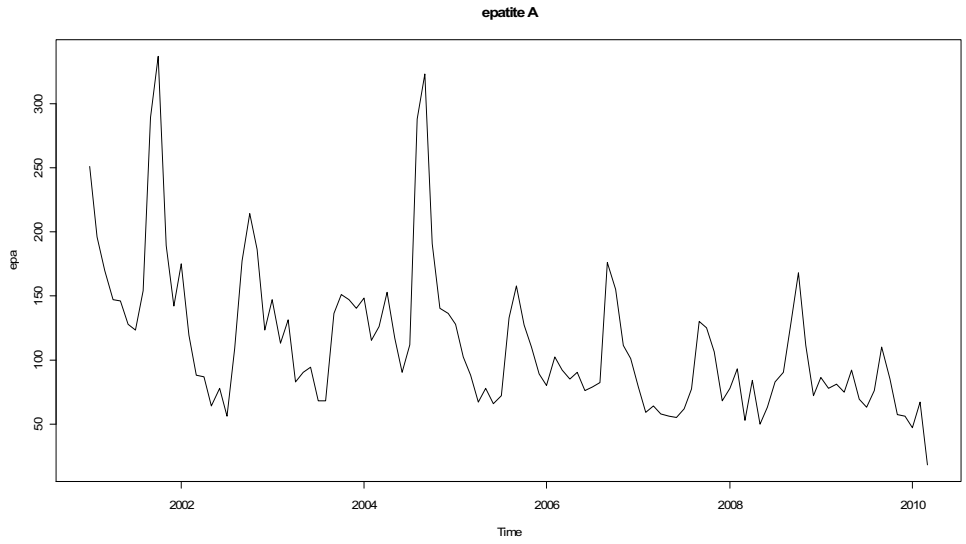

Fig. 3.13: *serie storica relativa all'infezione da Epatite A*.

Si vedano i grafici relativi alla componente stagionale:

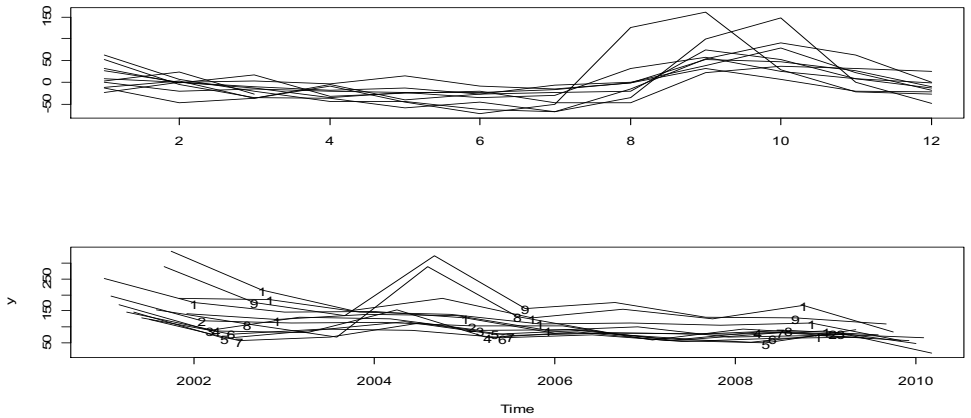

Fig. 3.14: *sottoserie mensili e profilo stagionale della serie storica relativa all'infezione da Epatite A.*

Dalla Figura 3.14, si nota un po' di irregolarità nel *profilo stagionale*, in particolare negli anni 2004 -2005, mentre, con riferimento al primo grafico, la malattia sembra colpire nei mesi autunnali con picchi che interessano i mesi di settembre e ottobre di ogni anno, e minimi che si verificano nei mesi estivi tra giugno e luglio. Si nota un andamento decrescente dei dati durante i nove anni di osservazione e la stagionalità sembra di tipo moltiplicativo. Di seguito si riporta l'esito degli adattamenti dei modelli:

```
   drift sea inn np nlog.lik      BIC      AIC rankAIC
1      n c/m   m 14 1213.406 1279.340 1241.406       1
2      n   m   m 15 1215.084 1285.727 1245.084       2
3    c/m c/m   m 15 1217.135 1287.778 1247.135       4
4      m c/m   m 16 1213.652 1289.004 1245.652       3
5    c/m   m   m 16 1219.526 1294.879 1251.526       5
6      n c/a   m 14 1230.909 1296.842 1258.909       6
7      n   a   m 15 1232.002 1302.645 1262.002       7
8      n c/m   a 14 1258.256 1324.189 1286.256       9
9    c/m c/m   a 15 1254.558 1325.200 1284.558       8
10   c/a c/m   a 15 1256.592 1327.235 1286.592      10
11   c/m   m   a 16 1254.627 1329.979 1286.627      11
12     m c/m   a 16 1254.821 1330.173 1286.821      12
```

I due criteri sono concordi solo per i primi due modelli e differiscono solo per il tipo di stagionalità proposta: il primo suggerisce una stagionalità di tipo costante/moltiplicativo, mentre il secondo una stagionalità di tipo moltiplicativo. Stimiamo allora il primo modello:

```
Call: esFit(y = epa, drift = "n", seasonality = "c/m", innovation = "m")
drift=none, seasonality=c/multiplicative, innovation=multiplicative
      alpha     l.start        s[1]        s[2]        s[3]        s[4]
  0.5822963 199.6754449   1.1564306   1.1020540   0.9477769   0.9362414
       s[5]        s[6]        s[7]        s[8]        s[9]       s[10]
  0.8705805   0.7989277   0.8165655   1.3004493   1.8673161   1.8853488
      s[11]       s[12]       sigma
  1.4246242   1.1318508   0.2114963
-2log(likelihood)= 1213.392  AIC= 1241.392   BIC= 1279.325
```

Come osservato dal grafico della Figura 3.14, si osserva che il coefficiente stagionale più elevato è quello stimato per il mese di ottobre pari ad 1.88, seguito immediatamente da quello associato al mese di settembre, periodo in cui, come detto, la malattia colpisce maggiormente nel mese di ottobre, registrando il massimo numero di casi.L'analisi grafica dei residui del modello è riportata in Figura 3.15:

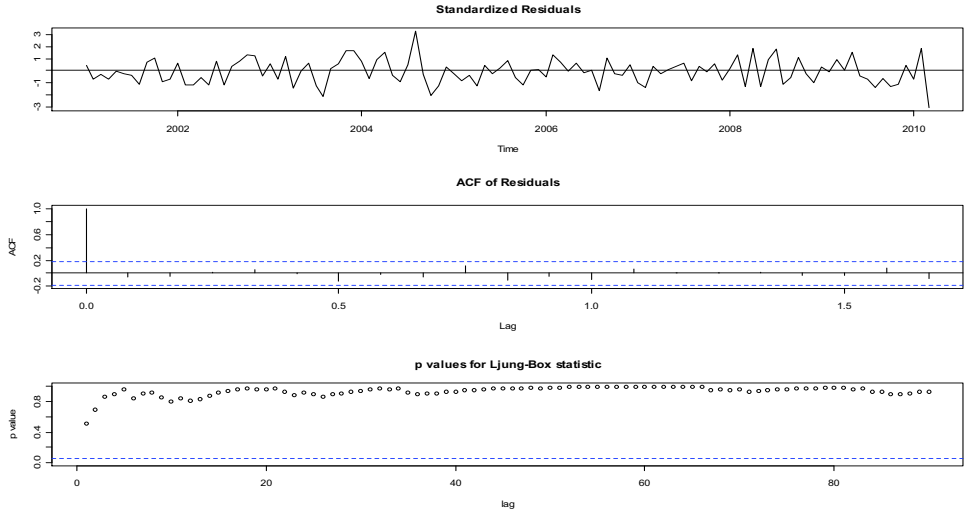

Fig. 3.15: *grafici diagnostici relativi al modello stimato* (*nessuna deriva, stagionalità costante/moltiplicativa e innovazione moltiplicativa).*

Il modello sembra cogliere molto bene la serie analizzata: infatti i residui non presentano alcuna correlazione seriale: test di Ljung – Box ha un p–value pari a 0.912. Il grafico delle sue componenti è mostrato in Figura 3.16:

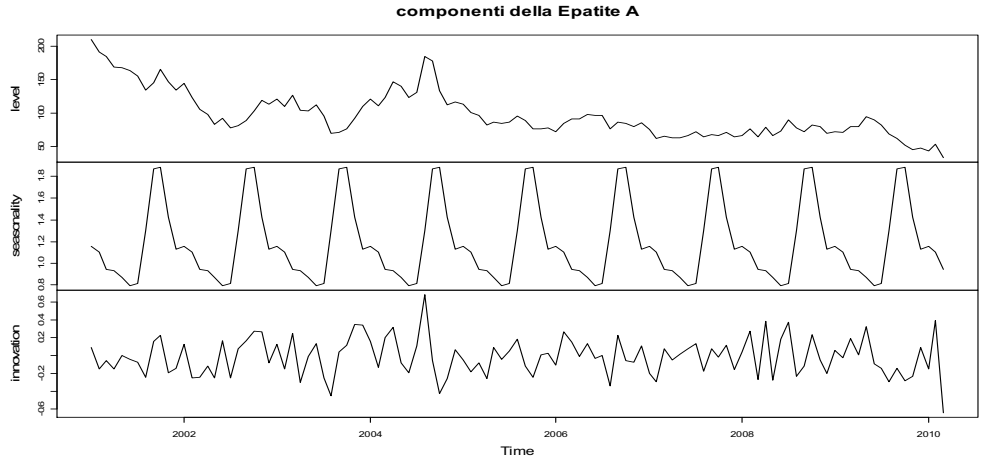

Fig. 3.16 *grafici delle componenti della serie storica relativa all'Epatite A stimate dal modello: nessuna deriva, stagionalità costante/moltiplicativa e innovazione moltiplicativa.*

Visto la buona adattabilità del modello, si sono sovrapposti ai dati osservati i dati stimati, dal modello basato sulle medie mobili. Una

media mobile di periodo *L* consiste in una serie di medie aritmetiche calcolate su una finestra temporale di valori osservati di lunghezza *L*. Il lisciamento consiste dunque nel calcolare queste medie in corrispondenza di ciascuna osservazione, conservando ogni volta la stessa ampiezza della finestra temporale. Proprio per la scelta dell'ampiezza temporale, questa è una tecnica altamente soggettiva, in quanto dipende dalla lunghezza del periodo scelto per la costruzione delle medie. Volendo eliminare le fluttuazioni cicliche della serie, l'analista dovrebbe in qualche modo stimare la durata media dei cicli all'interno della serie e sulla base di questa stima procedere al calcolo delle medie mobili.

Il comando *stlId*(·) è una procedura per la stima simultanea del trend e della stagionalità di una serie temporale e, in particolare con *stlId*(·), la serie viene scomposta nel miglior modo possibile scegliendo in maniera ottimale i parametri di lisciamento[42]. Poiché il primo modello a lisciamento esponenziale suggeriva una stagionalità costante moltiplicativa, si propone una stagionalità moltiplicativa come secondo argomento della funzione *stlId*(·) e, una volta applicato il comando, si visualizza il grafico delle componenti della serie stimate secondo il lisciamento delle medie mobili.

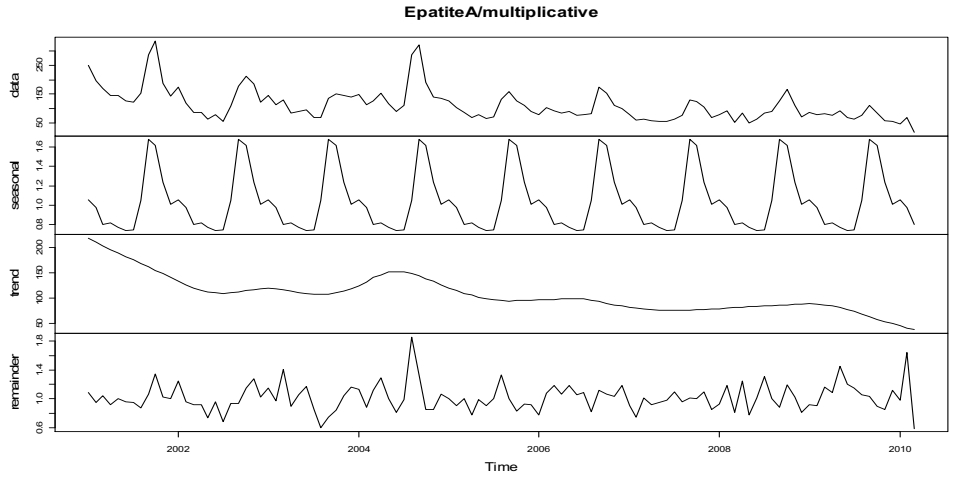

Fig. 3.17: *scomposizione della serie storica relativa all'Epatite A stimate secondo il metodo delle medie mobili con stagionalità moltiplicativa.*

Il grafico della Figura 3.17 visualizza rispettivamente la serie originale, la componente stagionale di tipo moltiplicativo, secondo quanto

[42] Per maggiori dettagli sull'argomento,si rinvia a: "*Diario delle lezioni in laboratorio informatico di "analisi delle serie temporali*", Masarotto G. (2004) "*Analisi delle Serie Temporali lucidi delle lezioni a.a. 2004/05*"

suggerito nella funzione *stlId*(·), il trend e la parte residuale. La scomposizione con questa tecnica risulta molto buona: infatti l'analisi dei residui restituisce un p-value del test Ljung – Box pari a 0.9448, quindi il modello coglie molto bene la correlazione seriale. Con le stime del trend e della stagionalità attuate tramite il lisciamento con le medie mobili, si sono costruiti i grafici in Figura 3.18:

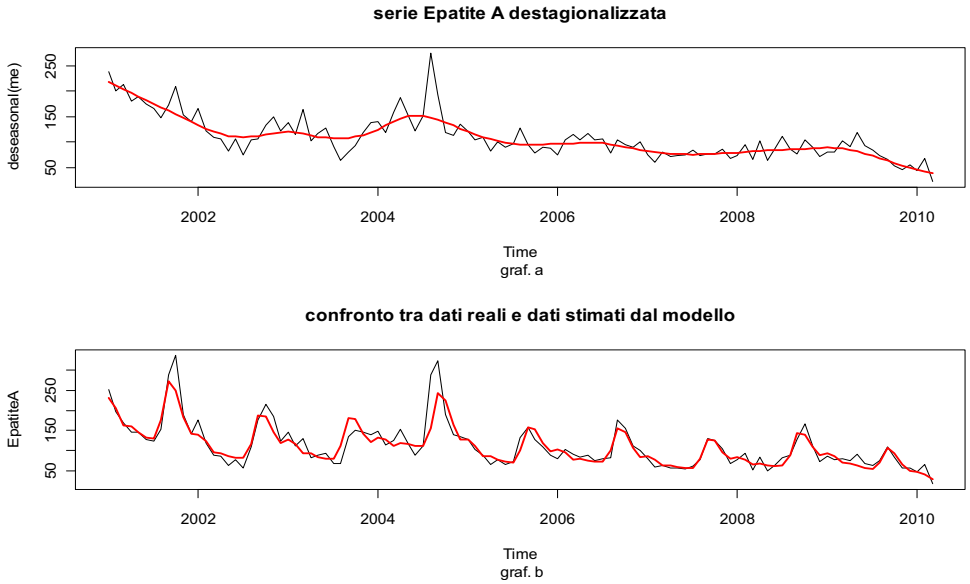

Fig. 3.18: *Fig.a: grafico della serie storica destagionalizzata*
Fig.b: grafico della serie originale e stimata secondo il modello.

A conferma di quanto detto, il grafico della Figura 3.18 evidenzia come la serie stimata approssimi bene la serie storica osservata; in particolare il grafico a. evidenzia la serie storica destagionalizzata, ossia il trend che, come si vede, segue molto bene la tendenza dei valori osservati durante i nove anni di studio, mentre il grafico b. sovrappone la serie stimata (linea rossa) a quella originale.

3.6 Epatite B.

L'epatite B è un virus che colpisce principalmente il fegato e che, se non trattato, può portare a lungo andare cirrosi epatica o tumore del fegato. Si differenzia dall'epatite A per le modalità di trasmissione e per il decorso clinico: l'epatite B infatti si trasmette per via parenterale (esempio: sangue infetto) e per via sessuale; inoltre la malattia ha un decorso cronico.

Questi i dati e il corrispondente grafico in Figura 3.19, riferiti sempre allo stesso periodo di analisi delle malattie infettive esaminate :

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001	221	271	219	220	187	216	262	162	150	168	145	128
2002	178	126	116	111	122	104	115	110	102	144	120	73
2003	130	88	91	99	113	118	123	105	116	111	118	102
2004	128	119	106	105	117	94	104	98	123	111	85	84
2005	96	104	90	93	97	111	132	135	103	78	101	96
2006	102	102	110	85	88	113	112	87	103	90	104	89
2007	94	88	106	75	101	80	86	80	72	80	71	69
2008	67	69	70	59	69	76	64	75	78	74	64	54
2009	58	69	75	62	61	69	73	56	61	57	45	64
2010	55	59	23									

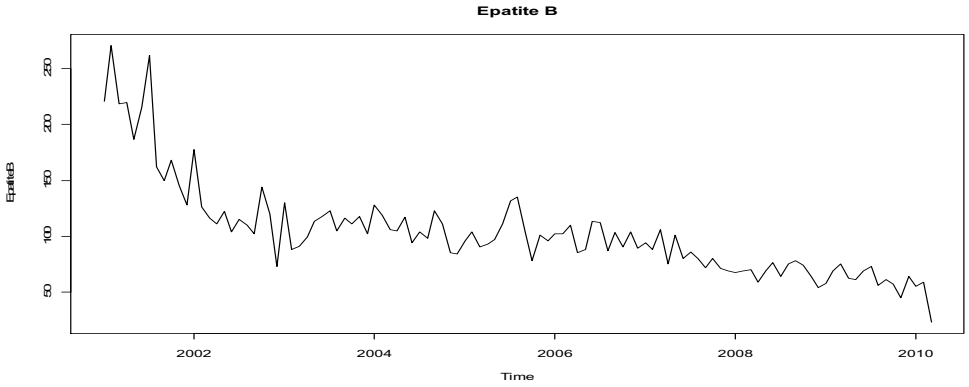

Fig. 3.19: *serie storica relativa all'infezione da Epatite A.*

L'andamento della Figura 3.19 è piuttosto irregolare; inoltre non sembra essere una malattia a carattere stagionale, visto che, dai dati, il virus colpisce indifferentemente nei diversi mesi dell'anno. Si nota un andamento decrescente del numero dei casi: in particolare un notevole calo degli infetti fino al 2003, poi sembra esserci un andamento più o meno costante negli anni che vanno dal 2004 fino al 2007 circa, per poi registrarsi un nuovo leggero calo fino alla fine del periodo considerato, marzo 2010.

Passiamo ad analizzare la componente stagionale, anche se, come già detto, non sembra essere presente.

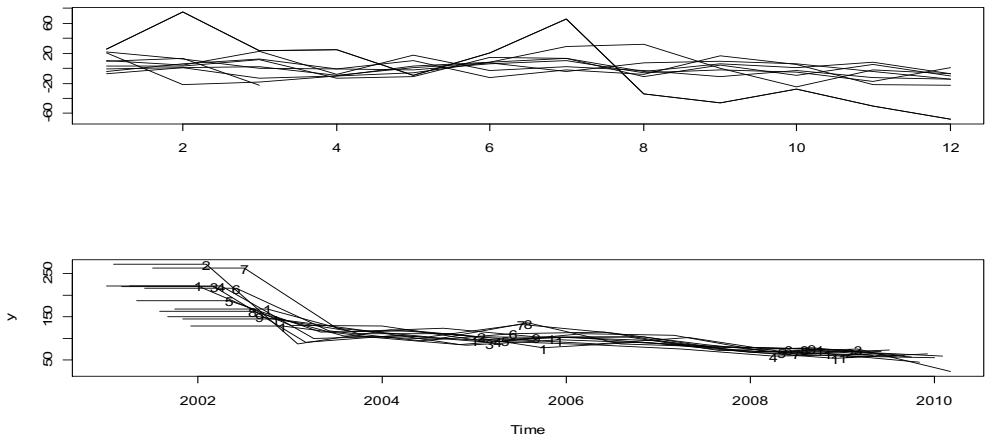

Fig. 3.20: *sottoserie mensili e profilo stagionale della serie storica relativa all'infezione da Epatite B.*

Dal primo grafico della figura 3.20, si nota infatti molta irregolarità e non esiste un periodo in particolare durante il quale il virus registra il massimo numero di infetti. Dal secondo grafico invece, si osserva, per l'intero periodo, una diminuzione dei casi prima commentata con la figura 3.19.

Vediamo ora i modelli a lisciamento esponenziale proposti:

	drift	sea	inn	np	nlog.lik	BIC	AIC	rankAIC
1	n	n	m	2	1132.946	1142.365	1136.946	2
2	c/m	n	m	3	1129.289	1143.418	1135.289	1
3	m	n	m	4	1133.522	1152.360	1141.522	4
4	m	n	a	4	1156.409	1175.247	1164.409	13
5	c/m	n	a	3	1162.434	1176.562	1168.434	14
6	c/m	c/m	m	15	1110.007	1180.650	1140.007	3
7	n	c/m	m	14	1114.896	1180.829	1142.896	5
8	d	n	m	5	1159.584	1183.132	1169.584	15
9	c/a	n	a	3	1169.170	1183.299	1175.170	17
10	n	n	a	2	1174.814	1184.233	1178.814	18
11	a	n	a	4	1165.776	1184.614	1173.776	16
12	n	c/a	m	14	1119.398	1185.331	1147.398	6
13	n	m	m	15	1121.341	1191.984	1151.341	8
14	d	c/m	m	17	1114.398	1194.460	1148.398	7
15	c/m	c/m	a	15	1126.781	1197.424	1156.781	11
16	m	c/m	a	16	1123.900	1199.252	1155.900	9
17	c/m	m	m	16	1123.983	1199.336	1155.983	10
18	c/m	m	a	16	1128.990	1204.343	1160.990	12

Come nel caso dell'Epatite A, anche per la B i due criteri non concordano nella scelta del modello migliore: quello che per il criterio AIC è il primo, è il secondo per il criterio BIC e viceversa, mentre sono quasi diametralmente opposti nella scelta dei modelli a partire dalla terza posizione.

I primi due modelli proposti, si differenziano per il tipo di deriva suggerita: secondo il criterio BIC non esiste alcuna deriva, mentre per il criterio AIC la deriva esiste ed è di tipo moltiplicativo; entrambi invece concordano sul fatto che non esiste alcuna stagionalità, d'altronde come si era prima osservato.

Visto l'evidente calo di casi notato dalle prime analisi grafiche della serie si è provato ad adattare direttamente il secondo modello che corrisponde al primo secondo il criterio AIC. Le stime dei parametri ottenute sono le seguenti:

```
Call: esFit(y = epb, drift = "c/m", seasonality = "n", innovation = "m")
drift=c/multiplicative, seasonality=none, innovation=multiplicative
      alpha     l.start     d.start       sigma
  0.5264403 238.6380048   0.9996000   0.1668763
-2log(likelihood)= 1159.764  AIC= 1165.764   BIC= 1173.893
```

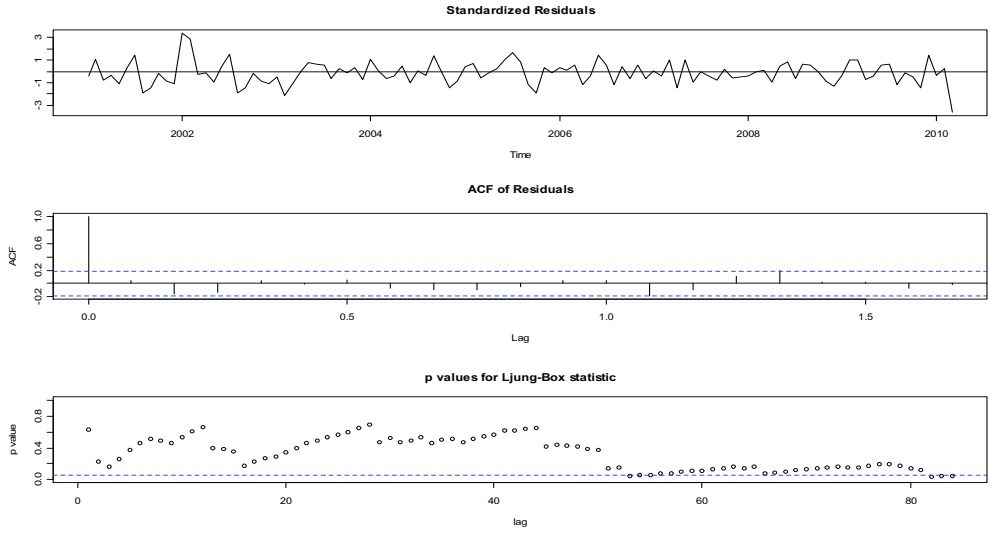

Fig. 3.21: *grafici diagnostici relativi al modello stimato* (*deriva costante/moltiplicativa, nessuna stagionalità e innovazione moltiplicativa)*

I grafici dei residui mostrato in Figura 3.21 sono relativamente buoni, almeno fino al cinquantesimo ritardo, ossia passati circa quattro anni dal periodo iniziale, che è gennaio 2001; rimane poi della correlazione residuale e inoltre c'è parecchia variabilità nei residui. Il test di Box-Ljung non è molto soddisfacente: per un p-value inferiore al 4% i residui infatti non si distribuiscono come un *whyte-noise*; rimane perciò della correlazione residua. Pertanto il modello non riesce a cogliere la variabilità della serie. Le cose peggiorano se si adatta il primo modello secondo il criterio BIC, ossia quello che non riconosce alcuna deriva.

3.7 Legionellosi.

La Legionellosi, è una malattia dovuta a diversi tipi di un batterio molto diffuso nei nostri ambienti (la legionella), che vive bene in ambiente umido. Si presenta in due forme: la febbre di Pontiac, molto simile a un'influenza e quindi raramente identificata; una polmonite, con febbre, tosse, dolori muscolari e in alcuni casi anche difficoltà respiratorie.

Generalmente le forme più gravi si presentano in persone debilitate come anziani, alcolisti, malati cronici. La legionella si trasmette per via aerea, inalando particelle di acqua aeresolizzata, come avviene durante la doccia o in ambienti climatizzati.

Di seguito si riportano i dati

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001	16	20	16	18	15	32	47	37	35	42	26	28
2002	26	27	25	26	27	31	40	59	48	53	37	19
2003	29	28	20	26	23	28	42	48	42	52	31	27

2004	38	21	29	30	36	32	58	46	58	52	41	36
2005	34	33	31	32	36	49	51	79	57	58	40	59
2006	42	41	42	38	39	50	50	56	75	60	53	31
2007	26	30	30	24	30	71	81	67	64	51	34	29
2008	29	22	27	33	40	64	61	64	57	43	54	31
2009	35	36	39	22	29	42	61	63	67	51	31	27
2010	95	32	12									

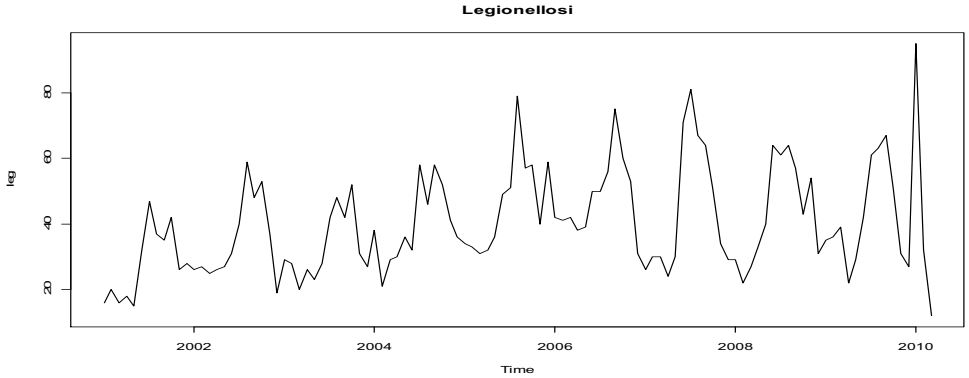

Fig. 3.22: *serie storica relativa all'infezione da Legionella.*

La serie riportata in Figura 3.22, sembra presentare un leggero trend crescente a partire dal 2001 fino al 2007 circa, per poi più o meno stabilizzarsi. La stagionalità invece è un po' irregolare, anche se la malattia sembra più colpire nei mesi autunnali: in particolare il massimo numero dei casi, durante il decennio considerato, si registra tra i mesi di agosto-novembre. Vediamo allora la stagionalità, in Figura 3.23

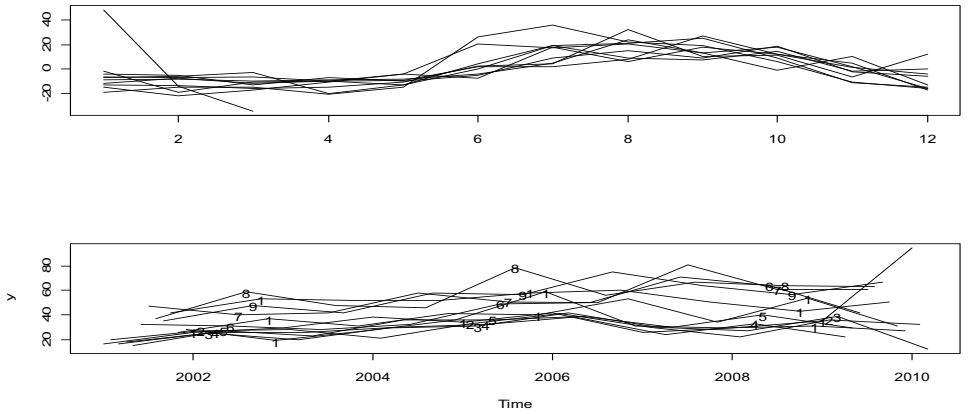

Fig. 3.23: *sottoserie mensili e profilo stagionale della serie storica relativa all'infezione da Legionella.*

La componente stagionale presenta un po' di irregolarità e varia durante gli anni. Si passi all'individuazione del modello migliore.

	drift	sea	inn	np	nlog.lik	BIC	AIC	rankAIC
1	n	c/m	m	14	1013.624	1079.557	1041.624	2
2	c/m	c/m	m	15	1010.666	1081.309	1040.666	1
3	n	m	m	15	1013.446	1084.089	1043.446	3
4	n	c/a	m	14	1019.886	1085.819	1047.886	5
5	n	a	m	15	1019.885	1090.528	1049.885	7
6	n	c/m	m	16	1015.483	1090.835	1047.483	4
7	n	c/m	a	14	1026.971	1092.905	1054.971	9

```
8      m   m   m 17 1013.920 1093.982 1047.920       6
9    c/a c/m   a 15 1024.816 1095.459 1054.816       8
10   c/m c/m   a 15 1024.984 1095.627 1054.984      10
11     n c/a   a 14 1031.111 1097.044 1059.111      13
12   c/a c/a   a 15 1028.272 1098.915 1058.272      12
13     a c/m   a 16 1024.427 1099.779 1056.427      11
```

Anche in questo caso i due criteri non sono molto in accordo; comunque tra i primi due modelli, si consideri il primo secondo il criterio AIC, ossia quello suggerisce la presenza di una deriva e di una stagionalità di tipo costante/moltiplicativa, e una innovazione di tipo moltiplicativo.

```
Call: esFit(y = leg, drift = "c/m", seasonality = "c/m", innovation = "m")
drift=c/multiplicative, seasonality=c/multiplicative, innovation=multiplicative
     alpha    l.start    d.start       s[1]       s[2]       s[3]       s[4]
 0.1593573 26.9786137  1.0124763  1.0162457  0.6851657  0.6519840  0.6572538
      s[5]       s[6]       s[7]       s[8]       s[9]      s[10]      s[11]
 0.7031052  1.0451309  1.2857534  1.3163197  1.2300593  1.1801064  0.8636021
     s[12]      sigma
 0.7521891  0.2294561
-2log(likelihood)= 1014.285  AIC= 1044.285   BIC= 1084.928
```

In base alle stime dei parametri del modello, con riferimento alla stagionalità, si osserva che il parametro più alto è associato al mese di agosto. Si passi ora all'analisi dei residui del modello, mostrato in Figura 3.24

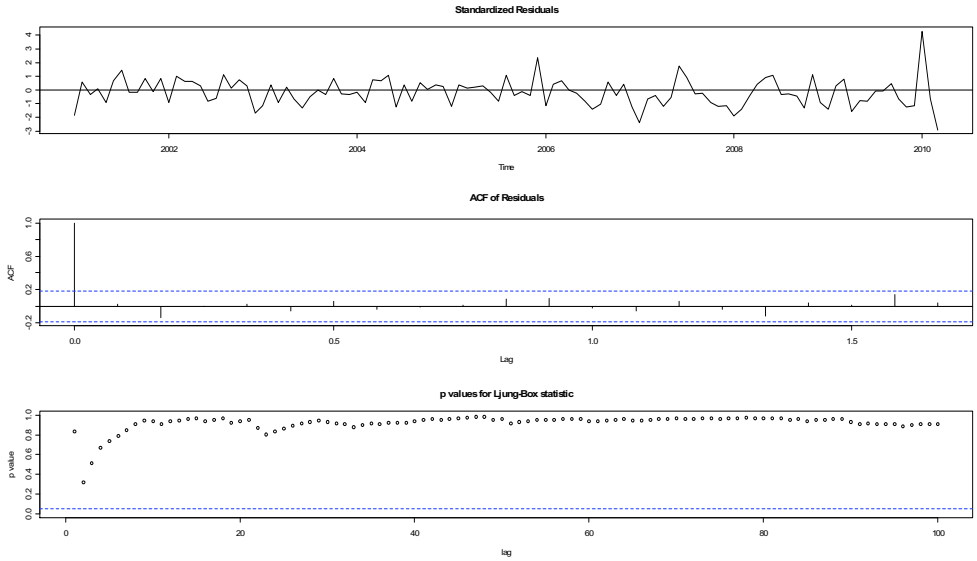

Fig. 3.24: *grafici diagnostici relativi al modello stimato* (*deriva e stagionalità costante/moltiplicativa e innovazione moltiplicativa).*

Il modello sembra adattarsi bene alla serie storica considerata; anche il test di Box-Ljung è moto buono presentando un p-value pari a 0.91. Considerando il primo modello secondo il criterio BIC, le cose non cambiano molto per quanto riguarda la stima dei parametri, che rimane pressoché invariata, mentre migliora leggermente il test di Box-Ljung, che restituisce un p-value pari a 0.97 allo stesso numero di ritardi.

Le componenti della serie stimata secondo il primo modello del criterio BIC sono mostrate in Figura 3.25

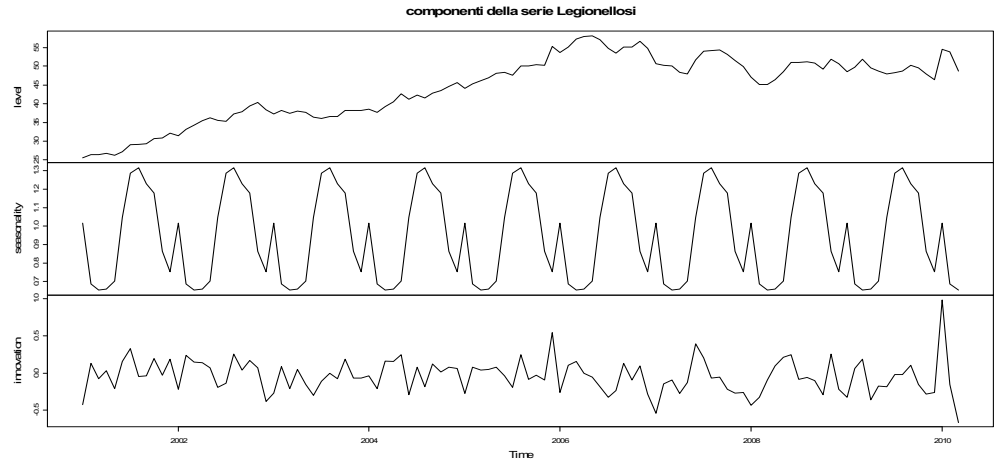

Fig. 3.25: *grafici delle componenti della serie storica relativa alla Legionellosi.*

Notiamo dalla prima Figura 3.25 come la serie presenti un andamento come quello descritto all'inizio del presente paragrafo: prima un aumento continuo fino a circa il 2006 per poi più o meno stabilizzarsi dopo un lieve calo tra il 2006 e il 2007.

Vista la buona adattabilità del modello alla serie, si è provato a sovrapporre alla serie originali le componenti di stagionalità e trend stimate però con la funzione *stlId*$(\cdot)$ come si è fatto con la serie storica relativa all'Epatite A del paragrafo 3.5, considerando una stagionalità di tipo moltiplicativo. Il grafico delle componenti stimate usando il metodo di lisciamento delle medie mobili, è riportato in Figura 3.26.

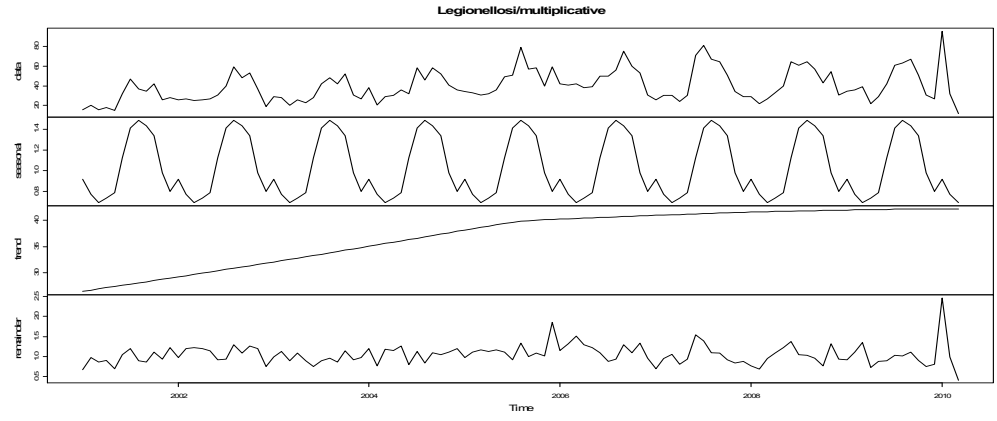

Fig. 3.26 *scomposizione della serie storica relativa alla Legionellosi, stimata secondo il metodo delle medie mobili con stagionalità moltiplicativa.*

Come si vede dal grafico della Figura 3.26, la scomposizione riesce a cogliere bene la correlazione seriale dei dati, confortata anche da un buon p-value del test di Box-Ljung sui residui che è pari a 0.935. Come

si nota nel secondo pannello dello stesso grafico, la componente stagionale non varia nel tempo: infatti i coefficienti mensili stimati rimangono uguali per tutti gli anni, vale cioè:

$$S_t = S_{t-12}.$$

Inoltre, vale la pena notare che i residui del terzo grafico della Figura 3.26 presentano, alla fine del periodo, un discreto aumento rispetto all'andamento degli anni precedenti; si può ad esempio trattare di un cambiamento nelle modalità di rilevazione che la stima non riesce molto a cogliere. Infine, possiamo verificare se il trend stimato si adatta bene alla serie: per questo si visualizza il grafico della serie destagionalizzata e sovrapponendo il trend stimato; come mostrato in Figura 3.27.

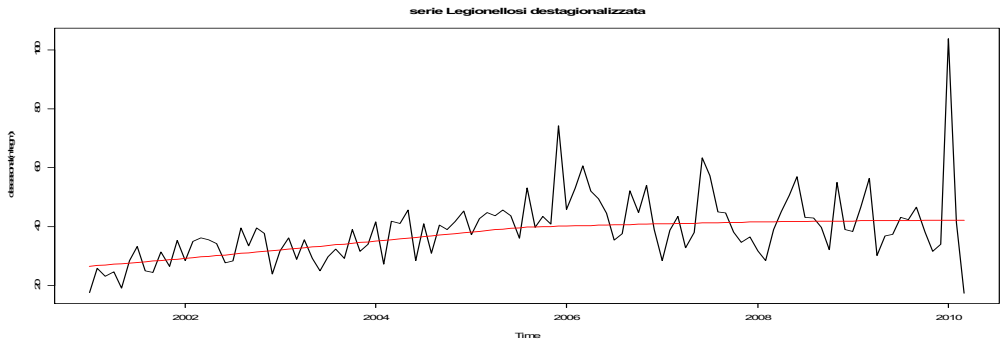

Fig. 3.27 : *serie Legionellosi: serie destagionalizzata e stima del trend prodotti dal stlId*(·).

La stima del trend sembra seguire abbastanza bene l'andamento della serie destagionalizzata: qui si vede meglio che il trend è sempre crescente, anche se verso la fine del 2005 si nota che la sua crescita quasi si appiattisce continuando così fino alla fine del periodo. Continuando con la verifica della stima di *stlId*(·), si consideri adesso solo la componente stagionale, togliendo cioè alla serie originale la componente di trend e visualizzando, appunto, solo quella stagionale.

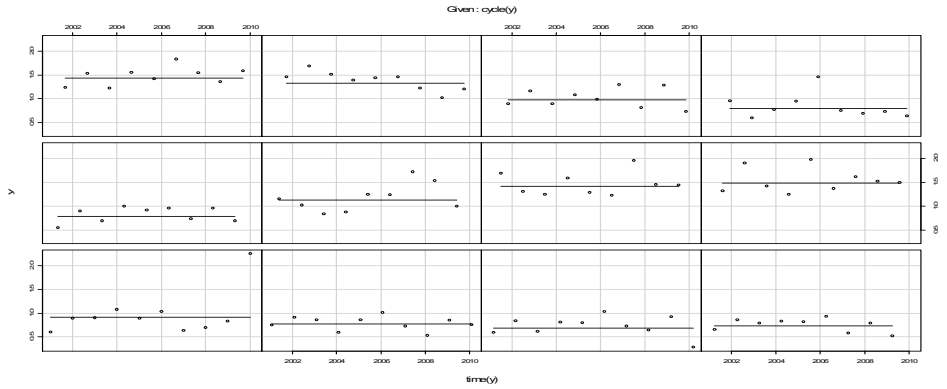

Fig. 3.28: *sottoserie mensili della serie Legionellosi con il trend stimato eliminato e componente stagionale stimata.*

In Figura 3.28, vediamo che la stagionalità nei vari mesi rimane costante proprio per il fatto che, come detto in precedenza, i coefficienti stagionali rimangono costanti lungo tutto il decennio considerato. Inoltre, con il grafico della Figura 3.28, si nota bene come nel primo quadrimestre (i primi quattro pannelli in basso a partire da sinistra) i coefficienti siano molto simili tra di loro, mentre nel secondo quadrimestre comincino ad aumentare per poi decrescere nell'ultimo quadrimestre. È da notare come il coefficiente più alto sia quello associato al mese di agosto, esattamente come succedeva con il modello a lisciamento esponenziale. Si è detto che il modello coglie bene la serie temporale. Un modo per verificarlo è anche quello di visualizzare la funzione di autocorrelazione dei residui.

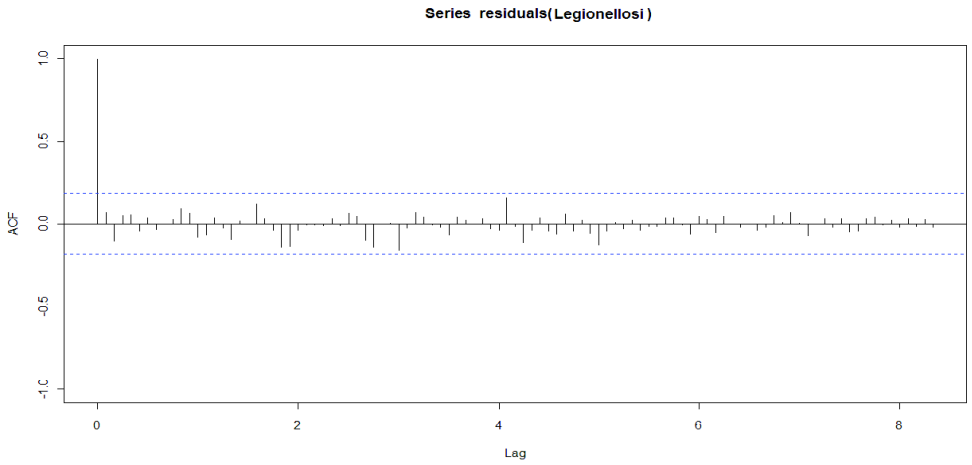

Fig. 3.29 : *stima della funzione di autocorrelazione della componente irregolare.*

Come si vede dalla Figura 3.29 la funzione di autocorrelazione dei residui è buona[43]: infatti, a parte il primo coefficiente[44], tutti gli altri si trovano dentro le due bande: questo significa che i residui sono *whyte-noise*, ossia non rimane alcuna correlazione residua e pertanto la stima delle due componenti coglie bene la correlazione seriale.

Infine, per concludere, si sovrappongono alla serie originale, la serie stimate nelle sue componenti di trend e stagionale secondo il modello moltiplicativo:

[43] La funzione di autocorrelazione è stata calcolata fino a 100 ritardi.
[44] Al ritardo zero infatti, h=0, perché si ha la correlazione della prima osservazione con sé stessa, in questo caso la correlazione del primo residuo con sé stesso.

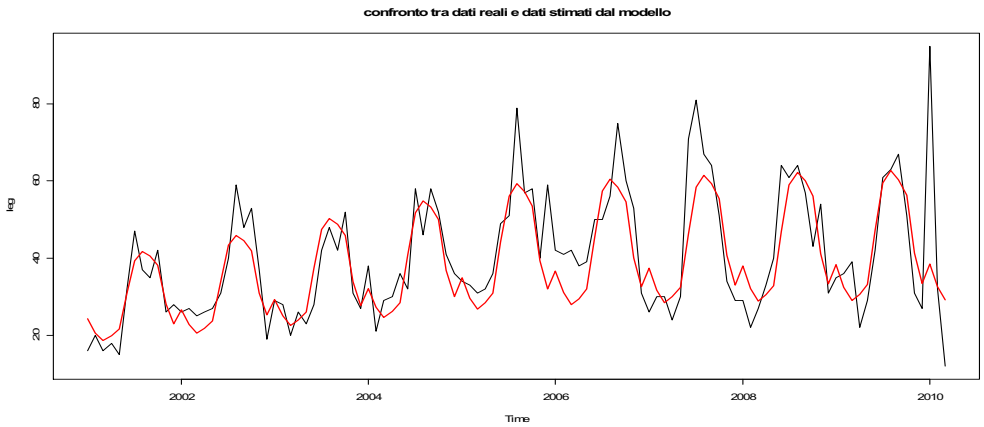

Fig. 3.30: *serie storica originale Legionellosi e serie stimata con il lisciamento a medie mobili.*

Come si vede dal grafico della Figura 3.30 la serie stimata approssima bene quella originale, anche se proprio alla fine del periodo si nota un cambiamento che non sembra essere molto recepito dalla stima, d'altro canto questo lo si era osservato analizzando i residui del grafico della Figura 3.29.

3.8 Tifo.

Il tifo è una grave malattia infettiva di origine batterica, che coinvolge l'intero organismo e che ancora è presente (endemica) in molte parti del mondo. Per questo motivo si può contrarre facilmente nei paesi dove vi è scarsa igiene o dove l'acqua da bere non è sufficientemente sicura. Una percentuale molto piccola degli individui che contraggono la malattia muore e alcuni possono rimanere portatori per molti anni. In alcune aree del mondo il batterio che provoca il tifo sta diventando resistente agli antibiotici rendendo il trattamento difficile. Il tifo è causato dal batterio *Salmonella typhi*. Il tifo si trasmette per via oro-fecale ovvero attraverso cibi e bevande contaminati dalle feci di un malato di tifo o di un portatore del batterio.

Di seguito si riportano i dati:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001	5	5	5	7	12	8	5	13	12	8	7	2
2002	6	3	0	6	8	6	4	6	8	5	5	2
2003	6	5	7	9	6	2	3	6	12	4	4	2
2004	6	3	4	6	7	5	11	8	17	8	4	3
2005	3	7	11	14	7	0	3	4	10	7	5	9
2006	4	5	11	6	9	5	4	5	7	7	8	4
2007	3	4	2	4	3	2	1	14	7	0	12	7
2008	3	4	3	4	7	7	3	12	11	8	3	4
2009	4	4	5	5	6	2	1	7	11	6	8	6
2010	5	4	1									

ed il relativo grafico in Figura 3.31.

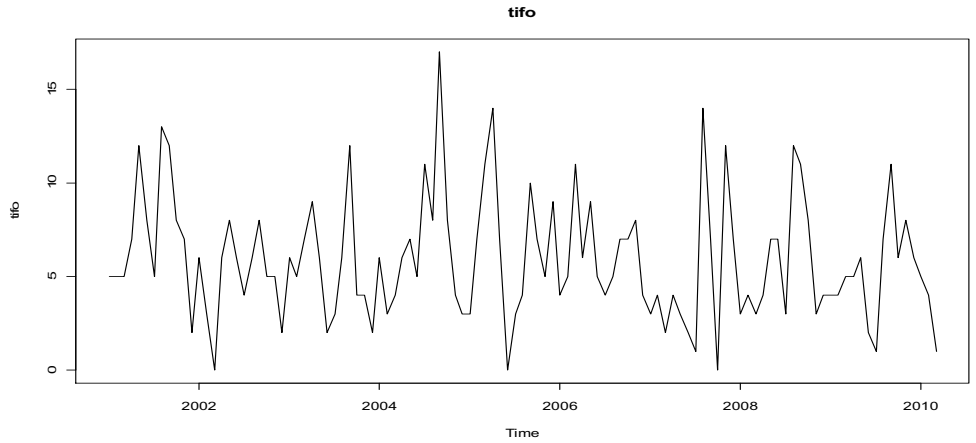

Fig. 3.31: *serie storica relativa all'infezione da Tifo.*

La serie non sembra presentare una evidente stagionalità e nemmeno un trend; è piuttosto irregolare. Vediamo la funzione di autocorrelazione parziale della serie storica, mostrata in Figura 3.32.

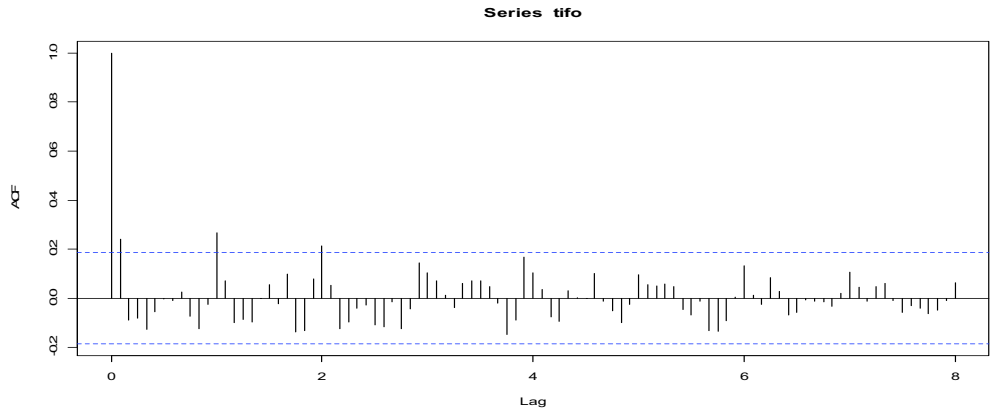

Fig. 3.32: *Funzione di autocorrelazione della serie storica relativa all'infezione da Tifo.*

A parte agli inizi del primo e secondo anno, le osservazioni non sono autocorrelate tra di loro: infatti quasi tutti i coefficienti sono dentro le bande disegnate. Vediamo cosa propone la funzione *esId*(·):

```
   drift sea inn np nlog.lik      BIC      AIC rankAIC
1      n   n   a  2 786.8288 796.2478 790.8288       8
2    c/a   n   a  3 789.6278 803.7564 795.6278      10
3      n c/a   a 14 740.7498 806.6832 768.7498       2
4    c/a c/a   a 15 737.9308 808.5738 767.9308       1
5      d   n   a  5 785.3789 808.9266 795.3789       9
6      n   a   a 15 742.0085 812.6515 772.0085       5
7      a   n   a  4 795.2049 814.0430 803.2049      11
8    c/a   a   a 16 738.8926 814.2451 770.8926       3
9      d c/a   a 17 737.6561 817.7181 771.6561       4
10     a c/a   a 16 751.0153 826.3678 783.0153       6
11     a   a   a 17 754.2388 834.3009 788.2388       7
12     d   a   a 18 769.9561 854.7276 805.9561      12
```

I due criteri non concordano affatto sulla stima delle componenti della serie: secondo il primo modello del criterio BIC (l'ottavo per il AIC), la

serie non presenta alcuna deriva e stagionalità, mentre il primo modello suggerito da AIC (il quarto per il BIC) riconosce una deriva e stagionalità di tipo costante moltiplicativa. Si adatti il primo modello individuato dal criterio BIC:

```
Call: esFit(y = tifo, drift = "n", seasonality = "n", innovation = "a")
drift=none, seasonality=none, innovation=additive
      alpha     l.start       sigma
0.003529081 5.480341943 3.285466380
-2log(likelihood)= 786.8288  AIC= 790.8288   BIC= 796.2478
```

Si noti che l'alfa stimato dal modello risulta essere molto piccolo; quindi l'osservazione corrente influenza ben poco quella successiva, e questo conferma quanto già visto nella funzione di autocorrelazione parziale della serie. Verifichiamo in Figura 3.33 la bontà del modello:

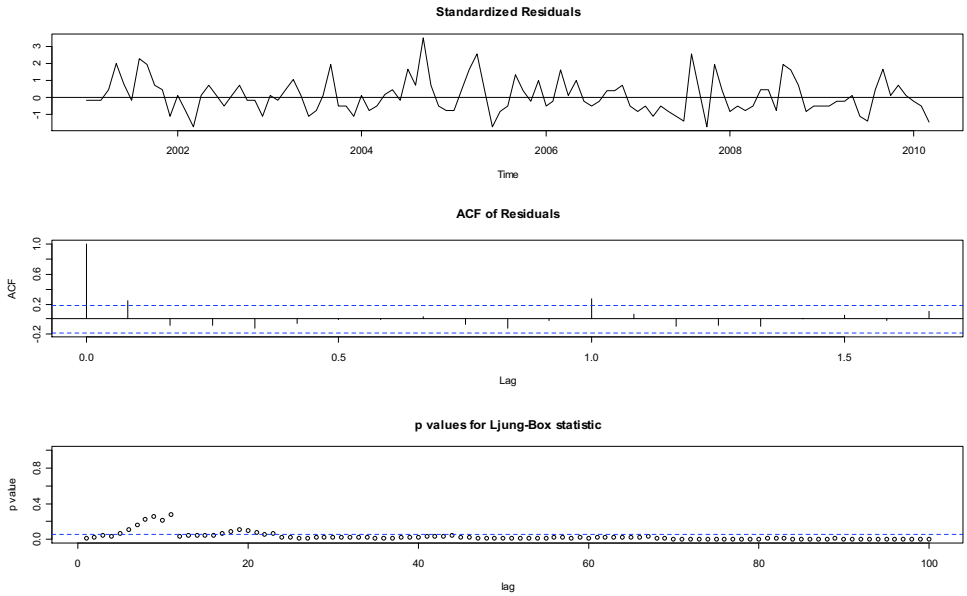

Fig. 3.33: *grafici diagnostici relativi al modello stimato secondo il criterio BIC: nessuna deriva e stagionalità e innovazione additiva.*

Effettivamente, come si vede dal terzo pannello del grafico della Figura 3.33 quasi tutti i coefficienti di autocorrelazione dei residui stanno al di sotto delle banda per cui il modello non coglie molto bene la correlazione seriale dei dati osservati. Le cose non migliorano di molto se si applica il primo modello secondo il criterio BIC, la cui analisi dei residui è mostrata in Figura 3.34

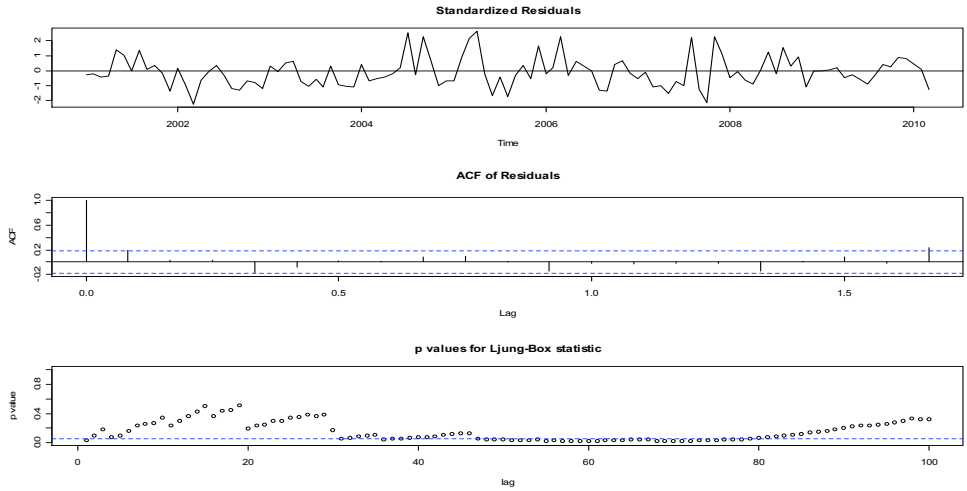

Fig. 3.34: *grafici diagnostici relativi al modello stimato secondo il criterio BIC deriva e stagionalità costante/additiva e innovazione additiva.*

Rispetto al grafico della figura 3.33, vediamo che nel grafico della Figura 3.34, più coefficienti stanno al di sopra della banda, tuttavia nemmeno in questo caso il modello coglie in maniera soddisfacente l'andamento della serie storica. Molto probabilmente questo dipende dal fatto che la serie non ha carattere di stazionarietà, e poiché tutte le stime e le diagnostiche si basano sull'ipotesi di stazionarietà, è evidente che le scomposizioni così effettuate portano a risultati non soddisfacenti. Un modo per rendere più stazionaria la serie potrebbe essere quello di passare alle differenze e successivamente applicare modelli stocastici autoregressivi.

3.9 E. Coli Enterite.

La *E.Coli enterite* è una malattia infettiva delle vie digerenti, che colpisce soggetti provenienti da paesi ad elevato tenore igienico quando si recano in aree in via di sviluppo. Le zone a maggior rischio sono il Sud-Est asiatico, l'India, il Bangladesh ed alcuni paesi dell'Africa e dell'America Centrale, in particolare il Messico. La trasmissione è per lo più fecale-orale: l'agente infettante viene eliminato con le feci dal soggetto malato e chi è contagiato viene a contatto per via orale con il materiale contaminato delle stesse feci infette. La trasmissione della malattia è perciò legata soprattutto alla qualità dell'acqua e delle bevande assunte.

Di seguito i dati:

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001	355	390	363	374	340	382	443	493	585	559	429	362
2002	432	422	417	388	365	392	441	644	576	537	492	256
2003	386	411	423	332	369	391	527	569	669	566	498	337

```
2004 369 354 423 307 361 387 512 595 658 613 547 458
2005 408 396 398 449 419 522 616 583 658 479 511 444
2006 452 532 571 444 431 460 600 543 668 652 601 518
2007 419 467 472 379 457 582 660 566 698 612 604 520
2008 478 506 462 424 427 560 653 755 843 705 673 516
2009 444 417 522 380 427 463 581 665 776 642 498 410
2010 391 375 174
```

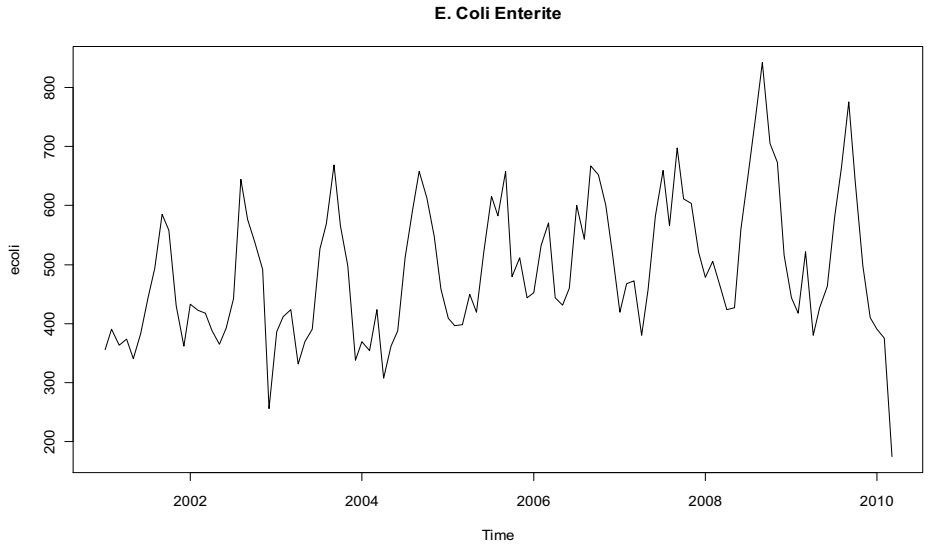

Fig. 3.35: *serie storica relativa all'infezione da E. Coli Enterite.*

Dalla Figura 3.35 si nota un trend leggermente crescente e una stagionalità della serie evidenziata anche dalla funzione di autocorrelazione parziale della serie, in Figura 3.36.

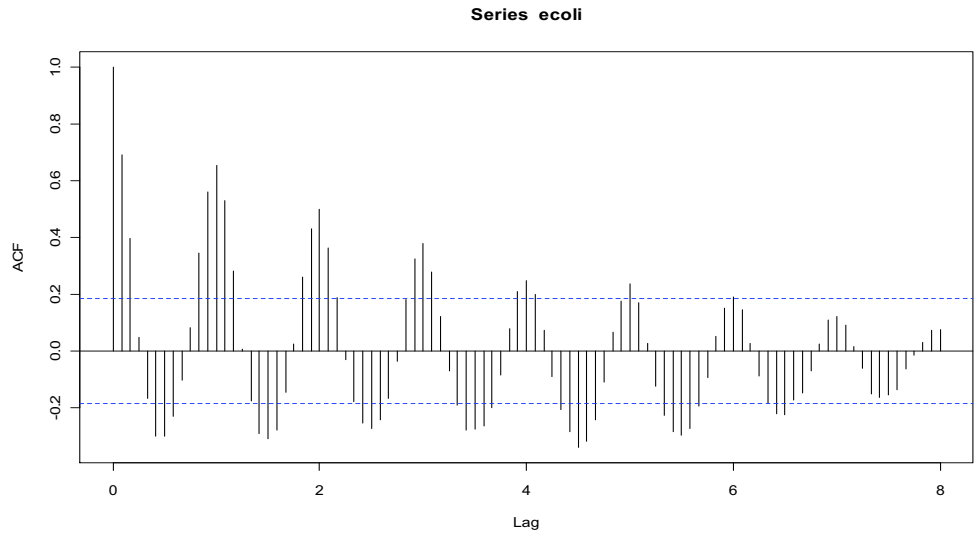

Fig. 3.36: *Funzione di autocorrelazione della serie storica relativa all'infezione da E. Coli Enterite.*

Vediamo la componente stagionale della serie:

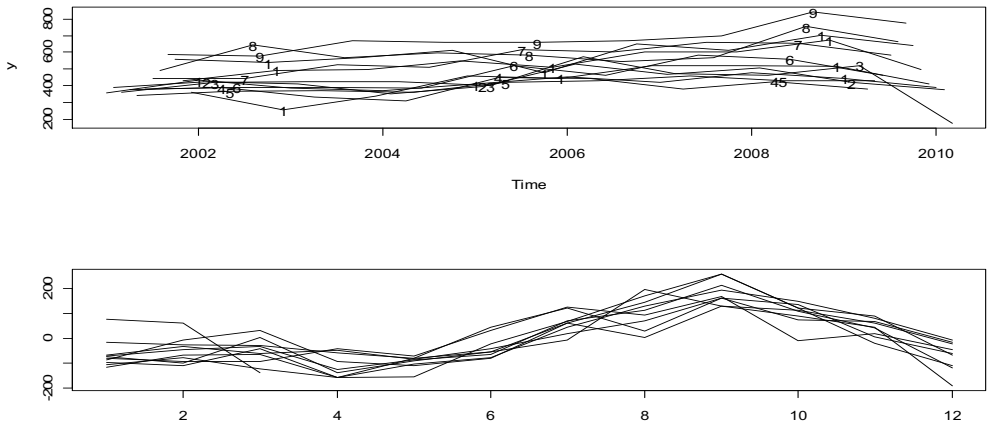

Fig. 3.37: *sottoserie mensili e profilo stagionale della serie storica relativa all'infezione da E. Coli Enterite .*

Come si osserva dalla Figura 3.37, la malattia colpisce in modo particolare durante i mesi estivi e autunnali, facendo registrare il massimo numeri di casi nel mese di settembre, mentre l'andamento di ciascun mese per ogni anno considerato è stato mutevole in particolare durante le stagioni invernali e primaverili. Vediamo i modelli proposti.

```
   drift sea inn np nlog.lik      BIC      AIC rankAIC
1      n c/m   a 14 1403.142 1469.076 1431.142       1
2      n c/a   a 14 1405.665 1471.599 1433.665       4
3    c/m c/m   a 15 1402.826 1473.469 1432.826       2
4    c/a c/m   a 15 1403.030 1473.673 1433.030       3
5      n c/m   m 14 1410.233 1476.167 1438.233       7
6    c/a c/a   a 15 1405.554 1476.197 1435.554       5
7      n c/a   m 14 1414.914 1480.848 1442.914      13
8    c/m c/a   m 15 1410.404 1481.047 1440.404       9
9      n   m   m 15 1410.588 1481.231 1440.588      10
10     a c/a   a 16 1406.848 1482.201 1438.848       8
11     d c/m   a 17 1403.479 1483.541 1437.479       6
12   c/m c/m   m 15 1413.045 1483.688 1443.045      14
13     a c/m   a 16 1409.175 1484.527 1441.175      11
14     m c/m   a 16 1410.519 1485.871 1442.519      12
```

I due modelli sono concordi sulla prima formulazione che propone assenza di deriva, una stagionalità costante moltiplicativa ed una innovazione di tipo additivo. Si stimano allora le componenti della serie:

```
Call: esFit(y = ecoli, drift = "n", seasonality = "c/m", innovation = "a")
drift=none, seasonality=c/multiplicative, innovation=additive
      alpha     l.start        s[1]        s[2]        s[3]        s[4]
  0.4947707 523.7182086   0.6910206   0.7181885   0.7186096   0.6321557
       s[5]        s[6]        s[7]        s[8]        s[9]       s[10]
  0.6556523   0.7613549   0.9261164   0.9860248   1.1241107   0.9822426
      s[11]       s[12]       sigma
  0.8926779   0.7115343  52.7569330
-2log(likelihood)= 1403.142  AIC= 1431.142   BIC= 1469.076
```

Come osservato nel grafico della Figura 3.37 il coefficiente stagionale più elevato è quello associato al mese di agosto. Analizziamo la bontà della scomposizione proposta dai due criteri.

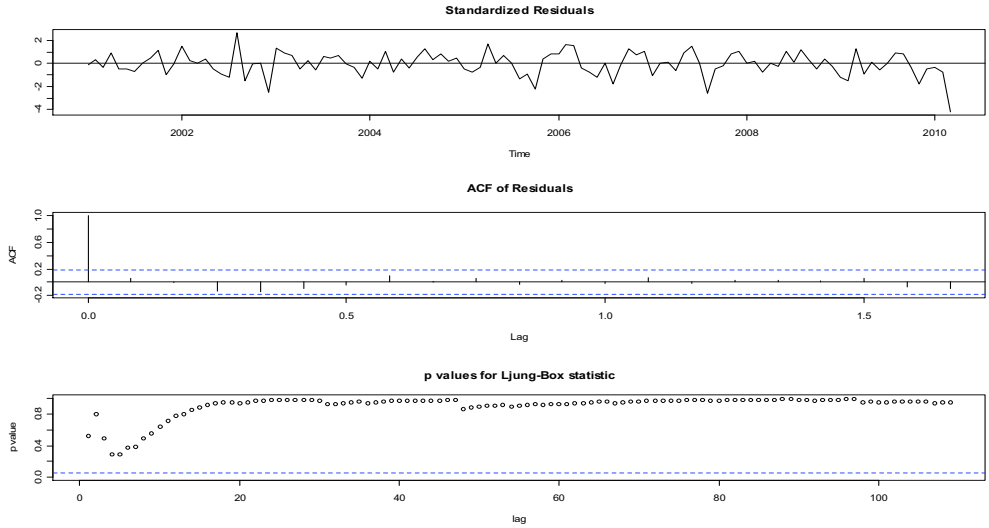

Fig. 3.38: *grafici diagnostici relativi al modello stimato secondo il criterio BIC nessuna deriva, stagionalità costante/moltiplicativa e innovazione additiva.*

Dalla Figura 3.38, si nota chiaramente che la scomposizione del modello coglie molto bene la correlazione seriale dei dati con un test di Box-Ljung molto buono, a cui è associato un p-value pari a 0.98.

Anche in questo caso, visto che il modello coglie bene la serie storica, si sono calcolate le componenti stagionali e di trend mediante lisciamento con medie mobili sovrapponendo poi alla serie osservata i dati stimati in base al lisciamento. Visto quanto suggerito dal modello a lisciamento esponenziale, si è usata una stagionalità moltiplicativa.

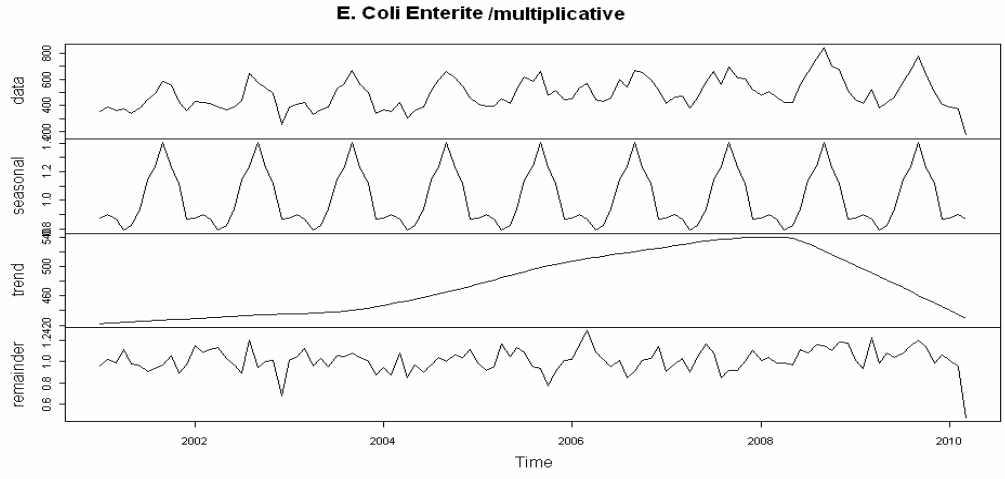

Fig. 3.39: *scomposizione della serie storica relativa alla E. Coli Enterite, stimata secondo il metodo delle medie mobili con stagionalità moltiplicativa.*

Dalla Figura 3.39 si osservano alcune cose che prima non erano cosi evidenti, come ad esempio il trend, la cui crescita comincia ad accentuarsi dal 2005 raggiungendo un massimo agli inizi del 2008, per

poi calare fino alla fine del periodo di osservazione. Si sovrappongano ora il trend alla serie originale destagionalizzata.

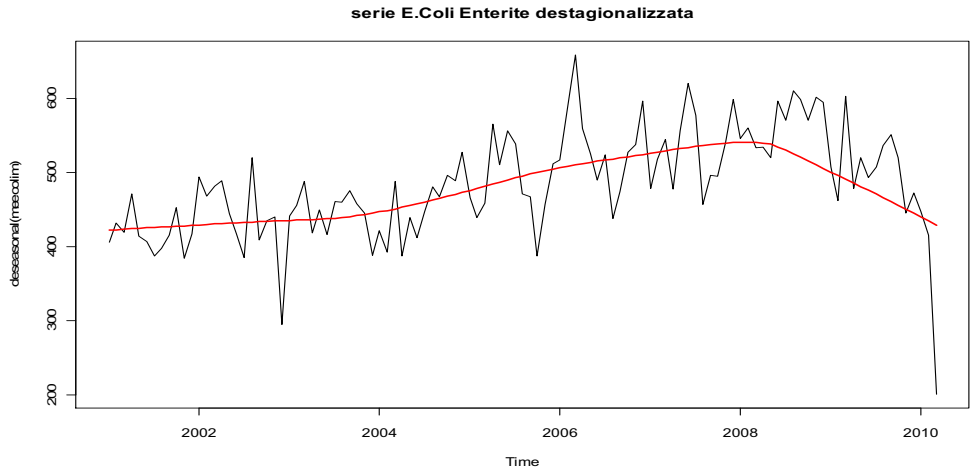

Fig. 3.40: *serie E. Coli Enterite: serie destagionalizzata e stima del trend prodotti dal stlId*(·).

Il grafico della Figura 3.40 sembra cogliere molto bene il trend della serie stimato in base a lisciamento con medie mobili. Infine proviamo a sovrapporre alla serie originale la serie stimata nelle sue due componenti:

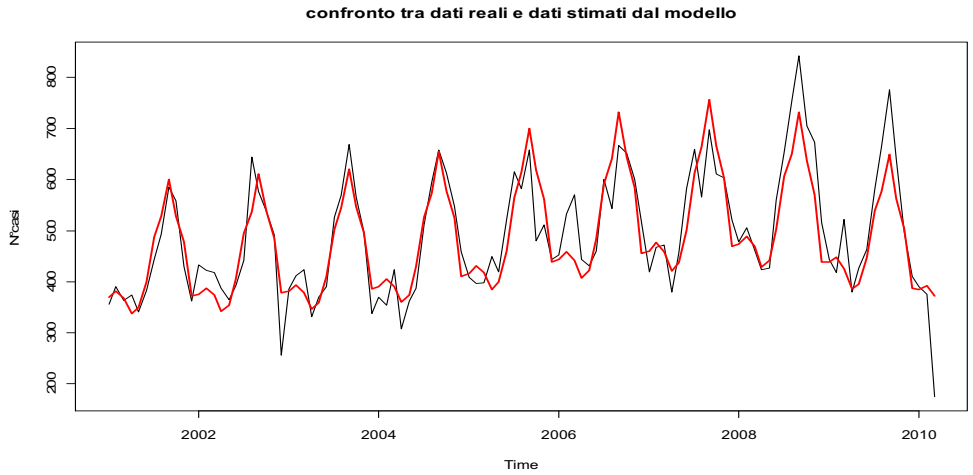

Fig. 3.41 : *serie storica originale E. Coli Enterite i e serie stimata con il lisciamento a medie mobili.*

Come si vede dalla Figura 3.41, anche con il lisciamento a medie mobili la serie viene ben approssimata riuscendo a cogliere quasi tutta la correlazione seriale dei dati: anche il test di Box-Ljung risulta molto soddisfacente con un p-value pari a 0.9432.

3.10 Influenza B.

L'influenza è una malattia respiratoria acuta dovuta alla infezione da virus influenzali. È una malattia stagionale che, nell'emisfero occidentale, si verifica durante il periodo invernale. Ci sono tre tipi di virus: il virus tipo A e il virus tipo B, responsabili della sintomatologia influenzale classica, e il tipo C, di scarsa rilevanza clinica (generalmente asintomatico).

Il virus può essere trasmesso per via aerea dal momento del contagio fino ai tre-quattro giorni successivi ai primi sintomi che si manifestano a distanza di uno-quattro giorni dall'infezione. Questo significa che il virus può essere trasmesso anche da persone apparentemente sane. Si diffonde molto facilmente negli ambienti affollati.

Vediamo ora i dati relativi all'influenza B e la loro rappresentazione grafica, mostrata in Figura 3.42

	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
2001	36	45	49	47	8	4	6	5	4	6	7	6
2002	75	331	685	155	24	2	2	0	0	2	3	2
2003	7	130	356	143	15	3	0	0	2	3	6	6
2004	7	14	8	2	0	0	0	1	2	1	1	19
2005	66	533	1431	185	12	1	1	0	3	0	3	1
2006	5	85	961	371	21	2	0	0	1	1	1	0
2007	6	67	116	30	3	1	2	0	4	1	12	29
2008	566	2575	2375	541	58	7	2	2	6	6	8	71
2009	896	3934	1767	163	44	24	27	25	20	15	43	12
2010	7	4	11									

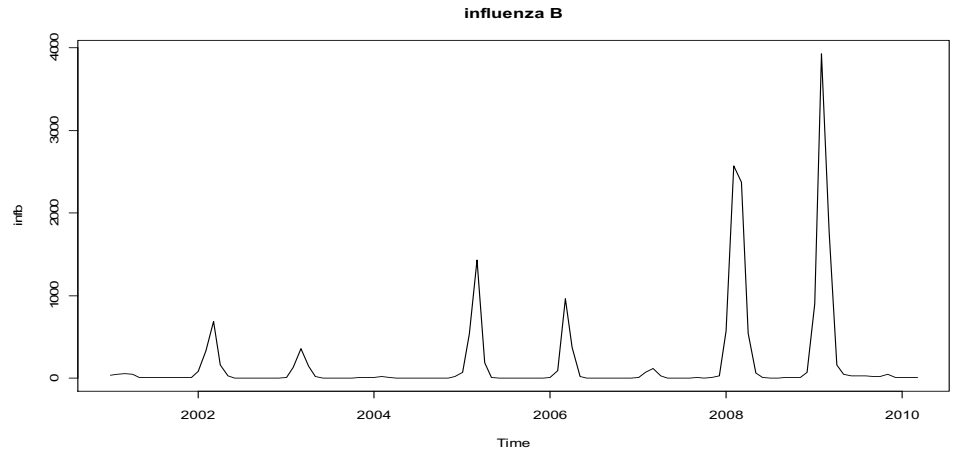

Fig. 3.42: *serie storica relativa all'Influenza B.*

Come si nota dalla Figura 3.42, la malattia non presenta un trend, mentre la componente stagionale risulta essere un po' ambigua: infatti in base ai dati, si osserva che la malattia colpisce principalmente nei mesi di febbraio e marzo, tuttavia l'incidenza nei diversi anni varia di molto. Ci sono infatti alcuni anni, come ad esempio il 2003, 2004, 2006

e 2007, in cui il massimo numero di casi registrati, pur corrispondendo ai mesi di febbraio e marzo, non è così elevato come invece succede nel resto del periodo considerato. Proviamo ad evidenziare la sottoserie mensile dei dati e il profilo stagionale.

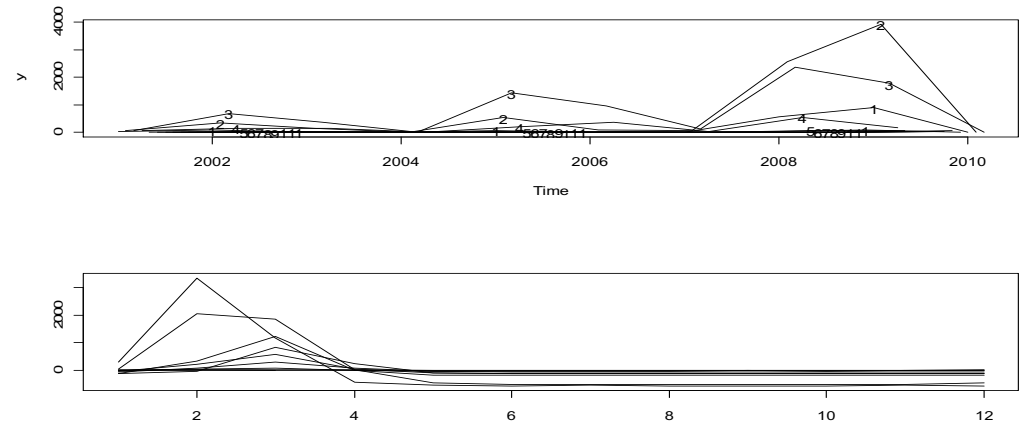

Fig. 3.43: *sottoserie mensili e profilo stagionale della serie storica relativa all'infezione da Influenza B .*

Il secondo grafico della Figura 3.43 conferma quanto detto prima a proposito dei mesi in cui la malattia registra il massimo numero di casi, ossia quelli di febbraio e marzo; tuttavia durante quei mesi si nota anche una certa variabilità dei casi che, verso la fine del periodo analizzato, aumentano in maniera considerevole. Vediamo allora se ci viene in aiuto a questo proposito, la funzione di autocorrelazione parziale delle osservazioni.

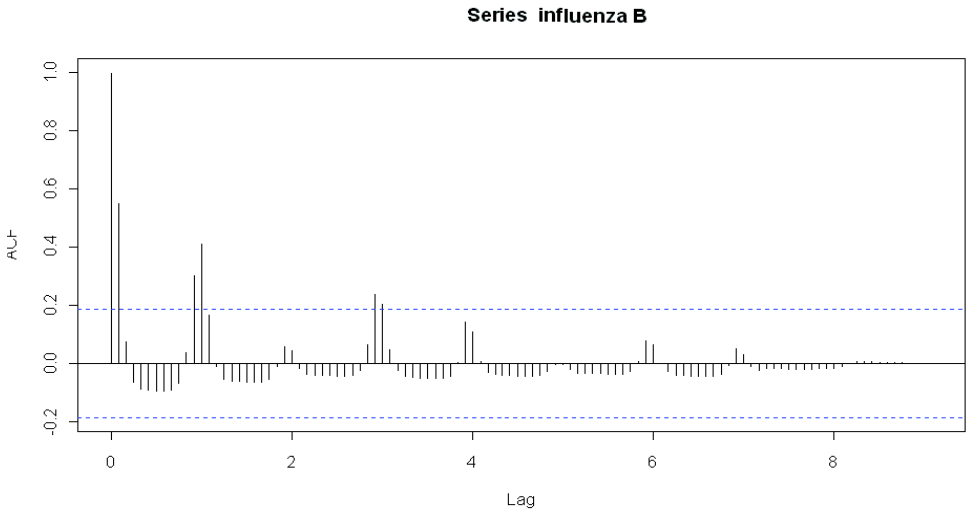

Fig. 3.44: *correlogramma della serie storica relativa all'Influenza B.*

Effettivamente nemmeno la funzione di autocorrelazione parziale della Figura 3.44 non è di molto aiuto per quanto riguarda la componente stagionale. Inoltre il grafico evidenzia che si tratta di osservazioni poco correlate tra di loro: infatti a parte i coefficienti dei primi 14 ritardi (corrispondenti a poco più di un anno di ritardo), si osserva che quasi tutti stanno sotto la soglia del 5%. Vediamo allora che modelli vengono suggeriti dalla funzione *esId*(·):

```
   drift sea inn np nlog.lik      BIC      AIC rankAIC
1      n   n   a  2 1909.776 1919.195 1913.776       8
2    c/a   n   a  3 1909.776 1923.905 1915.776       9
3      a   n   a  4 1909.799 1928.638 1917.799      11
4      d   n   a  5 1909.403 1932.950 1919.403      12
5      n c/a   a 14 1872.413 1938.346 1900.413       1
6    c/a c/a   a 15 1872.395 1943.038 1902.395       2
7      n   a   a 15 1878.166 1948.809 1908.166       4
8      d c/a   a 17 1870.599 1950.661 1904.599       3
9    c/a   a   a 16 1878.192 1953.544 1910.192       5
10     a c/a   a 16 1880.449 1955.801 1912.449       7
11     a   a   a 17 1878.031 1958.093 1912.031       6
12     d   a   a 18 1880.681 1965.452 1916.681      10
```

Si nota innanzitutto che i due criteri sono molto contrastanti tra di loro: il primo modello per il criterio AIC corrisponde quinto per il criterio BIC, mentre il primo per quest'ultimo corrisponde all'ottavo per l'AIC. In ogni caso il miglior modello suggerito dai due criteri si differenzia solamente per il tipo di stagionalità proposta che, secondo il criterio BIC non esiste, mentre viene riconosciuta dal primo modello classificato in base al criterio AIC.

Vista l'incertezza osservata dall'analisi dei grafici delle figure 3.43 e 3.44, si stimano entrambi i modelli:

BIC

```
Call: esFit(y = infb, drift = "n", seasonality = "n", innovation = "a")
drift=none, seasonality=none, innovation=additive
    alpha   l.start     sigma
  1.00000  38.63503 516.88908
-2log(likelihood)= 1909.776  AIC= 1913.776   BIC= 1919.195
```

AIC

```
Call: esFit(y = infb, drift = "n", seasonality = "c/a", innovation = "a")
drift=none, seasonality=c/additive, innovation=additive
       alpha      l.start         s[1]         s[2]         s[3]         s[4]
   0.7942697 -360.9349282  191.2885448  798.5644528  804.9991261  141.2006049
        s[5]         s[6]         s[7]         s[8]         s[9]        s[10]
 -18.3712420  -30.9095946  -27.7327247  -24.9607594  -20.6167319  -18.2307559
       s[11]        s[12]        sigma
  -9.7510057    0.0942663  436.8221673
-2log(likelihood)= 1872.413  AIC= 1900.413   BIC= 1938.346
```

Quest'ultimo modello associa il parametro di stagionalità più alto al mese di marzo, mentre si nota che la varianza stimata da entrambi i modelli è abbastanza elevata evidenziando, come già osservato, una

certa variabilità dei dati. Verifichiamo adesso l'adattamento dei modelli mediante l'analisi dei residui.

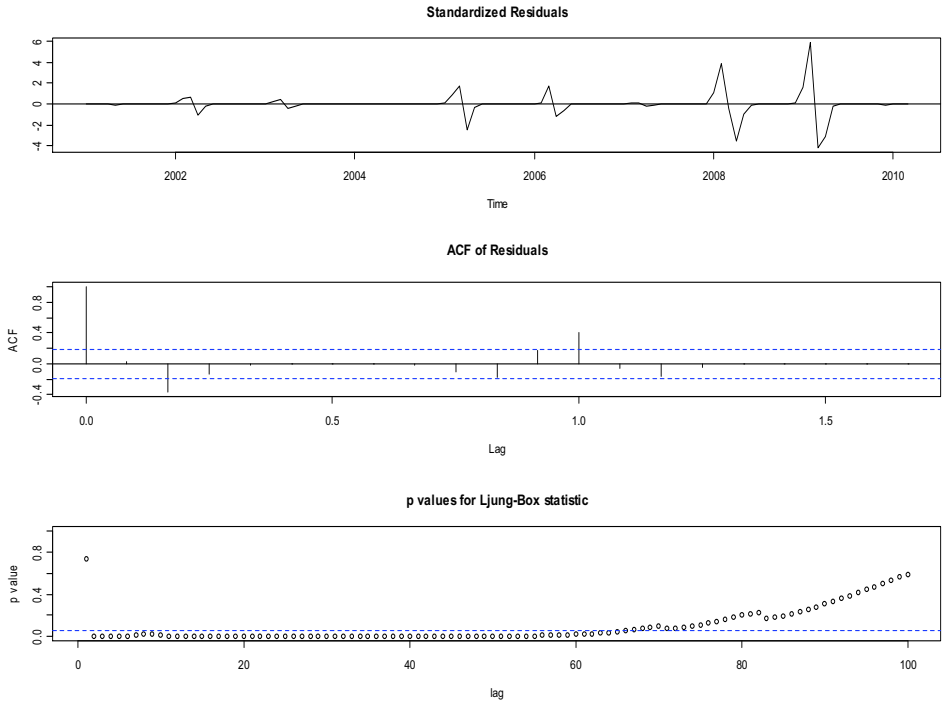

Fig. 3.45: *grafici diagnostici relativi al modello stimato secondo il criterio BIC nessuna deriva e stagionalità, innovazione additiva.*

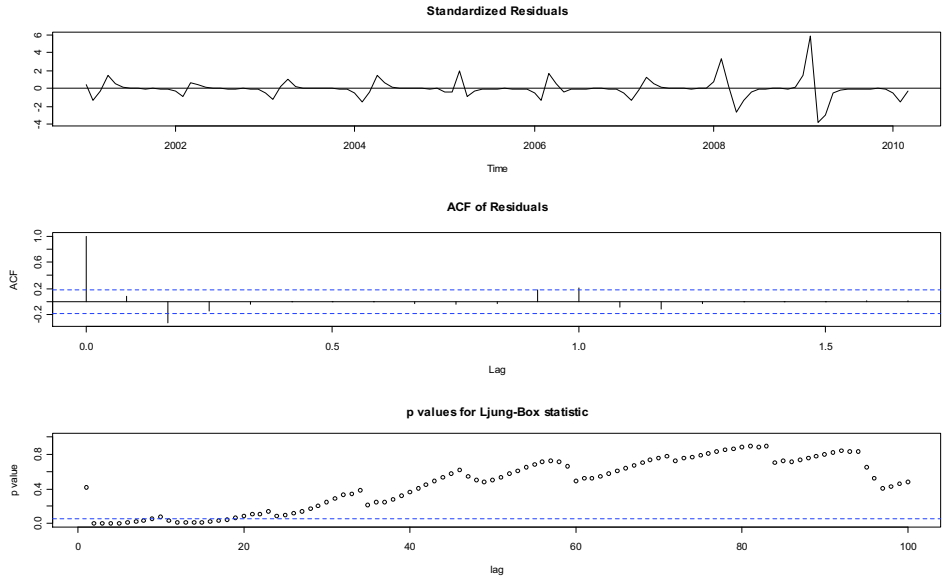

Fig. 3.46: *grafici diagnostici relativi al modello stimato secondo il criterio AIC nessuna deriva, stagionalità costante/moltiplicativa, innovazione additiva.*

Come si nota dalle Figure 3.45 e 3.46 i modelli suggeriti non sembrano cogliere molto bene la correlazione seriale delle osservazioni. Infatti i coefficienti di autocorrelazione dei residui stanno al di sotto della banda al 5%: in particolare si nota che il criterio BIC propone un modello peggiore dell'AIC. Inoltre vale la pena notare la discontinuità che si

osserva nel grafico dei residui standardizzati di entrambe le figure: questo dipende dal fatto che durante il periodo di osservazione dei dati, si alternano anni di elevata incidenza della malattia ad anni in cui l'incidenza è relativamente bassa.

L'applicazione di modelli a lisciamento esponenziale in questo caso non sembra essere la tecnica più adatta per questo tipo di dati osservati: come già detto in precedenza, essa presuppone l'ipotesi di processi stazionari che, nel caso dei dati relativi all'influenza B, non è verificata. I dati sono molto variabili durante i nove anni di osservazione, per cui, probabilmente, sarebbe più conveniente attuare una trasformazione ai dati e successivamente differenziare la serie così trasformata per renderla più stazionaria. Qualora anche in questo caso le tecniche di lisciamento esponenziale non dovessero ancora funzionare, allora si potrebbe pensare di passare all'applicazione di modelli probabilistici autoregressivi come i modelli ARIMA.

3.11 Conclusioni.

In questo capitolo sono state analizzate alcune malattie infettive più o meno comuni, i cui dati sono stati raccolti dall'Istituto Robert Koch di Berlino, che li rende disponibili tramite un database elettronico integrato, amministrato statalmente e sviluppato dall'istituto stesso, accessibile al sito internet *http://www3.rki.de/survstat*.

Tramite l'applicazione di alcune tecniche di scomposizione delle serie temporali, si sono isolate la componente stagionale e il trend, là dove queste erano presenti e, quando i modelli proposti risultavano soddisfacenti attraverso un'adeguata analisi diagnostica dei residui, si sono sovrapposte alle serie storiche originarie, le loro componenti stimate.

Le analisi dei precedenti paragrafi hanno evidenziato che la maggior parte delle malattie infettive ha caratteristiche di stagionalità, nel senso che colpiscono maggiormente in alcuni particolari mesi dell'anno, durante i quali si registrano il massimo numero di persone colpite dal virus, mentre solamente alcune presentano un trend evidente. Le tecniche di analisi hanno portato ad una buona approssimazione delle serie relative alla Salmonellosi, Epatite A, Legionellosi e E. Coli Enterite, tutte caratterizzate da andamenti stagionali, e quasi tutte prive di deriva, tranne che per la serie relativa alla Legionellosi che presenta una

deriva costante/moltiplicativa. Per le restanti malattie infettive, invece, le tecniche di analisi qui utilizzate non sono state soddisfacenti nel proporre modelli che cogliessero bene la correlazione seriale dei dati osservati, in parte perchè le ipotesi di base presuppongono serie stazionarie, ossia con media invariante nel tempo e con correlazione che dipende solo dal numero di ritardi considerati. In questi casi, invece, alcune serie presentavano una forte variabilità durante gli anni del periodo osservato, come ad esempio l'influenza B, che pur essendo un fenomeno tipicamente stagionale, non è stato colto dai modelli applicati a causa della violazione delle ipotesi di base.

Capitolo 4:
sorveglianza delle malattie infettive; algoritmo di Farrington.

4.1 La sorveglianza delle malattie infettive.

La patologia infettiva, in virtù della diffusibilità degli agenti che la determinano, costituisce, da sempre, uno dei principali problemi di sanità pubblica: ciò sia in termini sostanziali, cioè di impatto quali-quantitativo sulla salute della popolazione, sia per le ricadute sociali, soprattutto in relazione alla percezione di rischio ad esse correlato.
La sorveglianza sanitaria nei confronti delle malattie infettive e diffusive assume, dunque, una notevole importanza strategica nell'ambito del sistema sanitario: una buona sorveglianza consente sia di conoscere e, pur con certi limiti, prevedere l'andamento epidemiologico delle malattie, sia di programmare e valutare l'efficacia dei servizi addetti alla prevenzione ed al controllo del contagio.
La principale attività della sorveglianza in sanità pubblica consiste nell''utilizzazione dei dati per la prevenzione e il controllo delle malattie infettive e per il monitoraggio dei programmi di attività inteso come continua valutazione della relazione intervento-cambiamento. Essa si basa su tre caratteristiche fondamentali: raccolta sistematica dei dati, aggregazione e analisi dei dati raccolti, ritorno e diffusione delle informazioni.
Lo scopo del sistema di sorveglianza delle malattie infettive, è quello di essere in grado di fornire la maggior parte delle informazioni necessarie per la definizione delle strategie di controllo e per il monitoraggio dei programmi di intervento; in particolare esso deve garantire il raggiungimento dei seguenti obiettivi della sorveglianza delle malattie infettive:

1. seguire l'evoluzione dell'incidenza delle infezioni e delle loro conseguenze (complicanze, esiti,ecc);
2. individuare e descrivere le epidemie;
3. orientare le misure di prevenzione;
4. monitorare e valutare i programmi di prevenzione;
5. seguire i fattori di rischio (alimentare, sessuale, viaggi, iatrogeno, ecc.);
6. sorvegliare i trattamenti (TB, resistenza alle terapie).

Per garantire il raggiungimento di questi obiettivi il sistema di sorveglianza di sanità pubblica delle malattie infettive è affidata principalmente al Sistema Informativo delle Malattie Infettive (**SIMI**), basato sulle notifiche dei medici curanti, che comprende segnalazioni immediate per allertare gli operatori della sanità pubblica e riepiloghi mensili di tutte le malattie infettive notificate, compilati da ogni Azienda Sanitaria Locale.(ASL). Il SIMI è attualmente regolato dal Decreto ministeriale 15 dicembre 1990[45] e successiva modifica relativa alla tubercolosi e alla micobatteriosi (Decreto ministeriale 29 luglio 1998[46]), con il quale il Ministero della Sanità ha aggiornato e modificato l'elenco delle malattie infettive e diffusive che danno origine a particolari misure di sanità pubblica, sulla base delle esigenze di controllo epidemiologico e di integrazione del sistema informativo sanitario nazionale.

Il flusso informativo previsto si svolge attraverso il medico, ospedaliero o di base, che diagnostica la malattia infettiva e successivamente ha obbligo per legge di effettuarne la segnalazione alla ASL di competenza, le Aziende Sanitarie Locali incaricate della adozione di eventuali misure di profilassi a tutela della salute pubblica, la Regione (Agenzia di Sanità Pubblica) con azione di supervisione e coordinamento, gli Organismi Centrali (Ministero della Salute, ISTAT, Istituto Superiore di Sanità) ed eventualmente internazionali (UE, OMS).

Il SIMI stabilisce l'obbligo di notifica (definendone modalità e tempi) per 47 malattie infettive classificate in 4 classi in base alla loro rilevanza di sanità pubblica ed al loro interesse sul piano nazionale ed internazionale; prevede inoltre una quinta classe che comprende malattie non specificamente menzionate nei gruppi precedenti e le zoonosi indicate dal regolamento di Polizia Veterinaria. Secondo tale sistema le malattie infettive a obbligo di notifica sono state differenziate in base alle informazioni da raccogliere e alla tempestività di invio dei dati.

Due sono dunque le fonti di dati che consentono di disporre di informazioni su tutte le malattie: il sistema di notifica nazionale, gestito appunto attraverso il SIMI, e la registrazione delle dimissioni ospedaliere.

La notifica obbligatoria delle malattie infettive costituisce il flusso informativo alla base di tutto il sistema di sorveglianza, perché permette di definire e confrontare tra le ULSS e con le altri Regioni o i diversi

[45] http://www.salute.gov.it/imgs/C_17_normativa_1357_allegato.pdf
[46] http://www.salute.gov.it/imgs/C_17_normativa_1358_allegato.pdf

Paesi l'incidenza delle malattie infettive, mentre la valutazione delle schede di dimissione ospedaliera consente, per le malattie più gravi, una ulteriore fonte indipendente di informazione. Questa seconda fonte permette di aumentare la sensibilità del sistema attraverso l'identificazione di casi non notificati e di stimare l'incidenza vera e la efficacia dei flussi informativi ricorrendo al sistema di cattura-ricattura. La suddivisione in classi risponde anche a criteri di rilevanza epidemiologica e a esigenze differenziate di profilassi.

Esistono infine i Sistemi di Sorveglianza di Laboratorio per le diarree infettive (D.G.R. 4259 del 04/08/98), le meningiti e le altre forme invasive da batteri (D.G.R. 4260 del 04/08/98), le micobatteriosi e la legionellosi (D.G.R. 2488 del 11/05/99) che permettono una migliore accuratezza diagnostica e facilitano l'indirizzo di eventuali azioni di profilassi da intraprendere.

I dati ufficiali italiani sulla sorveglianza delle malattie infettive sono consultabili sul sito del ministero del Lavoro, Salute e Politiche sociali[47] e vengono aggiornati ogni anno. Essi si riferiscono ai dati raccolti attraverso il Simi. È possibile consultare i dati sul sito disaggregati per Regione, sesso e fasce d'età. Tuttavia si tratta di dati annuali o per lo più mensili e non a frequenza settimanale come quelli elaborati nel capitolo precedente ed estrapolati dal sito dell'Istituto Robert Koch.

4.2 Sistemi di sorveglianza univariata: alcune questioni statistiche.

Nel paragrafo precedente si è visto che la sorveglianza pubblica delle malattie infettive rappresenta uno strumento utile per identificare eventuali epidemie e valutare le possibili strategie di prevenzione, controllo e intervento. Per attuare ciò è però necessario disporre di adeguati strumenti informatici e statistici che consentano, attraverso l'applicazione di algoritmi automatici, la rilevazione di anomalie per identificare situazioni di allarme e testare rapidamente ipotesi esplicative sullo stesso.

Esistono diversi metodi statistici di rilevazione di focolai epidemici. In questo capitolo verrà data attenzione ad algoritmi di rilevazione di focolai applicati a serie temporali univariate: in particolare sarà descritto e applicato il metodo proposto da Farrington e Andrews il cui scopo principale è quello di prevedere il valore osservato al tempo attuale t_0

[47] http://www.salute.gov.it/malattieInfettive.

attraverso l'utilizzo di *valori di riferimento* presi dalla serie storica osservata.

Prima però di descrivere dettagliatamente l'algoritmo proposto da Farrington e Andrews, si illustreranno brevemente alcune principali questioni riguardanti le tecniche statistiche di rilevazione dei focolai epidemici.

Affinché un sistema di monitoraggio e rilevazione continua dei dati relativi alle malattie infettive sia efficace, è necessario che esso individui i focolai prima che essi si siano sviluppati completamente, quando ancora il numero dei casi è scarso. Inoltre deve rilevare i focolai tempestivamente, spesso prima che le relazioni siano complete o che i dati siano stati completamente validati. I principali requisiti per un sistema di scansione sistematico sono infatti la tempestività, sensibilità e specificità, insieme con risultati prontamente interpretabili. La tempestività e la sensibilità sono necessarie per garantire che i focolai siano rilevati in tempo per attuare interventi.

La normativa attuale fornisce una definizione operativa di *focolaio epidemico* come "il verificarsi di due o più casi della stessa malattia in un gruppo di persone appartenenti ad una stessa comunità (famiglia, scuola, caserma, istituto di ricovero ecc.) o comunque esposti ad una comune fonte di infezione".

Da un punto di vita metodologico, questa definizione di focolaio richiede la specificazione del calcolo di una *soglia* statistica o valore limite, al di sopra della quale stabilire che i dati osservati sono considerati anomali o "anormali", e di un metodo di decisione per valutare se l'attuale conteggio è significativamente superiore o meno a tale linea di confine. La maggior parte dei sistemi di rilevazione di un focolaio calcola dei valori di soglia in funzione di quelli ottenuti sulla base dei storici.

Quando il numero di soggetti segnalati supera la soglia, allora vengono attuate ulteriori indagini epidemiologiche, che determinano i comuni fattori epidemiologici, e se il focolaio viene confermato, si attuano le opportune misure di controllo.

In teoria, indicando con θ il valore di soglia, questo potrebbe essere scelto in modo tale da soddisfare il criterio del tipo:

$$P(X > \theta \mid \text{nessun focolaio}) = \alpha,$$

Dove X è il numero dei malati in un determinato periodo di tempo e α è il tasso dei falsi positivi, ossia la probabilità di avviare un' indagine,

quando invece non si è verificato alcun focolaio. In realtà, però è più realistico soddisfare un criterio più debole:

$$P(X > \theta \mid \text{nessuna anomalia}) = \alpha,$$

dove α è la probabilità di avviare un 'indagine, quando non esiste una vera e propria anomalia, cioè la probabilità che la soglia venga superata per caso (a seguito della variabilità casuale dei dati). La soglia θ è più comunemente definita come l'estremo superiore di un intervallo di previsione per l'osservazione corrente ed è calcolata sulla base delle osservazioni passate tranne quella corrente osservata.

4.3. Metodi statistici di rilevazione dei focolai.

Esistono principalmente due metodi di rilevazione dei focolai epidemici: quello prospettico e il metodo retrospettivo. Il primo rileva i casi nel momento in cui essi si verificano e differisce da quello retrospettivo non solo nel problemi metodologici di analisi che si pongono, ma anche nei tipi di ipotesi da testare. L'interesse si concentra sulla distribuzione nulla dei casi di malattia al momento attuale t, data la storia passata del processo prima del tempo t. Ciò richiede che debba essere specificata un adeguata distribuzione svincolata però dai focolai passati, e che sia formulata una regola di decisione in base alla quale l'attuale valore osservato y_0 sia classificato come anomalo o meno. Nel metodo prospettico, i dati vengono analizzati così come si presentano, principalmente dalla data della manifestazione del caso che corrisponde alla data dell'infezione, o se questa non è disponibile, alla data di inizio o di raccolta del campione.

In questo caso però ci possono essere dei ritardi nelle segnalazioni: infatti tra la data in cui un individuo è infetto, e quella in cui l'infezione viene notificata completa dei dettagli dell'agente infetto responsabile, si verifica inevitabilmente un ritardo. Durante il periodo di infezione infatti, c'è tutta una serie di attività che vanno dal prelievo dei vari campioni dall'individuo infetto, ad una successiva analisi per determinare l'agente infettivo. Questo processo può però richiedere diversi giorni.

Pertanto le serie temporali dei casi di malattia sono inevitabilmente soggette a distorsioni, in quanto i casi segnalati nelle ultime settimane rappresentano solo una frazione dei veri valori ed è molto probabile che le infezioni più recenti non vengano rilevate, sottostimando così i casi correnti di malattia e quelli più recenti; questo può essere un problema serio, perché sono proprio questi ultimi, quelli su cui si basa la

rilevazione del focolaio. In questo metodo di rilevazione, è dunque fondamentale avere una chiara comprensione del sistema di segnalazione e ritardi ad essa dovuti.

Con il metodo retrospettivo, invece, il problema non si presenta, in quanto la rilevazione del focolaio è effettuato con dati completi, per il quale i ritardi di segnalazione possono essere ignorati. La maggior parte delle procedure di rilevazione che usano il metodo retrospettivo è progettata per identificare i casi anomali in un intervallo di tempo (t_1, t_2) con $t_l < t_2 < t$, dove t è l'istante attuale mentre, t_l, t_2 spesso non sono ben specificati. Molti di questi metodi possono essere utilmente applicati in maniera prospettica.

4.4 Metodi di regressione.

Un modo semplice per rilevare in maniera prospettica le anomalie nelle serie storiche di dati di sorveglianza, è quello di stimare la distribuzione dei casi al tempo t, $f(y|t)$, mediante la regressione lineare dei casi osservati y_i, i= 1,2,……n sul tempo passato t_i. Per casi numerosi, si può assumere $f(y|t)$ come una normale $N(\mu,\sigma^2)$, mentre per casi poco frequenti può essere più appropriata una distribuzione di Poisson con parametro di dispersione pari a pari a μ o $\varphi\mu$. In tal caso si ottiene il valore atteso μ dei casi al tempo t dato dal valore predetto stimato dal modello di regressione, nonché il livello di soglia sempre ottenuto dall'intervallo di previsione del modello. Il vantaggio di questo metodo è che l'equazione di regressione può facilmente incorporare i coefficienti relativi al *trend* e alla componente *stagionale*, mentre ha lo svantaggio di ignorare un'eventuale autocorrelazione tra i casi. Per virus rari, la correlazione seriale non è un problema perché la maggior parte dei casi è sporadica, almeno in condizioni tali da non provocare un focolaio.

4.5 Algoritmo di Farrington.

Nei modelli di regressione lineare i trend sono sempre inseriti nella stima di una variabile temporale. Viene tenuto conto della stagionalità basando il calcolo della soglia soltanto su casi provenienti da periodi passati tra di loro confrontabili. Questo metodo è usato ufficialmente dal *Centers for Desease Control and Prevention* (CDC): il sistema non considera il *trend* temporale, e poiché i dati di sorveglianza spesso presentano caratteristiche di stagionalità, questa viene affrontata usando valori di riferimento, ossia confrontando i casi del mese corrente con quelli di altri mesi confrontabili, il mese precedente e quello successivo nei 5 anni

passati. Per esempio, il numero di casi relativi a marzo 2002 è confrontato con i casi di febbraio, marzo e aprile degli anni che vanno dal 1997 al 2001. Il valore predetto è pertanto la media dei 15 valori di base. Questi periodi tra di loro confrontabili e solitamente espressi in settimane, (*base-line weeks*) sono specificati dagli interi *b* e *w*. Se la settimana attuale è *x* dell'anno *y*, verranno considerati come valori di riferimento solo i dati appartenenti alla finestra temporale ($x - w$; $x + w$) per tutti gli anni che vanno da ($y - b$) a ($y - 1$), dando così un totale di $b(2w + 1)$ settimane prese come riferimento, *base-line weeks*, ai fini di incorporare gli effetti stagionali.

In generale, sia $y_{0:t}$ il numero dei casi della settimana corrente (indicata come settimana *t* nell'anno *0*), *b* il numero di anni con cui andare indietro nel tempo e *w* il numero di settimane in un intorno di *t* che saranno incluse negli anni precedenti. Per l'anno zero, si usa w_0 il numero delle settimane precedenti da includere nel calcolo dei valori di riferimento (*reference values*), in genere si ha $w = w_0$. Complessivamente l'insieme dei valori di riferimento è:

$$R(w,w_0,b)=\left(\bigcup_{i=1}^{b}\bigcup_{j=-w}^{w} y_{-i:t+j}\right)\bigcup\left(\bigcup_{k=-w_0}^{-1} y_{0:t+k}\right).$$

Per quanto sia auspicabile, ai fini della precisione delle stime, una elevata numerosità *n* dei casi osservati, è necessario sottolineare che l'ampiezza della finestra temporale, deve essere scelta in modo tale da rispecchiare la variabilità cronologica degli andamenti stagionali legati alla malattia sottoposta ad osservazione, mentre la scelta di *b*, cioè il numero di anni con cui andare a ritroso nel tempo, deve essere fatta in modo tale da ottenere dati tra di loro confrontabili che risentano il meno possibile dei cambiamenti nelle procedure di notifica delle malattie e nei metodi di analisi dei laboratori. I valori usati di default nell'algoritmo, sono b = 5, w = 3 per un totale di 35 valori di base. Pertanto fino a che i metodi di classificazione delle malattie infettive vengono aggiornati occasionalmente, si può dunque estendere i confronti fino a 5 anni.

4.6 Il modello di regressione stimato.

Una volta scelto i valori di riferimento (*reference values*) y_i, corrispondenti alle settimane scelte con i parametri *b* e *w* (*base-line week*), si assume che le y_i si distribuiscano con media μ_i e varianza $\varphi\mu_i$ e siano indipendenti tra di loro. L'unica componente sistematica lineare che viene inclusa nel modello è il trend temporale che caratterizza la

frequenza delle notifiche. Il modello principale è descritto da un modello di regressione log-lineare di quasi verosimiglianza di Poisson:

$$\log \mu_i = \alpha + \beta_{ti} \quad E(Y_i) = \mu_i \quad V(Y_i) = \varphi\mu_i$$

con α, β, e $\varphi > 0$, t_i è misurato in settimane, solitamente prendendo i valori ad esempio 1-7, 53-59 e così via. Le stime sono ottenute con il metodo di quasi-verosimiglianza dove il parametro di dispersione φ viene così stimato:

$$\hat{\varphi} = \max\left\{\frac{1}{n-p}\sum_{i=1}^{n}\omega_i\frac{(y_i-\hat{\mu}_i)^2}{\hat{\mu}_i},1\right\},$$

con ω_i pesi che saranno descritti nel paragrafo 4.8, mentre p=1 o p=2 a seconda che il trend sia stimato a meno. Sia t_0 la settimana attuale e y_0 il corrispondente numero di casi osservato. Allora il numero atteso dei casi osservati sarà data da:

$$\hat{\mu} = \exp(\hat{\alpha} + \hat{\beta} t_0).$$

Nel modello il trend temporale viene incluso solo se la serie temporale va a ritroso nel tempo di almeno di tre anni e se, nel caso risulti significativo al 5%, sia abbia:

$$\hat{\mu}_0 \le \max\{y_0 : i = 1.....n\}.$$

Questa condizione serve ad escludere stime non realistiche del trend temporale.

4.7 Il calcolo della soglia.

Una volta adattato il modello, si calcola il valore di soglia con cui confrontare i casi osservati al tempo t_0 e stabilire pertanto se si è in presenza o meno di un focolaio. In questo caso però, bisogna distinguere le malattie infettive caratterizzate da una certa frequenza di casi da quelle poco comuni, che si manifestano raramente e con casi poco numerosi. Queste ultime infatti, presentano una distribuzione molto inclinata e asimmetrica, di cui deve essere tenuto conto nel calcolo della soglia, allo scopo di mantenere il più possibile costante la percentuale di falsi positivi sul maggior numero di valori attesi μ_0.

Un modo per tener conto di ciò, correggendo la forte inclinazione delle distribuzioni in presenza di malattie infettive caratterizzate da pochi casi, è quello di applicare una trasformazione dei dati elevandoli a 2/3 che consente di ottenere una distribuzione di Poisson approssimativamente simmetrica. Sulla base di questa trasformazione viene poi calcolato l'intervallo di fiducia al 100(1- α)% e i risultati ottenuti trasformati nuovamente alla loro scala originaria. Pertanto abbiamo:

$$E(y_0^{2/3}) = \mu_0^{2/3},$$
$$\text{var}(y_0^{2/3}) = \frac{4}{9}\phi\mu_0^{2/3}$$

e

$$\text{var}(\hat{\mu}_0^{2/3}) = \frac{4}{9}\mu_0^{-2/3}\,\text{var}(\hat{\mu}_0).$$

La varianza dell'errore di previsione su scala potenza di 2/3 è:

$$\text{var}(y_0^{2/3} - \hat{\mu}_0^{2/3}) = \frac{4}{9}\tau\mu_0^{1/3},$$

dove

$$\tau = \phi + \frac{\text{var}(\hat{\mu}_0)}{\mu_0}.$$

L'intervallo di previsione (L,U) è dato dunque da:

$$U = \hat{\mu}_0\left\{1 + \frac{2}{3}z_\alpha\left(\frac{\hat{\tau}}{\hat{\mu}_0}\right)^{1/2}\right\}^{3/2},$$
$$L = \hat{\mu}_0\max\left\{\left\{1 - \frac{2}{3}z_\alpha\left(\frac{\hat{\tau}}{\hat{\mu}_0}\right)^{1/2}\right\}^{3/2}, 0\right\},$$

dove $\hat{\mu} = \exp(\hat{\alpha} + \hat{\beta}t)$.

Questo è l'intervallo approssimato di previsione di y_0 al 100(1- α)%, dove z_α è il 100(1- α) percentile di una distribuzione normale. I casi di malattia che superano il livello superiore dell'intervallo che definisce la soglia sono considerati un possibile focolaio. Tutte le componenti del calcolo possono essere ottenute mediante il risultato finale di una regressione. Supponiamo ad esempio che l'intervallo di previsione per il valore corrente sia stato ottenuto dai precedenti casi rilevati con frequenza settimanale: 3, 2, 5, 8 ,6, 7, 5 e 9 nelle settimane 0 - 7 (la settimana corrente è la 8). Adattando un modello di regressione lineare

generalizzata con distribuzione di Poisson si ottengono le stime seguenti[48]: $\hat{\alpha} = 1.222, \hat{\beta} = 0.1315, \text{var}(\hat{\alpha}) = 0.100094, \text{var}(\hat{\beta}) = -0.018654$, da ciò segue: $\hat{\mu}_8 = \exp(\hat{\alpha} + 8\hat{\beta}) = 9.718$ e $\text{var}(\hat{\mu}_8) = 0.0875$. Supponendo $\tau = 1$, il limite superiore di previsione al 95% sarà:

$$9.718 \times \left(1 + \frac{2}{3} \times 1.96 \times \sqrt{\frac{1}{9.718} + 0.0088}\right)^{3/2} = 19.1 .$$

Tutti i casi maggiori di 19.1 saranno considerati anormali, con un tasso nominale di falsi positivi pari a 2.5%.

Come già detto, la trasformazione dei dati alla potenza di 2/3 è opportuna solo per casi poco frequenti per i quali si ipotizza una distribuzione di Poisson[49], mentre per casi di malattia più frequenti non si attua tale trasformazione e si può ipotizzare una distribuzione normale.

4.8 L'influenza dei passati focolai.

Una delle maggiori difficoltà che si deve affrontare quando si adatta un modello di regressione lineare (usando la distribuzione normale o di Poisson), è come riuscire a ridurre l'influenza dei focolai passati: non tener conto di questi ed includerli nel calcolo della soglia, porta ad una sovrastima della stessa e ad una corrispondente riduzione della sensibilità[50]. In termini statistici, i focolai passati corrispondono ad

---

[48]
```
y=c(3,2,5,8,6,7,5,9)
t=c(0,1,2,3,4,5,6,7)
e=glm(y~t,"poisson")
summary(e)

Call:
glm(formula = y ~ t, family = "poisson")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0490  -0.4042   0.1346   0.1985   1.2157

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.22199    0.31671   3.858 0.000114 ***
t            0.13150    0.06688   1.966 0.049290 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 7.6681  on 7  degrees of freedom
Residual deviance: 3.6943  on 6  degrees of freedom
AIC: 35.718

Number of Fisher Scoring iterations: 4
predict(e,newdata=data.frame(t=8),type="response")
9.718194
vcov(e)
            (Intercept)            t
(Intercept)  0.10030333 -0.018689639
t           -0.01868964  0.004473586
```

[49] o quasi-poisson qualora $\phi \neq 1$.

[50] La sensibilità è la capacità di identificare correttamente i casi che superano la soglia come focolai, ossia che l'ipotesi funzioni correttamente.

outliers della serie temporale e il modo più semplice per correggere i dati da questi valori anomali è quello di ometterli nel calcolo della soglia. Alternativamente, i focolai passati possono essere ponderati utilizzando degli appositi *pesi*.

L'algoritmo di Farrington contiene, a tal proposito, una serie di ulteriori approfondimenti ai fini di una migliore previsione di y_{t0}, ad esempio correggendo i valori presi a riferimento, i *reference values*, dai passati focolai, verificando l'ipotesi sulla significatività del trend, rappresentato dal coefficiente β del predittore lineare e correggendo l'asimmetria della distribuzione usata per effettuare le previsioni

Utilizzando le iniziali stime di μ_i e del parametro di dispersione ϕ, ottenuto con $\omega_i = 1$, i residui sono così definiti:

$$s_i = \frac{3}{2\sqrt{\hat{\phi}}} \frac{y_i^{2/3} - \hat{\mu}_i^{2/3}}{\hat{\mu}_i^{1/6}\sqrt{(1-h_{ii})}},$$

dove h_{ii} sono gli elementi della diagonale principale della matrice di varianze e covarianze. Nel caso di dati che si distribuiscono come una Poisson e per i quali $\phi = 1$, s_i sono i residui standardizzati di Ascombe. I pesi ω_i sono così definiti:

$$\omega_i = \begin{cases} \gamma s_i^{-1} & \text{se } s_i < 1 \\ \gamma & \text{altrimenti} \end{cases},$$

dove γ è una costante tale che $\Sigma\, \omega_i = n$.

Con questa funzione di ponderazione viene dunque applicato un filtro lineare allo scopo di ridurre l'influenza dei focolai passati senza eliminarli del tutto; in particolare ai valori di base, quelli cioè presi a riferimento, che presentano i residui più alti, si assegnano i pesi ω_i più bassi e viceversa. Inoltre i valori di soglia così calcolati non risentono della ponderazione anche quando nei "reference value" non sono presenti passati focolai e questo a beneficio della *specificità*[51] delle stime.

4.9. L'ambiente statistico di R per la rilevazione di focolai epidemici.

Si è detto che lo scopo dell'algoritmo per la sorveglianza dei dati relativi alle malattie infettive, è quello di rilevare tempestivamente delle anomalie nei dati. Oltre a dover disporre di un sistema routinario di monitoraggio continuo delle malattie infettive in grado di fornire in

[51] La capacità di dare il meno possibile falsi allarmi.

tempo reale i dati necessari per attuare attività di controllo ed intervento, un buon sistema di sorveglianza, necessita di ambienti informatici in grado di gestire in maniera efficiente una grande mole di dati in tempi brevi e con il minor numero di errori possibile. Ciò richiede l'utilizzo di software appropriati attraverso i quali implementare algoritmi che consentano di elaborare e segnalare eventuali anomalie nei dati che possono essere considerati dei focolai. L'ambiente statistico R offre a tal proposito la possibilità di applicare diversi algoritmi per la rilevazione di focolai epidemici attraverso l'utilizzo del pacchetto *surveillance.* Si tratta di un pacchetto contente l'applicazione di metodi statistici per la rilevazione modellazione e simulazione di serie storiche relative alle malattie infettive. Esso è principalmente adatto per la rilevazione di focolai epidemici relativi a serie temporali di dati provenienti soprattutto da sistemi di sorveglianza pubblica delle malattie infettive, ma anche da altri ambienti e scienze sociali.

In particolare, il pacchetto *surveillance* consente di monitorare ed elaborare dati relativi alle malattie infettive mediante l'applicazione di diversi algoritmi che si basano sia metodo prospettico che quello retrospettivo[52]. Gli algoritmi presenti sono:

1) per la rilevazione prospettica di focolai per serie temporali univariate:
    a. *cdc* – Stroup et al. (1989)
    b. *farrington* – Farrington et al. (1996)
    c. *rogerson* – Rogerson and Yamada (2004)
    d. *cusum* – Rossi et al. (1999) and extensions
    e. *glrnb* – H. and Paul (2008)

2) per la rilevazione retrospettiva:
    a. *hhh* – Held et al. (2005), Paul et al.(2006)
    b. *twins* – Held et al. (2005)

4.10 La struttura dei dati.

Si indica con $\{y_t \; ; \; t = 1,\ldots\ldots n\}$ la serie temporale dei casi di malattia. Siccome però i dati vengono raccolti o mensilmente o su base settimanale, viene usata una notazione alternativa che tiene conto dell'unità temporale, ossia $\{y_{i:j}\}$ con $j \in \{1,\ldots\ldots\ldots,12\}$ o $\in \{1,\ldots\ldots\ldots,52\}$ dell'anno $j \in \{-b,\ldots\ldots,-1,0\}$. Con quest'ultima notazione, gli anni sono indicizzati in maniera tale che l'anno più recente, o attuale, sia quello

[52] si veda paragrafo 4.3.

con indice zero. Inoltre, se la disponibilità dei dati lo permette, l'algoritmo consente di evidenziare in corrispondenza di ciascuna settimana o mese, se ci sono stati valori tali da essere considerati un focolaio.

Pertanto, nel pacchetto *surveillance* la serie temporale viene indicata con due vettori:

$\{y_t ; x_t; t = 1,\ldots\ldots n\}$ appartenenti alla classe di oggetti *disProg* (diseas progress) che consiste in una lista contenente due vettori: il numero dei casi osservati, e un vettore booleano di stato che indica se in quel determinato mese o settimana c'è stato un focolaio[53].

Il pacchetto consente anche di fare simulazioni di dati in base ai vari algoritmi.

Le serie temporali del pacchetto *surveillance* $\{y_{ij} ; i = 1,\ldots\ldots n, j = 1,\ldots\ldots..,m\}$ sono rappresentate mediante una classe di oggetti di tipo *disProg* con i seguenti argomenti:

`obdserved`	matrice *n* x *m* che rappresenta i casi y_{ij};
`start`	vettore di lunghezza 2 contenente l'origine della serie temporale, es.: c(anno; settimana);
`Freq`	indica l'unità temporale di rilevazione dei dati: ad es. 52 se i dati sono raccolti settimanalmente, 12 se rilevati mensilmente e così via;
`alarm`	matrice booleana n x m contenente i risultati dell'algoritmo applicato alla serie temporale;
`upperbound`	matrice *m* x *n* contenente il numero dei casi che risultano segnalati come allarme (la sua specifica interpretazione dipende dal tipo di algoritmo applicato);
`control`	elenco degli argomenti di controllo usati dall'algoritmo.

Per importare i dati in R, ad esempio da un in formato *excel* o *txt*, si possono usare i classici comandi *read.csv* o *read.able* e successivamente trasformali in un oggetto sts a partire dalle osservazioni organizzate in forma matriciale.

Supponendo che le osservazioni raccolte con frequenza mensile di una certa malattia infettiva, ad esempio l'influenza, siano disponibili in formato excel per il periodo che val dal 1' gennaio 2007 al 1' maggio

[53] è un vettore di lunghezza n, dove il valore 1 indica che nella settimana *j* il numero dei casi registrati pari a $y_{t=j}$ rappresenta un focolaio, 0 altrimenti.

2009 , chiamando con *influ* il file che raccoglie le informazioni, i comandi da eseguire saranno:

```
R> influ.casi =  as.matrix(read.csv2("influ"))
R> influ=create.disProg(week=1:nrow(influ),observed=influ,
+      state=matrix(0,nrow(influ),ncol(influ)),start=c(2001,1),freq=52)
```

Detto questo, tutti i grafici e le applicazioni dell'algoritmo del pacchetto *surveillance* operano con la classe di oggetti *disProg*. Per visualizzare il grafico di un oggetto *disProg*, il pacchetto *surveillance* usa la funzione `plot` che consente alcune rappresentazioni grafiche ad esempio:

```
R>     plot(influ)
```

Una volta ottenuto i dati e fissato la lista di valori facenti parte dell'argomento dell'algoritmo, questo seguirà i seguenti passi:

1. stima un primo modello iniziale ottenendo le stime della media $\hat{\mu}_i$ e del parametro di sovradispersione $\hat{\phi}$;
2. calcola i pesi ω_i per la correzione dei focolai passati;
3. ristima il modello;
4. ottiene una nuova stima di ϕ;
5. riscala il modello applicando la trasformazione di potenza pari a 2/3;
6. elimina il trend se non è significativo;
7. ripete l'intera procedura;
8. calcola il valore di soglia;
9. elenca i valori che eccedono la soglia, se vengono analizzati più virus, altrimenti sarà evidenziato un solo valore.

Questo è dato da:

$$X = \frac{y_0 - \hat{\mu}_0}{U - \hat{\mu}_0}.$$

Questo valore viene posto uguale a zero se nelle ultime 4 settimane sono stati ricevute meno di 5 notifiche: questo perché si vuole ridurre la probabilità di segnalare come allarmi casi sporadici che si riferiscono a malattie infettive poco frequenti (virus rari). Se X > 1, allora lo si considera un allarme di focolaio e quindi sottoposto ad un maggior approfondimento.

4.11 Analisi di casi studio tramite l'applicazione dell'algoritmo di Farrington.

Nel paragrafo precedente si è descritto l'algoritmo di Farrington e i singoli passaggi da esso effettuati nel rilevare un focolaio epidemico. In questo paragrafo si applicherà l'algoritmo ad alcune serie storiche analizzate nel capitolo precedente. In particolare, si analizzerà la serie storica relativa all'influenza B.

Nel paragrafo 4.7 del capitolo , si è visto che la serie presentava un andamento stagionale con picchi raggiunti principalmente nei primi mesi dell'anno durante la stagione invernale, anche se nel corso degli anni considerati, si osservava una certa variabilità come evidenziato in Figura 4.1:

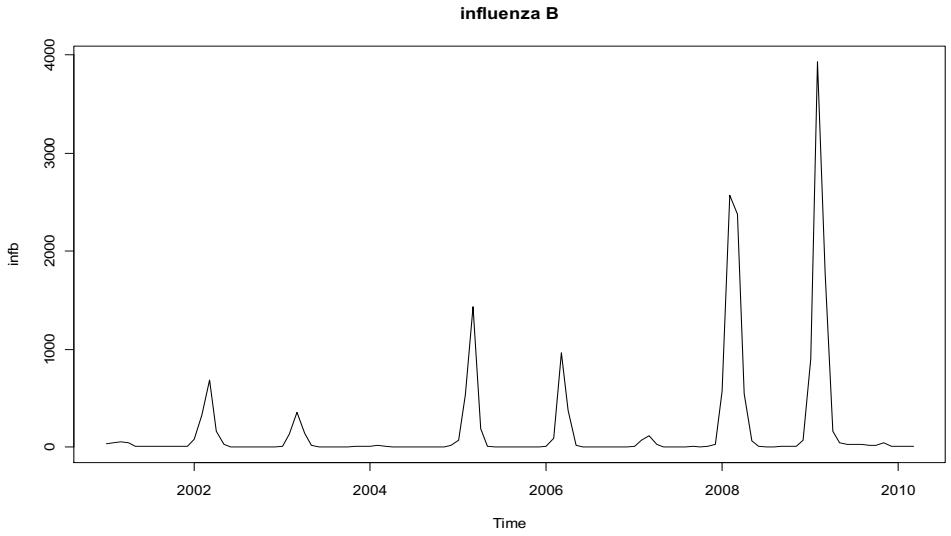

Fig. 4.1: *serie storica relativa all'Influenza B.*

In base ai dati, tentiamo di monitorare l'andamento della serie a partire dalle ultime settimane dell'anno 2008 in particolare, a partire dalla 47' settimana del 2008 quando ancora il virus da influenza B non ha ancora assunto le caratteristiche di epidemia.

L'applicazione dell'algoritmo tenterà dunque di rilevare delle anomalie nella distribuzione dei dati e segnalare degli allarmi in caso di rischio di epidemia. Per la sua implementazione in R useremo la funzione *algo.farrington*(·) che come primo argomento usa i dati strutturati come oggetto di classe *disProg*[54], mentre come secondo argomento contiene una lista di opzioni specifiche per l'algoritmo: in particolare l'intervallo temporale che viene monitorato, un α pari 0.01 per fissare il limite superiore che definisce la soglia, *b* è il numero di anni con cui andare

[54] Si veda paragrafo 4.10.

indietro nel tempo nell'uso dei valori osservati della serie storica per il calcolo della soglia, e w = 2 rappresenta la finestra temporale espressa in settimane centrata nel valore che si vuole prevedere y_0, per generare così i valori di riferimento necessari per calcolare la soglia. Come detto nel paragrafo 4.5, per l'istante temporale $t_0=(t_0^{sett}, t_0^{anno})$, l'algoritmo di Farrington attua una previsione del numero osservato dei casi $y_{t0.}$ Nell'esempio che stiamo considerando, poiché si parte a monitorare i dati a partire dalla 47' settimana del 2008, ponendo w = 2, si considerano le due settimane antecedenti e seguenti alla settimana t_0^{sett} = 47, ossia vengono presi i valori osservati corrispondenti alle settimane 45-49, compresi quelli della settimana stessa. Questo viene poi fatto per ciascuno degli anni passati della serie storica: ossia l'algoritmo prenderà come valori di confronto (*reference values*) tutti quelli corrispondenti alle settimane 45-49 degli anni precedenti al 2008: $t_0^{anno-1},.....,t_0^{anno-b}$, cioè in questo caso, fino al 2001. Pertanto l'algoritmo estrarrà un totale di $b \cdot (2w + 1) = 7 \cdot (2 \cdot 2 + 1) = 35$ *reference values*.

A questo punto, l'algoritmo di Farrington procede secondo la sequenza spiegata nel paragrafo. 5, fino ad arrivare al calcolo della soglia data dal limite superiore dell'intervallo di previsione per y_{t0} al livello di confidenza pari a (1 – α)100% sulla base di un modello lineare generalizzato di quasi verosimiglianza di Poisson[55] con funzione legame log lineare[56]. Questa operazione sarà ripetuta per tutti i casi osservati dalla 47' settimana del 2008, ossia per le 82 osservazioni della serie storica che arrivano sino ad aprile 2010, calcolando così la soglia per il range di osservazioni scelte.

Il modello così stimato è il seguente:

$$E(Y_i) = \mu_i \text{ dove } \log \mu_i = \alpha + \beta_{ti} \quad \text{e} \quad V(Y_i) = \varphi\mu_{i.}$$

L'intervallo di previsione che con il limite superiore definisce la soglia, è il seguente:

[55] E' questo un metodo di stima applicabile *senza effettuare assunzioni distributive*. La funzione di quasi verosimiglianza viene utilizzata per produrre stimatori con caratteristiche simili a quelle degli stimatori di massima verosimiglianza e funzioni test simili a quelle ottenute tramite il rapporto di verosimiglianze. Le assunzioni alla base di questo metodo non riguardano l'intera distribuzione delle osservazioni, ma piuttosto i suoi momenti secondi.

[56] Il metodo della quasi–verosimiglianza consente di affrontare i *problemi di sovradispersione.* Si parla di *sovradispersione* (*overdispersion*) quando la varianza della variabile Y è maggiore di quella teorica: si può infatti specificare var(Y_i) in modo tale da consentire una maggiore variabilità rispetto a quella imposta dalla famiglia esponenziale di riferimento.

$$u_{t0} = \hat{\mu}_{t0} + z_{1-\alpha/2} \cdot \sqrt{Var(y_{t0} - \hat{\mu}_{t0})} = \hat{\mu}_{t0} \cdot \left(1 + z_{1-\alpha/2} \cdot \sqrt{\frac{\hat{\varphi}\hat{\mu}_{t0} + Var(\hat{\mu}_{t0})}{\hat{\mu}_{t0}^2}}\right),$$

dove $z_{1-\alpha/2}$ è il quantile 1-α/2 di una distribuzione normale standardizzata, mentre $\hat{\mu}_{t0}$, $\hat{\phi}$ e $Var(\hat{\mu}_{t0})$ sono i parametri stimati dal modello di quasi verosimiglianza di Poisson. A questo punto, se il valore osservato y_{t0} è maggiore di u_{t0}, allora l'istante temporale è segnalato come un allarme. Usando una funzione indicatrice per ciascun istante temporale t_0, abbiamo:

$$\hat{x}(t_0) = I\left(\frac{y_0}{u_0} > 1\right),$$

dove $I(\cdot)$ è una funzione indicatrice che segnala l'allarme se il rapporto tra il dato osservato e quello stimato secondo il limite superiore definito dalla soglia è maggiore di uno.

Prima di procedere all'applicazione dell'algoritmo descritto nel paragrafo 4.5, il pacchetto surveillance contiene altre funzioni che consentono ad esempio di variare la trasformazione della distribuzione mediante la potenza a 1/2 o 2/3 in caso di serie storica caratterizzata da casi poco numerosi.

Applichiamo dunque l'algoritmo a partire dalla 47' settimana e a tutte quelle successive fino alla fine della serie storica, ottenendo il grafico mostrato in Figura 4.2

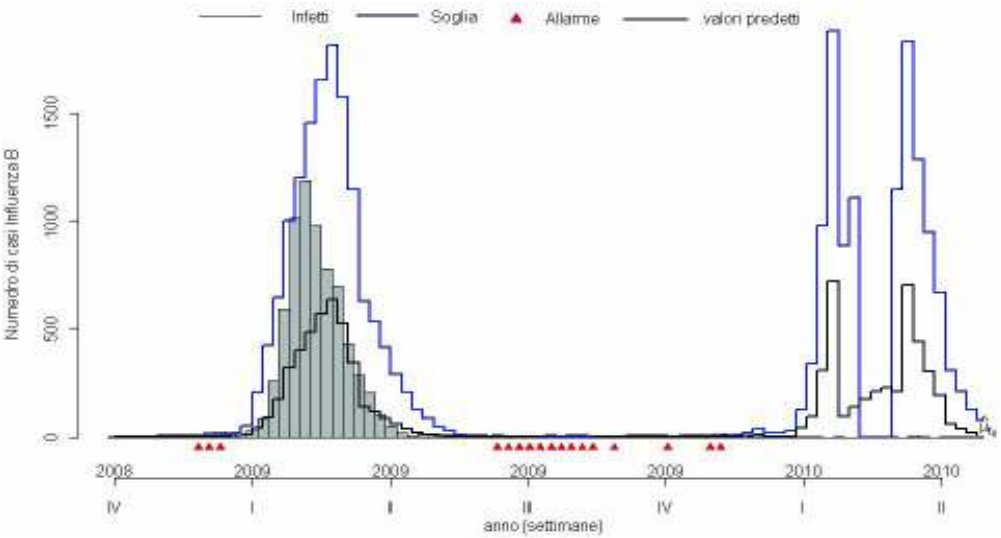

Fig. 4.2: *rilevazione dei focolai epidemici con il metodo di Farrington. La linea blu indica la soglia, la linea nera indica il numero di casi attesi previsti dal modello per ciascun istante temporale t_0: $\hat{\mu}_{t0}$. I triangoli indicano un allarme*

Nella Figura 4.2 per ciascun istante temporale monitorato, l'algoritmo, ha calcolato il livello di soglia per il numero dei casi osservati, utilizzando i valori passati della serie storica. Si è scelto un α = 0.01 per il calcolo

del limite superiore della distribuzione di previsione: ciò significa che se in un determinato istante temporale (in una determinata settimana in questo caso) si osserva che $y_{t0} > u_{t0}$, allora la probabilità che ciò sia dovuto al caso e quindi che si tratti di un falso allarme, è pari all'1%. Sulla base dei dati della serie storica e osservando la Figura 4.2, si nota che il vero valore osservato del numero dei casi infetti nella 48' settimana del 2008 è pari a 3, che, se confrontato con il limite superiore calcolato dall'algoritmo: u_{t0}=6.2, è inferiore alla soglia, per cui l'algoritmo non segnala alcun allarme. Se invece si passa all'istante temporale immediatamente successivo, ossia alla 49' settimana, il numero dei casi osservati, che è pari a 8, risulta superiore alla soglia calcolata, che è pari a u_{t0}=7.7, pertanto l'algoritmo segnala un allarme indicato dal triangolo rosso della figura 4.2. Lo stesso vale per i successivi due istanti temporali, con valori osservati rispettivamente pari a 22 e 23 e con soglie calcolate pari a 8.4 e 9.1. L'allarme viene segnalato in corrispondenza delle ultime tre settimane del 2008, proprio qualche settimana prima del mese di gennaio quando poi si incomincia a registrare un aumento considerevole dei casi fino a raggiungere il picco nella prima settimana di marzo. In questo caso lo scopo dell'algoritmo è stato raggiunto, in quanto l'allarme viene segnalato con anticipo permettendo così di essere verificato dalle autorità pubbliche sanitarie. Notare che la Figura 4.2 non segnala alcun focolaio in quanto risulta che il rapporto tra casi osservati e il livello di soglia calcolato è sempre inferiore a 1 nel range di istanti temporali monitorati. Notare inoltre che durante tutto il periodo di osservazione, l'algoritmo non segnala alcun focolaio.

Altri allarmi vengono poi segnalati dall'algoritmo verso la seconda metà dell'anno 2009, durante i mesi estivi, mentre le due ultime segnalazioni si registrano nel mese di dicembre 2009. Tuttavia, come si legge dai dati riportati in Figura 4.2, nel 2010 si osserva una notevole diminuzione del numero dei casi della malattia considerata rispetto agli anni precedenti, presentando un andamento completamente opposto rispetto alle stagioni passate caratterizzate da picchi stagionali della malattia. Ciò potrebbe essere la conseguenza di azioni preventive e di controllo attraverso vaccini o altri interventi che si sono dimostrati efficaci nel ridurre l'incidenza della malattia. L'allarme viene comunque segnalato in corrispondenza dei periodi tipici di maggior incidenza dell'influenza B, spetterà poi alle autorità valutare la possibilità di attuare interventi preventivi e/o curativi.

Per questo motivo, proviamo allora ad applicare nuovamente l'algoritmo mantenendo gli stessi controlli usati per costruire la Figura 4.2 e aggiungendo come nuovo argomento, il limite dato dal comando `limit54=c(cases,weeks)` usato per ridurre la possibilità che casi poco frequenti avvenuti nelle ultime 4 settimane siano segnalati come allarmi. In questo caso abbiamo messo come valori: *limit54=c(0,4)*. L'algoritmo infatti inserisce questo criterio proprio per tenere conto di quelle malattie che sono caratterizzate da pochi casi e per questo evitare che vengano rilevati falsi allarmi.

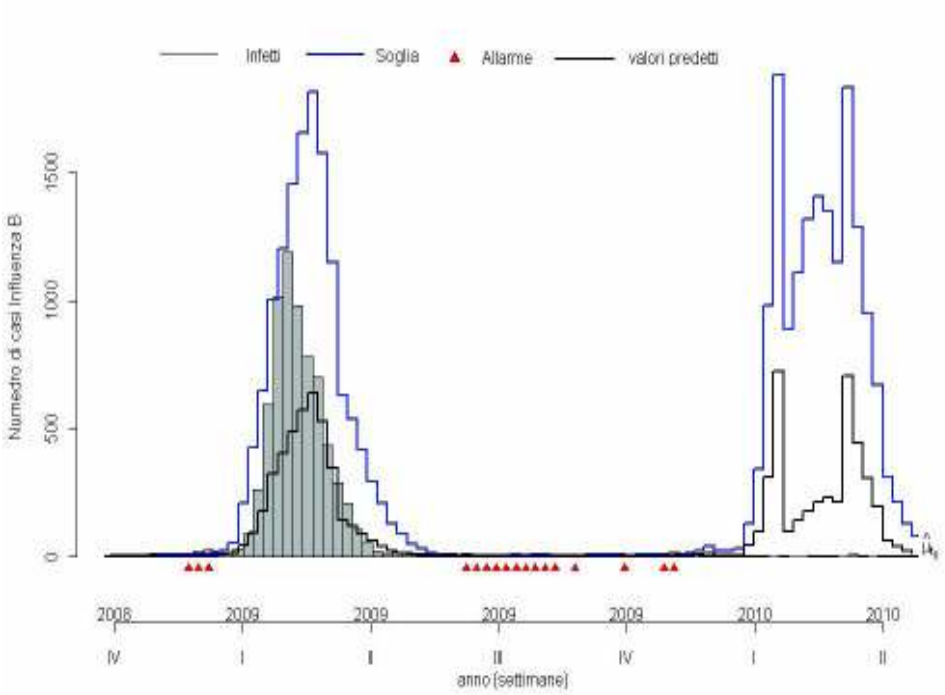

Fig. 4.3: *rilevazione dei focolai epidemici con il metodo di Farrington tenendo conto del basso numero di casi registrati nelle ultime 4 settimane.*

La Figura 4.3 calcola la soglia considerando il caso di in cui nelle ultime 4 settimane siano stati notificati casi poco frequenti. Come si osserva, la soglia è più elevata rispetto alla Figura 4.2, anche se la segnalazione degli allarmi rimane immutata.

Infine vediamo come cambia il livello di soglia calcolato dall'algoritmo, a seconda che vengano effettuate trasformazioni di 1/2, 2/3 o nessuna trasformazione sulla distribuzione. Ne paragrafo 4.7 si è detto che per casi di malattia poco frequenti, è opportuno attuare una trasformazione per rendere più simmetrica la distribuzione di Poisson; ad esempio la trasformazione alla potenza di 2/3 fa si che la probabilità dei falsi positivi rimanga costante per un ampio intervallo di valori.

Implementiamo allora l'algoritmo di Farrington sullo stesso intervallo di istanti temporali corrispondenti alle settimane monitorate, nonché gli stessi parametri impiegati inizialmente, ma considerando le tre diverse

opzioni: nessuna trasformazione della distribuzione; trasformazione alla potenza di 2/3; trasformazione alla potenza di ½.[57]. I risultati sono illustrati nella Figura 4.4.

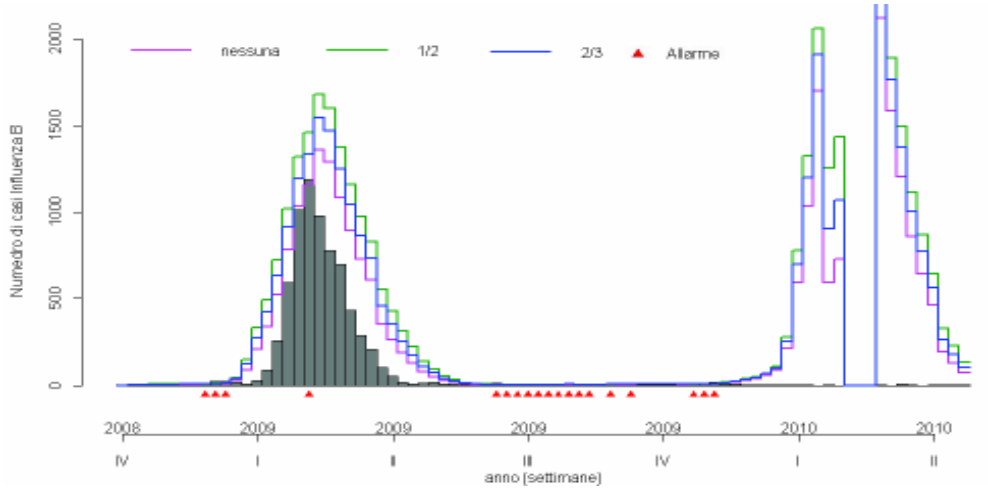

Fig. 4.4: *rilevazione dei focolai epidemici con il metodo di Farrington tenendo conto delle diverse tra formazioni alla distribuzione di Poisson.*

La Figura 4.4, illustra le tre diverse soglie calcolate dall'algoritmo a seconda che siano state applicate o meno le trasformazioni alla distribuzione di Poisson: come si nota, i valori più alti della soglia sono quelli associati al calcolo effettuato applicando la trasformazione di potenza pari ad 1/2 che, conseguentemente, è quella che segnala meno allarmi rispetto a quella pari a 2/3 di potenza e alla soglia calcolata senza alcuna trasformazione. Per queste ultime due non si rileva molta differenza tra la soglia che segnala più allarmi e quella calcolata con la distribuzione di Poisson senza trasformazione. I risultati sono illustrati in Tabella 4.1.

[57] La trasformazione che stabilizza la varianza della distribuzione di Poisson.

nessuna trasformazione settimana n°:	trasformazione pari ad ½ settimana n°:	trasformazione pari a 2/3 settimana n°:
413	-	413
414	414	414
415	415	415
423	441	441
441	442	442
442	443	443
443	444	444
444	445	445
445	446	446
446	447	447
447	448	448
448	-	-
449	449	449
450	450	450
452	-	452
454	-	454
460	-	-
461	461	461
462	462	462

Tab. 4.1: *numero delle settimane in corrispondenza degli allarmi segnalati dall'algoritmo di Farrington in base alle diverse trasformazioni.*

La Tabella 4.1 evidenzia il numero delle settimane relative al periodo monitorato, durante le quali è stato segnalato l'allarme, ovvero quando il valore osservato ha superato il livello della soglia calcolato: come si vede vengono coinvolti gli stessi periodi anche se il numero delle settimane si riduce se si attua una trasformazione della variabile.

Dopo aver applicato l'algoritmo di Farrington alla serie storica dell'influenza di tipo B, c'è tuttavia da ricordare che, come osservato nel precedente capitolo. al paragrafo 3.10, la serie storica osservata, non risultava stazionaria in quanto presentava una certa variabilità. Questo purtroppo può compromettere la buona riuscita dell'algoritmo in quanto si potrebbero produrre molti falsi negativi; ossia non segnalare allarmi quando in realtà dovrebbero essere segnalati, o falsi positivi; ossia segnalare allarmi quando in realtà non devono essere considerati tali.

Capitolo 5: *conclusioni*.

Tra i requisiti più importanti di un programma di controllo delle infezioni acquisite nelle organizzazioni sanitarie vi è la capacità di identificazione tempestiva e corretta gestione degli eventi epidemici. Le epidemie sono infatti eventi rari, ma attesi, la cui frequenza varia, a seconda del tipo di malattia. Se le epidemie vengono identificate tempestivamente, vengono rapidamente adottate appropriate misure di controllo e identificate fonti e meccanismi di trasmissione, è possibile ridurne in modo significativo l'impatto, e soprattutto è possibile modificare eventuali pratiche non corrette che possono averne condizionato l'insorgenza.

Da qui l'importanza di disporre di utili strumenti di rilevazione ed elaborazione dei dati relativi alle malattie infettive in grado di coadiuvare le politiche di intervento al fine di renderle efficaci nei loro obiettivi.

Il lavoro presentato in questa tesi, fornisce nella prima parte, una descrizione delle malattie infettive e una loro successiva analisi mediante la descrizione di alcuni principali modelli che ne descrivono il meccanismo di trasmissione tra gli individui, nell'ipotesi di popolazioni chiuse. Nella seconda parte, dopo aver analizzato diverse serie temporali relative ad alcune principali malattie infettive, si è illustrato il sistema di sorveglianza proposto da Farrington mediante l'applicazione di un algoritmo in grado di segnalare un allarme quando si è in presenza di un epidemia. Si è poi applicato l'algoritmo ad un caso di malattia infettiva molto comune, che è l'influenza B, i cui dati sono stati ricavati dall'Istituto Robert Koch di Berlino (RKI), che è l'istituto ufficiale preposto alla raccolta dei casi di malattia obbligatoriamente notificabili alle autorità sanitarie.

Bibliografa

Abbey H. *An examination of the Reed Frost theory of epidemics*. Human Biology, 24:201-233, 1952.

Alfred Clark, Jr. *S-I-R Model of Epidemics Part 1 Basic Model and Examples Revised September 22, 2005,* disponibile su http://www.me.rochester.edu/courses/ME406

Bailey N.T.J., *Some Problems in the Statistical Analysis of Epidemic Data,*Journal of the Royal Statistical Society. Series B (Methodological), Vol. 17, No. 1, pp. 35-68, 1955.

Bailey N.T.J., *The Mathematical Theory Of Epidemics*. London, Charles Griffin & Co., 1957

Bailey, N.T.J., *A simple stochastic epidemic*, Biometrika, 37, pp. 193-202. (1950).

Bartlett, M.S., *Deterministic and stochastic models for recurrent epidemics*, Proc. 3rd Berkeley Symposium Mathematical Statistics & Probability, 4, pp. 81-109 (Berkeley, University of California, (1956).

Bernoulli, Daniel, *Essai d'une nouvelle analyse de la mortalité causée par la petite vérole et des avantages de l'inoculation pour la prévenir*, in "Histoire et Mémoires de l'Académie des Sciences", 2, 1766, pp. 1-79 (mem. presentata nel 1760).

Brownlee, J., *Statistical studies in immunity: The theory of an epidemic*. Proceed-ings Royal Society Edinburgh, 26:484-521. 7, (1906).

Daryl J. Daley, Joe Gani, Joseph Mark Gani, *Epidemic Modelling: An Introduction,* Cambridge Studies in Mathematical Biology, Cambridge University Press, 1999.

De Pierro M., *La propagazione di una malattia*, da Appunti del corso di Informatica Applicata anno 2007 disponibile nel sito: http://www.di.unito.it/~depierro/educational/materiale.html

Evans, A.S., *Viral Infections of Humans* 2nd edn., Plenum Medical Book Company, New York. (1982)

Evans, G.H., *Some arithmetical considerations of the progress of epidemics*, Transactions of the Epidemiological Society, London, 1873-5, p.551, (1875)

Farrington, C., Andrews, N., Monitoring the health of populations. Oxford University Press, Ch. 8 *Outbreak Detection: Application to Infectious Disease Surveillance*, pp. 203–231,2003.

Farrington, C.P., Andrews, N.J, Beale A.D. and Catchpole M.A., *A statistical algorithm for the early detection of outbreaks of infectious disease,* J. R. Statist. Soc. A, 159, 547-563, (1996)..

Frost, W. *Some conceptions of epidemics in general*. American Journal of Epidemiology 103, (1976), pp.141–151.

Gillespie DT, *A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions.*Journal of Computational Physics, 22, 403- 434, (1976).

Giuseppe Gaeta, *Modelli Matematici in Biologia,* Springer Milano, 2007

Greenwood M., *On the Statistical measure of infectionsness*. J.Hyg. Camb., 31, 336-351. 1931

Hammer W.H. *Epidemic desease in England*. Lancet, 1, 733-739, 1906

Herbert W. Hethcote, *The Mathematics of Infectious Diseases*, SIAM Review, Volume 42, Pages: 599 - 653, Issue 4 (December 2000).

Herbert W. Hethcote, *Three Basic Epidemiological Models*, Biomathematics,Vol. 18, Springer-Verlag Berlin Heidelberg 1989, 131-132

Hohle M., *Aberration Detection in R Illustrated by Danish Mortality Monitoring,* 2010, disponibile nel sito: http://www.stat.uni-muenchen.de/~hoehle.

Hohle M*., Generating outbreak signals with R,*Department of Statistics, Ludwig-Maximilians-Universitat Munchen, Germany, 2009, disponibile sul sito: http://www.s-gem.se/publications/presentations/documents/hoehle_sgem2009.pdf

Höhle M., Paul M., Held L., *Statistical approaches to the surveillance of infectious diseases for veterinary public healt,* Ludwig Maximillians Universitat Munchen, Department of Statistic, Technical Report Number 014, 2007.

Istituto Superiore di Sanità, *Scenari di diffusione e controllo di una pandemia influenzale in italia*, Rapporti ISTISAN 06/33, 41 p., 2006

Kendall, D. G. (1956). *Deterministic and stochastic epidemics in closed populations.* In Proc. Third Berkeley Symp. Math. Statist. Prob., 1954--1955, Vol. IV, University of California Press, pp. 149--165.

Kermack W.O. e McKendrick A.G. *A Contribution to the Mathematical Theory of Epidemics* Proceedings of the Royal Society of London. Series A, Vol. 115, (1927), pp. 700-721.

Lewis P.D. *R for Medicine and Biology*, Jones and Bartlett Series in Biomedical Informatics, Series Editor Jules J. Berman, chap. 16 pp. 233-240, 2010.

Lilia Leticia Ramirez Ramirez, *On the dynamics of infectious diseases in non-homogeneous populations.* A thesis presented to the University of Waterloo in fullment of the thesis requirement for the degree of Doctor of Philosophy in Statistics - Waterloo, Ontario, Canada, 2008.

Luigi Vajani, *Moderni metodi statistici per lo studio delle epidemie*, Estratto da "Bollettino I.S.M.", volume 40, 1961. Università Cattolica di Milano.

Manfredi P., Tarini F., Modelli epidemiologi: teoria e simulazione. (I), Università degli Studi di Pisa Dipartimento di Statistica e Matematica Applicata all'Economia Report n. 91, 1995

Masarotto G., *Analisi delle Serie Temporali lucidi delle lezioni a.a. 2004/05,* Guido Masarotto Facoltà di Scienze Statistiche Università di Padova, 2005,disponibile sul sito: http://sirio.stat.unipd.it

Michael Thrusfiel: *Veterinary Epidemiology* Third Edition, Blackwell Publishing

Pecoraro A., Tiano A., Vendegna V., *Modello della Dinamica Epidemica in una Metapopolazione* in Società Italiana di Ecologia - XIII Congresso Nazionale disponibile su http://www.ecologia.it/congressi/XIII/articles, 2003

Pugliese A., *Modelli di epidemie in popolazioni omogenee,* 2004, disponibile sul sito: http://www.science.unitn.it/~anal1/biomat/aa0304/epi-omog.pdf

R.Ross, *The prevention of Malaria*, 2nd ed., New York, E.P. Dutton & Co., 1911.

Regione del Veneto, AA.VV referente Gallo G., *Sistema di sorveglianza delle malattie infettive*, disponibile sul sito http://www.regione.veneto.it, 2010

Roland Regoes, Stochastic Simulation of Epidemics, disponibile su: http://www.tb.ethz.ch/education/model/SIR-stoch

S. Sakino, C. Hayashi, *On the analysis of epidemic model I, (Theoretical approach)* Ann. Inst. Star. Math. Vol. X (1959), pp. 261-275.

Stroup, D.F., Williamson, G.D., Herndon, J.L., and Karon, J.M. *Detection of aberrations in the occurrence of notifiable diseases surveillance data*, Stat Med 8 (1989): 323–9; discussion 331-2.

Toniolo F. Gallo G., La sorveglianza di sanità pubblica delle malattie infettive,disponibile sul sito http://www.regione.veneto.it, 2010