DERIVAZIONE DEI MODELLI PREVISIONALI PER LE ROTATORIE URBANE. UN CASO STUDIO

Giuffrè O.
Full Professor - Department of Road Infrastructure Engineering – Università di Palermo, e-mail: ogiuffre@unipa.it
Granà A.
Assistant Professor - Department of Road Infrastructure Engineering – Università di Palermo, e-mail: anna.grana@unipa.it
Giuffrè T.
Postdoctoral researcher - Department of City and Land, Università di Palermo, e-mail: tullio.giuffre@polimi.it
Marino R..
PhD candidate in Road Infrastructure Engineering - Department of Road Infrastructure Engineering, Università di Palermo, e-mail: roberta.marino@unipa.it

ABSTRACT

Come è noto, i modelli previsionali d'incidentalità (in campo internazionale conosciuti come Safety Performance Functions) vengono ormai da tempo impiegati nelle valutazioni di sicurezza; ad essi è affidato il compito di condensare le conoscenze pregresse sulla sicurezza di entità (tronchi stradali, intersezioni, etc.) simili a quelli in considerazione.

Quando integrati in una procedura di stima empirico-bayesiana, i modelli previsionali d'incidentalità concorrono alla correzione della distorsione dovuta alla regressione verso la media ed all'accrescimento della precisione della stima.

Nella presente memoria viene presentato un esempio di determinazione di tali modelli per il caso di rotatorie urbane.

Attraverso lo stesso esempio, si intende mostrare l'opportunità di prendere in considerazione la correlazione temporale dei dati, in conseguenza della quale l'ipotesi di indipendenza che autorizza l'impiego di modelli lineari generalizzati (GLM) cade in difetto.

Contemporaneamente, si intende mostrare come uno sforzo addizionale che tenga conto della vera struttura di correlazione dei dati possa essere compensato da una migliore precisione nella stima dei parametri del modello.

Keywords: modelli di incidentalità, metodi di stima in presenza di correlazione temporale.

1. INTRODUZIONE

Il metodo empirico-bayesiano (EB) per le analisi di sicurezza stradale è ampiamente applicato per diverse finalità: indagini sui siti suscettibili di miglioramento in termini di sicurezza, valutazioni degli effetti sulla sicurezza degli interventi migliorativi, valutazioni dei benefici potenziali in termini di sicurezza ottenibili in seguito a

interventi di riqualificazione, etc. Infatti, il metodo EB rappresenta la struttura teorica fondamentale per correlare le informazioni sui dati di incidentalità pregressa alla conoscenza sulla sicurezza di elementi infrastrutturali simili a quelli da esaminare; ciò è reso possibile attraverso l'impiego di modelli previsionali dell'incidentalità, noti in campo internazionale come Safety Performance Functions (SPFs).

I principali vantaggi attribuibili al metodo EB, come descritti da Hauer (2002), sono riconducibili alla correzione del fenomeno della regressione alla media ed all'aumento di precisione nella stima.

Entrambi i due aspetti influenzano positivamente l'affidabilità complessiva della stima, di particolare interesse sia quando devono essere effettuate decisioni in merito a scelte progettuali con ricadute sulla sicurezza, sia pure quando si presenta la necessità di valutare in modo accurato gli effetti reali di un determinato intervento migliorativo.

Relativamente alla precisione della stima, può farsi notare che la varianza stimata con il metodo EB non è mai superiore alla varianza stimata solamente in base ai dati di incidenti ed è generalmente minore di quest'ultima. Stime più accurate ottenute con il metodo EB sono riconducibili principalmente all'efficienza delle stime con SPFs, che a loro volta dipendono dall'accuratezza dell'analisi di regressione dei dati di incidentalità utilizzati nei modelli previsionali.

Il più delle volte la messa a punto di un modello di incidentalità è fondata su osservazioni estese a diversi periodi temporali su elementi stradali (tronchi, intersezioni, etc.) che presentano caratteristiche simili.

Una struttura di dati di questo genere (individuati in letteratura come *longitudinal* o *panel data*) fa venir meno l'ipotesi di indipendenza della variabile dipendente e crea uno specifico problema nella messa a punto delle SPFs.

Con riferimento alla precisione della stima, infatti, nel caso di osservazioni ripetute più volte sullo stesso sito, può essere inficiata seriamente la stessa validità del modello dal momento che esso si fonda su coefficienti di regressione e sulla stima della relativa varianza anche fortemente distorti (Diggle et al., 2002).

E' noto in letteratura che in questo caso una stima dei regressori ottenuta attraverso metodi di tipo GEEs (*Generalized Estimating Equations*) può ancora fornire stime coerenti o "robuste" (o a *sandwich*) dei parametri della regressione anche se la matrice di correlazione non è correttamente specificata (Fitzmaurice, 1995; Zeger et al., 1986).

Tuttavia si è dimostrato che l'efficienza degli stimatori diminuisce quando la correlazione aumenta e diventa apprezzabile quando la correlazione è maggiore di 0.4.

Inoltre, le perdite in termini di efficienza – allorchè l'assunzione di indipendenza è falsa – compromette la significatività della stima nel caso di correlazioni tra i dati maggiori di 0.5. Errori significativi si commettono, pertanto, per correlazioni positive o negative molto elevate.

Altri ricercatori ritengono che la ricerca della matrice di correlazione corretta è importante solo quando si sviluppano modelli marginali che ricorrono a dati non completi (Lord et Persaud, 2000; Lord et al., 2005); ciò nonostante gli stessi ricercatori ritengono che i coefficienti di regressione sono solitamente sottostimati quando gli effetti temporali non sono inclusi nell'impostazione del modello (Lord et al., 2005; Hardin, J. J., and J. M. Hilbe, 2003).

A partire da queste considerazioni, lo studio si basa sull'idea che nella formulazione dei modelli previsionali di sicurezza (SPFs) uno sforzo addizionale per ottenere la reale

struttura di correlazione dei dati sia compensato da un miglioramento nella precisione nella stima dei parametri del modello e che, pertanto, le valutazioni successive (ad esempio nelle applicazioni di tipo empirico-bayesiane) risultino migliorate nella loro affidabilità.

Nella memoria, inoltre, utilizzando un campione di dati di incidenti occorsi in rotatoria, viene presentato un caso studio per illustrare i principali aspetti metodologici dell'approccio proposto.

2. CONSIDERAZIONI METODOLOGICHE

La ricerca della correlazione nelle variabili dipendenti non richiede azioni fondamentalmente diverse da quelle che solitamente vengono intraprese per la calibrazione di un modello predittivo attraverso gli usuali metodi probabilistici basati sul concetto della massima verosimiglianza (ad esempio, attraverso metodi GLM, di tipo lineare generalizzato).

La definizione del modello richiede in primo luogo la specificazione della sua forma funzionale e della distribuzione di probabilità della variabile dipendente; la valutazione del modello e la verifica della bontà di adattamento ai dati osservati dovranno poi essere effettuate mediante analisi statistica o altri strumenti basati sull'analisi dei residui.

Nonostante la similarità formale, quando si vogliono interpretare dati di tipo longitudinale (ad esempio osservazioni ripetute nel tempo sugli stessi siti), le attività di cui sopra devono essere condotte considerando che la struttura di correlazione della variabile dipendente non è nota e l'analisi, quindi, deve essere svolta non solo per determinare i parametri del modello, ma anche per approssimare la struttura di correlazione esistente tra i dati di partenza.

Gli autori non intendono intraprendere in questa sede un'analisi sui problemi in cui si incorre quando si applicano i modelli GEE all'analisi di regressione.

Ci si riferirà, pertanto, alla procedura sviluppata da Liang e Zeger (1986a; 1986b), così come agli studi sull'argomento condotti da Hardin e Hilbe (2003) e da Diggle et al. (2002); quanto all'applicazione dell'approccio GEE alla sicurezza stradale, sono anche di riferimento gli studi di Lord (2000) e di Lord e Persaud (2000).

I metodi di tipo GEE consentono l'interpretazione dei dati sperimentali tramite un modello marginale che restituisce una variabile media di risposta (funzione di un certo numero di variabili indipendenti) per le osservazioni che condividono le stesse covariate (Zeger et al., 1988); più specificamente un modello GEE restituisce le variazioni in media nella popolazione per ogni incremento unitario di una covariata.

Nella sua essenza la procedura GEE consiste in un'estensione dei modelli lineari generalizzati (GLMs) e può essere considerata come uno strumento efficace per saggiare un'ipotesi circa l'influenza di uno o più fattori su una variabile dipendente di tipo binario o distribuita in modo esponenziale, allorché si disponga di dati raccolti nel tempo su più siti.

Per la stima dei coefficienti, occorre risolvere la seguente equazione:

$$\sum_{i=1}^{k} \boldsymbol{D}_i' \boldsymbol{V}_i^{-1} \left[\boldsymbol{Y}_i - \boldsymbol{\mu}_i \right] = \boldsymbol{0} \qquad (1)$$

in cui per il sito *i-esimo (i = 1, 2 ,..., k)*,

$$\boldsymbol{D}_i = \frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial \mu_{1i}}{\partial \beta_1} & \frac{\partial \mu_{1i}}{\partial \beta_2} & \cdots \frac{\partial \mu_{1i}}{\partial \beta_p} \\ \vdots & \vdots & \vdots \\ \frac{\partial \mu_{ni}}{\partial \beta_1} & \frac{\partial \mu_{ni}}{\partial \beta_2} & \cdots \frac{\partial \mu_{ni}}{\partial \beta_p} \end{bmatrix} \quad (2)$$

con:
$\boldsymbol{V}_i$ = matrice di covarianza
$\boldsymbol{\beta}_p$ = parametri da stimare,
$\boldsymbol{\mu}_i$ = valore atteso dell'osservazione *i*.

Le precedenti espressioni considerano le osservazioni (Y_{ij}, $\boldsymbol{X}_{ij}$) ripetute al tempo t_{ij}, $j = 1,\ldots n_i$ per ogni siti i; Y_{ij} rappresenta la variabile dipendente e $\boldsymbol{X}_{ij}$ è il vettore delle covariata di dimensione $p \times 1$. Allo stesso tempo $\boldsymbol{Y}_i$ è il vettore $n_i \times 1$ $(Y_{i1},\ldots, Y_{in_i})'$ e $\boldsymbol{X}_i$ è la matrice $n_i \times p$ $(\boldsymbol{X}_{i1},\ldots, \boldsymbol{X}_{in_i})'$ per il generico sito. Con la notazione usuale impiegata nei modelli di regressione, $\boldsymbol{\mu_i} = \boldsymbol{g}^{-1}(\boldsymbol{X_i\beta})$, in cui $\boldsymbol{\beta}$ è un vettore $p \times 1$ di parametri da stimare e g rappresenta la relazione funzionale intercorrente tra variabile dipendente e le covariata (o variabili indipendenti).

La matrice R(••) di dimensione $n_i \times n_i$ (da specificare a priori in base alle informazioni disponibili) descrive la correlazione temporale intercorrente fra le diverse osservazioni.

Le matrici di covarianza possono essere espresse come segue:

$$\boldsymbol{V_i} = \boldsymbol{A_i^{1/2}}\boldsymbol{R_i}(\cdot\cdot)\boldsymbol{A_i^{1/2}} \quad (3)$$

e

$$\boldsymbol{cov}(\hat{\boldsymbol{\beta}}) = \sigma^2\left[\sum_{i=1}^{k} \boldsymbol{D_i'V_i^{-1}D_i}\right]^{-1} \quad (4)$$

in cui $\boldsymbol{A}_i$ è una matrice diagonale di dimensione ($n_i \times n_i$) comprendente le varianze degli elementi di $\boldsymbol{Y}_i$, calcolati a partire dai regressori $\boldsymbol{\beta}$.

La soluzione simultanea delle equazioni precedentemente descritte, attraverso il metodo dei minimi quadrati pesati (Green, 1984) fornisce la soluzione GEE per i regressori $\boldsymbol{\beta}$ e per il tipo di correlazione assegnata. Tenuto conto che non è possibile conoscere a priori il tipo di correlazione intercorrente tra i dati, Liang e Zeger (1986a) hanno proposto l'uso di una matrice di correlazione $\hat{\boldsymbol{V}}_{\boldsymbol{i}}$, basata sulla matrice di correlazione $\hat{\boldsymbol{R}}_{\boldsymbol{i}}$. In questo modo, si effettua la stima dei coefficienti sostituendo $\boldsymbol{V_i}$ con $\hat{\boldsymbol{V}}_{\boldsymbol{i}}$ nelle equazioni fondamentali sopra riportate.

Quanto sopra evidenzato pone in evidenza che la specificazione della forma della correlazione nella variabile dipendente rappresenta il problema centrale nell'ottenimento di stime più efficienti. Ciò rappresenta l'argomento trattato in questo lavoro. Infatti, sebbene i modelli GEE siano nella generalità dei casi "robusti" anche nel

caso di una non corretta specificazione della struttura di correlazione (Liang & Zeger, 1986), quando la struttura specificata è molto lontana da quella reale, ci si può aspettare perdita di efficienza nelle stime (Ballinger, 2004).

Per ottenere la reale struttura di correlazione (o meglio, per approssimarsi ad essa) è necessario verificare ipotesi diverse della correlazione nelle osservazioni.

Nel caso dei dati di incidentalità sembra essere appropriata una struttura dipendente di ordine (*n-1*); in alternativa è possibile non definire a priori la struttura di correlazione e ricavarla come esito computazionale del modello GEE. La scelta di una struttura dipendente di ordine (*n-1*) implica che $R(\cdot\cdot)$ sia definito dalle seguenti espressioni:

$$[\mathbf{R_i}]_{jk} = \begin{cases} \alpha^{|t_{ij}-t_{ik}|} & |t_{ij}-t_{ik}| \le m \\ 0 & |t_{ij}-t_{ik}| > m \end{cases} \qquad (5)$$

Quando non viene fissata a priori la struttura della matrice di correlazione (ovvero quando si procede mediante una "*unstructured matrix*") le stime di tutte le possibili correlazioni fra le osservazioni all'interno di ciascuna entità vengono incluse nella stima della varianza. Altre scelte per la matrice di correlazione possono ovviamente essere effettuate e fra queste, in particolare:

- che non vi sia un ordine logico nelle osservazioni ("*exchangeable matrix*");
- che tutte le osservazioni siano indipendenti le une dalle altre ("*independence matrix*").

In quest'ultimo caso si perde il vantaggio dell'uso della procedura GEE, poiché la suddetta struttura non tiene conto della correlazione tra le osservazioni e, pertanto, le stime dei parametri sono quelle che più agevolmente possono ottenersi attraverso un modello lineare generalizzato GLM. Ciò nonostante, le soluzioni sotto l'ipotesi di indipendenza o di assenza di correlazione nelle osservazioni risultano utili nella messa a punto del modello definitivo, rappresentando il più delle volte il modello di partenza.

In base a quanto detto, la selezione della forma funzionale del modello e la ricerca della struttura di correlazione più appropriata costituiscono il momento centrale di tutta la procedura GEE.

2.1 Criteri di valutazione della bontà di adattamento del modello ai dati osservati

Visto che le osservazioni non sono indipendenti le une dalle altre, è plausibile aspettarsi che i residui siano parimenti non indipendenti.

In situazioni di questo tipo i metodi basati sul criterio della massima verosimiglianza, o altri metodi comunemente impiegati per valutare la capacità interpretativa dei modelli, non risultano adeguati per valutare la bontà di adattamento del modello ai dati sperimentali.

Nella presente memoria, si è deciso, quindi, di considerare per la validazione dei risultati i seguenti criteri informativi:

– l'indicatore R^2 marginale che, in accordo a Zheng (2000), confronta lo scarto quadratico tra i valori stimati dal modello con il quadrato delle deviazioni delle osservazioni dal valore medio della variabile dipendente:

$$R_m^2 = 1 - \frac{\sum_{t=1}^{T}\sum_{i=1}^{k}\left(Y_{it} - \hat{Y}_{it}\right)^2}{\sum_{t=1}^{T}\sum_{i=1}^{k}\left(Y_{it} - \overline{Y}_{it}\right)^2} \quad (6)$$

in cui *t (t = 1,2,...,T)* rappresenta il generico anno di osservazione ed *i (i = 1, 2,..., k)* è il generico sito. La statistica R^2 viene solitamente interpretata come l'aliquota di varianza nella variabile di risposta che è spiegata dal modello implementato (Hardin e Hilbe, 2003).

– Il criterio di Pan (2001), noto come QIC (*Quasi-likelood under the indepence model criterion*), il quale rappresenta un'estensione del criterio informativo di Akaike ed è utile al fine di confrontare le matrici di covarianza dei modelli GEE con la matrice di covarianza generata sotto l'ipotesi di indipendenza. La migliore struttura di correlazione è quella che presenta il QIC più vicino allo zero.
Secondo Pan (2001) la discrepanza tra il modello effettivo M_1 e quello reale M^* - entrambi indicati dal vettore dei parametri $\boldsymbol{\beta}$ - può essere ottenuta dalla seguente relazione (Kullback and Leibler, 1999):

$$\Delta_0\left(\beta, \beta^*\right) = E_{M^*}\left[-2L\left(\beta; D\right)\right] \quad (7)$$

in cui il valore atteso E_{M^*} è funzione della reale distribuzione di *D* e $L(\boldsymbol{\beta}; D)$ denota il logaritmo della funzione di likelihood del modello M_1.
Se si ricorre al modello indipendente ($R = I$), le coppie di osservazioni (Y_{ij}, X_{ij}) in *D* sono considerate indipendenti ed il criterio di AIC (Akaike Information Criterion) può essere considerato uno stimatore non distorto del parametro $E_{M^*}[\Delta_0(\hat{\beta}, \beta^*)]$.
Ne segue che il modello da selezionare da un insieme di modelli è quello con il più piccolo valore $\Delta_0(\hat{\beta}, \beta^*)$.
Da un punto di vista concettuale il QIC è ottenuto da Pan (2001) sostituendo *L* con *Q*; l'entità dello scostamento (nell'ipotesi di indipendenza) è quindi espressa dalla seguente espressione:

$$\Delta_0\left(\beta, \beta^*, I\right) = E_{M^*}\left[-2Q\left(\beta; I, D\right)\right] \quad (8)$$

dove *I* denota la matrice di correlazione di lavoro sotto l'ipotesi di indipendenza.
Per una matrice di correlazione $R \neq I$, è stato dimostrato che lo stimatore di $E_{M^*}\left[\Delta_0\left(\hat{\beta}, \beta^*, I\right)\right]$ può essere determinato dalla relazione che segue:

$$QIC(R) \equiv -2Q\left(\hat{\beta}(R); I, D\right) + 2trace\left(\hat{\Omega}_I \hat{V}_r\right) \quad (9)$$

dove: $\left(\hat{\Omega}_I, \hat{V}_r\right)$ sono rispettivamente gli stimatori empirici di $\Omega_I = \sum_{1=1}^{k} D_i' V_i D_i$ e $\hat{V}_r$ è la stima robusta (o sandwich) della covarianza cov($\hat{\beta}$).

Si rimanda ad Hardin e Hilbe (2001) per ulteriori applicazioni del QIC finalizzate alla scelta della migliore struttura di correlazione per un modello marginale GEE.

3. APPLICAZIONE DELLA PROCEDURA AD UN CASO STUDIO

Al fine di illustrare la metodologia sopra esposta è stata calibrata una SPF da utilizzare nelle valutazioni di sicurezza inerenti possibili miglioramenti dell'infrastruttura. L'approccio metodologico è stato applicato ad un campione di incidenti occorsi in corrispondenza di 21 rotatorie urbane della città di Trento, la cui analisi descrittiva è presentata in questo stesso convegno da Mauro e Corradini (2008).

Le principali caratteristiche geometriche delle rotatorie considerate sono riportate nella Tabella 1. Ulteriori informazioni sono desumibili dalla successiva Figura 1.

La Figura 2 mostra i valori medi del traffico giornaliero medio annuo, calcolati per ciascun sito mediando i valori del traffico dall'anno di entrata in servizio sino al 2004, ed i valori medi degli incidenti accaduti.

Tabella 1 – Caratteristiche delle rotatorie del campione

rotatorie		Rami afferenti	D_{est} [m]	D_{int} [m]
1	Al Desert	4	49	31
2	Berlino	5	162	150
3	Cognola	4	19	4
4	De Gasperi	4	38	24
5	Lr. Don. Sangue	4	36	18
6	Lr. Med. d'Oro	3	40	20
7	Largo Prati	4	36	18
8	Loc. Stella di Man	4	48	34
9	Maccani	5	114	104
10	Monte Baldo	3	16	4
11	Piedicastello	4	52	39
12	Ponte Ravina	4	38	25
13	Ponte San Lorenzo	4	32	12
1	Roncafort	5	34	12
15	Soprasasso	4	37	24
16	Tridente	4	59	45
17	Via alla Cascata	4	31	20
18	Via Dallafior	4	32	20
19	Via Fersina	3	31	16
20	Via Galassa	3	38	24
21	Via Vittorio Veneto	2	23	5

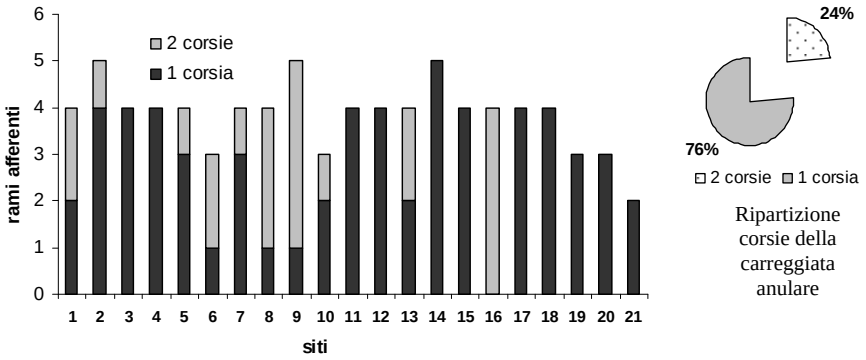

Figura 1 – Numero rami afferenti a ciascuna rotatoria e ripartizione corsie a ramo.

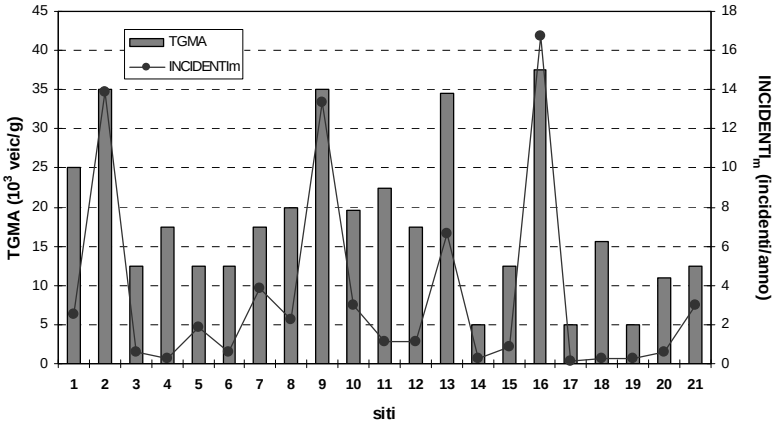

Figura 2 – Valori medi degli incidenti e del traffico giornaliero medio annuo nei siti esaminati.

La Tabella 2 riporta gli incidenti accaduti in corrispondenza delle rotatorie esaminate nell'arco temporale compreso tra il 1997 ed il 2004.

Oltre all'eterogeneità constatata nelle caratteristiche geometriche dei siti, quanto a diametro dell'isola centrale, numero di rami afferenti, numero di corsie in entrata e sulla carreggiata anulare), un'ulteriore differenza è riconducibile all'anno di entrata in servizio di ciascuna rotatoria, ciò che rende differente l'estensione del periodo di osservazione dei sinistri accaduti.

In ogni caso, l'estensione del campione è troppo modesta per consentire l'estrazione di gruppi omogenei o per finalizzare l'analisi unicamente ai siti caratterizzati dallo stesso periodo di osservazione; per queste ragioni, lo studio di cui si riferisce è stato sviluppato unicamente con propositi dimostrativi, ovvero per mostrare le problematiche procedurali relative al miglioramento dell'affidabilità delle stime nelle previsioni inerenti il fenomeno incidentale.

Premesso ciò, l'intento degli autori è risolvere alcuni nodi procedurali attinenti:

– la ricerca della migliore forma funzionale dell'equazione del modello previsionale;

- la definizione della distribuzione della variabile dipendente;
- la ricerca della probabile struttura di correlazione nella variabile dipendente.

Tabella 2 – Incidenti accaduti

rotatorie	1997	1998	1999	2000	2001	2002	2003	2004
1	-	-	-	-	-	-	5	0
2	9	15	12	12	24	9	10	15
3	1	0	1	0	1	1	1	0
4	-	0	0	0	1	1	0	0
5	1	2	0	1	0	0	6	5
6	1	0	0	3	0	0	0	1
7	7	6	2	4	5	4	1	2
8	2	10	3	0	2	1	0	0
9	12	15	12	12	15	13	14	14
10	-	-	-	-	4	3	5	0
11	0	2	0	3	2	0	1	1
12	1	1	2	2	0	1	2	0
13	9	9	8	5	5	6	6	5
14	0	0	0	0	2	0	0	0
15	-	3	0	2	1	0	0	0
16	-	-	-	-	9	22	19	17
17	0	0	0	0	0	1	0	0
18	0	1	0	0	0	1	0	0
19	-	-	-	-	1	0	0	0
20	-	0	2	0	1	1	0	0
21	-	-	-	-	-	-	-	3

(-) dato non disponibile

3.1 La forma funzionale dell'equazione del modello previsionale

Per la ricerca della migliore forma funzionale dell'equazione del modello previsionale si è ricorso al metodo Integrale – Differenziale (di seguito metodo ID) proposto da Hauer e Bamfo (1997) e applicato da Lord et al. (1999). Hauer e Bamfo (1997) descrivono il metodo ID, indicando che il metodo consente di speculare proficuamente sulla forma funzionale del modello anche quando la dispersione dei dati non mostra un andamento preciso.

Come variabile esplicativa è stata scelta il traffico annuo giornaliero medio entrante (TGMA) nell'intersezione. In accordo al metodo ID sopra menzionato, la funzione integrale empirica (*Empirical Integral Function* - *EIF*) consente una stima della funzione integrale effettiva della relazione funzionale del modello cercato (cfr. Figura 3a). I grafici delle Figure 3b e 3c riportano il *ln(EIF)* in funzione rispettivamente del *ln(TGMA)* e del *TGMA*; il metodo *ID* ha mostrato che sia la funzione potenza (F_1), sia quella esponenziale (F_2) possono essere assunte per rappresentare le forme funzionali del modello. Di seguito queste saranno denominate Modello 1 e Modello 2 rispettivamente.

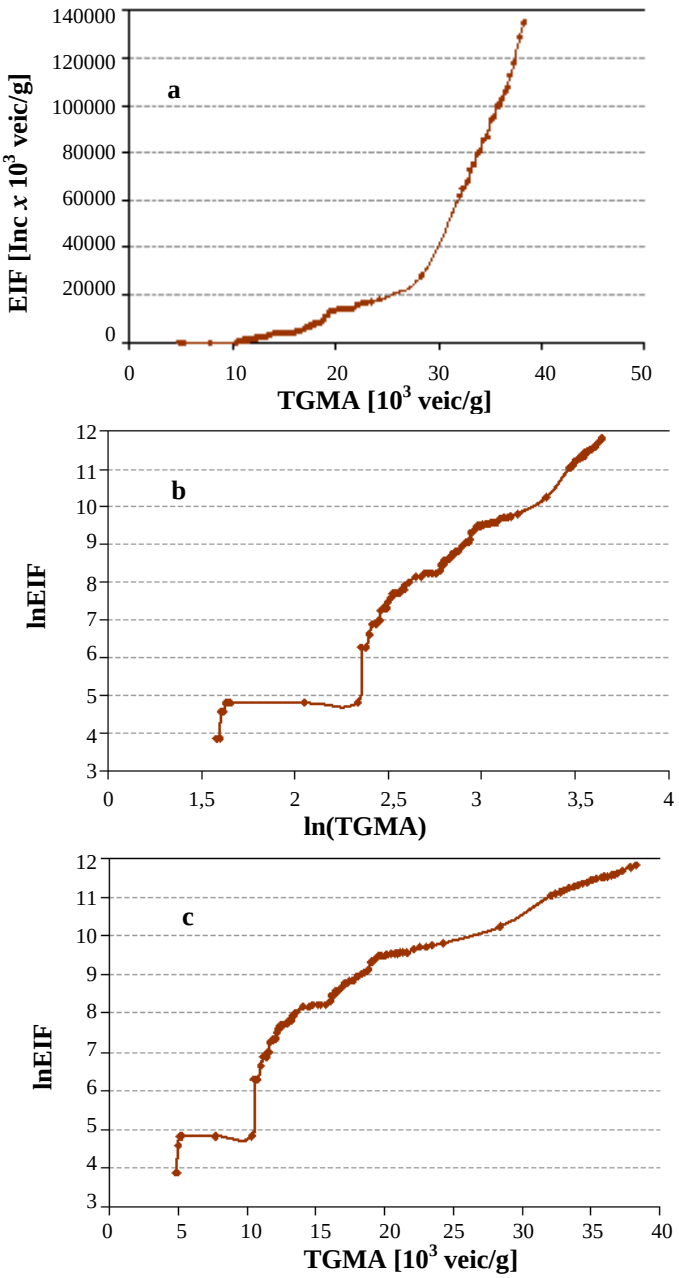

Figura 3 – Forma funzionale: a) EIF per i dati; b) funzione potenza F_1 (Modello 1); c) funzione esponenziale F_2 (Modello 2)

Ne segue che i modelli considerati sono:

Modello 1: $\mathrm{E}\{k\} = \alpha \cdot AADT_{ij}^{\beta}$ **Modello 2**: $\mathrm{E}\{k\} = \alpha \cdot e^{\beta \cdot AADT_{ij}}$

dove:

- *i (i = 1,2,...,k)* = sito generico
- *j (j = 1,2,...,n_i)* = anno di osservazione
- $\boldsymbol{E}\{k_{ij}\}$ = numero atteso di incidenti nel sito *i* nell'anno *j* (incidenti/anno)
- $TGMA_{ij}$ = traffico totale entrante nel generico sito *i* nell'anno *j* (veic/g)
- α, β = coefficienti da stimare.

La seguente tabella riporta le informazioni riepilogative del fit dei dati a partire dal metodo *ID* per i due modelli suggeriti dalla funzione itegro-differenziale empirica.

Tabella 3 – Fit dei dati per il Modello 1 e per il Modello 2.

	α	β	R^2
$F_1 = \alpha\, AADT^{\beta}$	0.14	3.76	0.97
$F_2 = \alpha\, e^{\beta AADT}$	113.64	0.203	0.89

3.2 Distribuzione della variabile di risposta

Quando si applicano i modelli GEE nell'interpretazione dei dati, così come per le equazioni lineari generalizzate (GLM), la specificazione della distribuzione della variabile di risposta tiene conto che la varianza sia espressa come una funzione della risposta media. La varianza viene, quindi, utilizzata per il calcolo della matrice di covarianza.

In considerazione che la mancata specificazione della distribuzione della variabile di risposta può avere conseguenze importanti sulle stime dei parametri di regressione, è stato implementato preliminarmente un test di sovradispersione per i due modelli considerati.

Il test ha avuto inizio stimando la regressione di Poisson per ciascun modello (cfr. Tabella 4) attraverso le equazioni lineari generalizzate e calcolando successivamente le stime $\hat{\mu}_{ij}$. A tal scopo è stato utilizzato il software GenStat (Payne et al., 2004).

Sono stati quindi utilizzati al fine di interpretare i dati i seguenti modelli ausiliari di regressione lineare dei minimi quadrati ordinari (OLS):

OLS$_1$: $\dfrac{\left(y_{ij} - \hat{\mu}_{ij}\right)^2 - y_{ij}}{\hat{\mu}_{ij}} = \gamma \hat{\mu}_{ij} + u_{ij}$ **OLS$_2$:** $\dfrac{\left(y_{ij} - \hat{\mu}_{ij}\right)^2 - y_{ij}}{\hat{\mu}_{ij}} = \gamma + u_{ij}$

in cui:

- $\hat{\mu}_{ij}$ = stime ottenute attraverso la regressione da GLM (cfr. Tabella 4);
- y_{ij} = incidenti accaduti nel sito *i* nell'anno *j*;
- γ = parametro da stimare attraverso la regressione OLS;
- u_{ij} = errore.

Tabella 4 – Modelli base di Poisson per i test di sovradispersione (GLM)

	Modello 1		**Modello 2**	
	stima	e.s.	stima	e.s.
constant	0.0012	0.427	0.210	0.162
AADT	2.567	0.128	0.114	0.005
$\mathbf{R^2_{dev}}$	0.68		0.69	

Per entrambe le forme funzionali ausiliarie, la statistica *t* per γ è asintoticamente normale sotto l'ipotesi nulla di assenza di sovradispersione, in confronto all'alternativa di sovradispersione.

In altre parole, i precedenti test ausiliari di regressione sono utilizzati per individuare la forma più appropriata per la distribuzione della variabile dipendente nell'ambito della famiglia delle curve binomiali negative, cioè per indicare se la varianza varia con il quadrato della media (NB2) ovvero se intercorre un semplice rapporto di linearità (NB1).

I risultati dei test di sovradispersione, riepilogati in Tabella 5, mostrano che l'ipotesi nulla di nessuna sovradispersione deve essere rifiutata rispetto all'alternativa di sovradispersione sia per la forma NB1, sia per la NB2. Ciò nondimeno, poiché le statistiche *t* riportate indicano che la forma NB1 è più appropriata della forma NB2, si assumerà nel prosieguo che la funzione varianza è espressa da $\omega_i = V\left[k_i \middle| \mathrm{x}_i\right] = (1+\varphi)\cdot\mu_i$ (o, come è usuale nella regressione lineare generalizzata, da $\omega_i = \varphi' \cdot \mu_i$), in cui φ è il parametro di dispersione e $\varphi' = 1 + \varphi$.

Tabella 5 – Risultati del test ausiliario di regressione.

	OLS_2 (Poisson *vs* NB1)				**OLS_1 (Poisson *vs* NB2)**			
	stima	**e.s.**	**t**	**t-Prob**	**stima**	**e.s.**	**t**	**t-Prob**
Modello 1								
Costante	1,18	0,33	3,57	<0,001	-	-	-	-
fit	-	-	-	-	0,13	0,06	1,97	0,051
Modello 2								
Costante	1,12	0,34	3,3	0,001	-	-	-	-
fit	-	-	-	-	0,11	0,06	1,75	0,082

3.3 La struttura di correlazione

Relativamente al modello selezionato attraverso la precedente analisi (Modello 2 con la distribuzione NB1), le regressioni GEE sono state condotte considerando differenti matrici di correlazione, cioè assumendo che le osservazioni ripetute fossero correlate in modi diversi. Anche in questo caso si è ricorso al software GenStat.

Come anticipato, sono state considerate successivamente:

- una matrice di correlazione di tipo indipendente (*independence matrix*), per la quale i risultati forniti dalla procedura GEE coincidono con quelli ottenibili da

regressione in ambiente GLM; in questo caso, semplice ma parimenti irrealistico $\mathbf{R}_i = \mathbf{I}_{ni}$ (in cui $\mathbf{I}_{ni}$ rappresenta la matrice identità $n_i \, x \, n_i$);
- in contrasto con l'ipotesi di indipendenza, è stata specificata una matrice di correlazione *unstructured* per consentire stime libere della correlazione tra le osservazioni;
- è stata, infine, assunta una struttura di correlazione di un processo stazionario *m*-dipendente.

I risultati della regressione GEE effettuata con riferimento alle diverse matrici di correlazione sono riportati in Tabella 6, in cui sono presenti i valori dell'R_m^2 e della statistica Pan (*QIC*). Per il calcolo del QIC, con la distribuzione NB1, la funzione di likelihood *L* è stata determinata con la seguente espressione:

$$ln\,L(\varphi,\beta) = \sum_{z=1}^{N}\left\{\left(\sum_{j=0}^{y_z-1} ln\left(j+\varphi^{-1}\hat{\mu}_z\right)\right) - ln\,y_z! - \left(y_z+\varphi^{-1}\hat{\mu}_z\right)ln\left(1+\varphi\right) + y_z\,ln\,\varphi\right\}$$

in cui y_z (z = 1,2,…, N) è la generica osservazione al sito *i* nell'anno *j* e $\hat{\mu}_z$ è la stima del valore atteso.

I risultati in Tabella 6 confermano che il Modello 2 interpreta i dati meglio del Modello 1, anche dopo le regressioni GEE. Contestualmente, un modello che tiene conto dell'ipotesi di dipendenza sembra approssimare la reale struttura di correlazione in modo appropriato. Tuttavia, lievi differenze nelle statistiche Pan, nei casi in cui si è ricorso alle matrici di correlazione *dependence* e *unstructured* (anche se entrambi i valori assoluti tendono a zero) non permettono di determinare la migliore struttura di correlazione in modo certo.

Tabella 6 – Quadro riassuntivo delle regressioni GEE con diverse matrici di correlazione.

	GEE – distribuzione binomiale negativa (NB1)					
	independence		*unstructured*		*dependence* (ordine=7)	
	stima	e.s.	stima	e.s.	stima	e.s.
Modello 1						
costante	0.0012	0.938	0.0055	0.919	0.0052	0.916
TGMA	2.567	0.281	2.146	0.284	2.168	0.284
Statistica Pan (QIC)	(44.27)		14.46		18.50	
R_m^2	0.712		0.714		0.716	
Modello 2						
costante	0.209	0.340	0.187	0.257	0.252	0.273
TGMA	0.115	0.010	0.117	0.009	0.110	0.009
Statistica Pan (QIC)	(45.70)		48.66		25.91	
R_m^2	0.739		0.738		0.736	

Si è deciso così di analizzare in modo approfondito l'adeguatezza del modello da un altro punto di vista. Coerentemente alle finalità della ricerca, si è pensato che l'ampiezza dell'intervallo di confidenza per la media (insieme all'R_m^2 ed al QIC potessero rappresentare un criterio per selezionare il modello, così come per decidere la

struttura di correlazione della variabile di risposta. Visto che le stime dei parametri GEE ($\hat{\beta}_{GEE}$) risultano asintoticamente normali, ne deriva che un intervallo di confidenza asintotico (1-α)100% per la media (x' $\hat{\beta}_{GEE}$) è dato da:

$$x'\hat{\beta}_{GEE} \pm z_{1-\frac{\alpha}{2}} \sqrt{\left(x' V_{GEE} x\right)}$$

La Tabella 7 mostra gli intervalli di confidenza al 99% di x' $\hat{\beta}_{GEE}$ per i modelli costruiti con riferimento all'ipotesi di struttura di correlazione *unstructured* (a), *dependance* (b) e *independence* (c).

Tabella 7 – Intervallo di confidenza di $x'\beta_{GEE}$ al 99%.

	$\sqrt{(x' V_{GEE} x)}$		
	media	min	max
a: unstructured	0.152	0.107	0.218
b: dependence	0.187	0.155	0.238
c: independence	0.193	0.111	0.297

Si può agevolmente osservare che l'ipotesi di una struttura di correlazione *unstructured* consente la migliore interpretazione dei dati in termini di precisione della stima della media. Infatti gli errori standard presentano in media valori più contenuti, (circa il 20%) se confrontati con quelli ottenuti assumendo le altre ipotesi.

4. CONCLUSIONI

La ricerca illustrata descrive un approccio metodologico che ingloba nella calibrazione delle SPFs l'andamento degli incidenti osservati in un arco temporale esteso; in particolare, lo studio ha previsto l'elaborazione dei dati di incidentalità registrati in corrispondenza di rotatorie urbane. L'esempio mostra che l'uso delle equazioni GEE per la stima dei parametri del modello nell'ipotesi di indipendenza per la matrice di correlazione può condurre a risultati poco significativi.

La ricerca della reale struttura di correlazione attraverso la procedura GEE e la valutazione della bontà della stima con criteri appropriati (*Quasi-Likelihood Independence Criterion* - *QIC*) possono migliorare la precisione delle stime ed influenzare positivamente l'affidabilità delle previsioni di sicurezza (ad esempio quando il modello previsionale è utilizzato in una procedura di tipo Empirico-Bayesiano).

In base all'ampiezza dell'intervallo di confidenza calcolato attraverso il modello GEE, oltre al criterio informatore di Pan, è stato formulato un criterio aggiuntivo per la selezione del modello e per l'approssimazione della reale struttura di correlazione della variabile dipendente.

Il caso studio esaminato ha permesso di mettere in evidenza gli aspetti salienti della procedura suggerita ed i punti nodali da affrontare nella messa a punto di modelli di incidentalità basati su osservazioni protratte nel tempo.

5. BIBLIOGRAFIA

- Ballinger, G. A. (2004). Using Generalized Estimating Equations for Longitudinal Data Analysis. Organizational Research Methods, Vol. 7, No. 2, pp. 127–150.
- Diggle, P. J., P. Heagerty, K.-Y. Liang e S. L. Zeger (2002). Analysis of Longitudinal Data, 2nd ed. Oxford University Press, Oxford, United Kingdom.
- Fitzmaurice, G. M. A Caveat (1995). Concerning Independence Estimating Equations with Multivariate Binary Data. Biometrics, Vol. 51, pp. 309–317.
- Green, P. (1984) Iterative Reweighted Least Squares for Maximum Likelihood Estimation and Some Robust and Resistant Alternative (with discussion). Journal of the Royal Statistical Society, Series B, Vol. 46, pp. 149–162.
- Hauer, E. e J. Bamfo (1997). Two Tools for Finding What Function Links the Dependent Variable to the Explanatory Variables. Presented at International Cooperation on Theories and Concepts in Traffic Safety Conference, Lund, Sweden, 1997.
- Hauer, E. (2002) Observational Before and After Studies in Road Safety. Pergamon.
- Hardin, J. J. e J. M. Hilbe (2003). Generalized Estimating Equations. Chapman & Hall/CRC, Boca Raton, Flo (USA).
- Kullback, S. e R. A., Leibler (1999). On information and sufficiency. Annals of Mathematical Statistics, Vol. 22, 1951, pp. 79-86
- Liang, K.-Y. e S. L. Zeger (1986). Longitudinal Data Analysis Using Generalized Linear Models. Biometrika, Vol. 73, pp. 13–22.
- Lord, D., E. Hauer, e J. Bamfo (1999). Application de Deux Nouvelles Méthodes pour Examiner la Relation entre les Accidents et les Variables Explicatives (in French). Routes et Transports, Vol. 28, No. 3, pp. 11–20.
- Lord, D. e B. N. Persaud (2000). Accident Prediction Models With and Without Trend: Application of the Generalized Estimating Equations Procedure. In Transportation Research Record: Journal of the Transportation Research Board, No. 1717, TRB, National Research Council, Washington, D.C., pp. 102–108.
- Lord, D. (2000). The Prediction of Accidents on Digital Network: Characteristics and Issues Related to the Application of Accident Prediction Models. PhD thesis. Department of Civil Engineering, University of Toronto, Toronto, Ontario, Canada.
- Lord, D., A. Manar e A. Vizioli (2005). Modeling Crash-Flow-Density and Crash-Flow-V/C Ratio Relationship for Rural and Urban Freeway Segments. Accident Analysis and Prevention, Vol. 37, pp. 185–199.
- Mauro R. e Corradini M. (2008). Una indagine di incidentalità su incroci a rotatoria in ambito urbano. XVII Convegno Nazionale SIIV, Enna 10-12 Settembre, 2008.
- Pan, W. (2001) Akaike's Information Criterion in Generalized Estimating Equations. Biometrics, Vol. 57, 2001, pp. 120–125.
- Payne, R. W., S. A. Harding, D. A. Murray e D. M. Soutar. (2004). GenStat Release 8. VSN International, Oxford, United Kingdom.
- Zeger, S. L., e K.-Y. Liang (1986b). Longitudinal Data Analysis for Discrete and Continuous Outcomes. Biometrics, Vol. 42, pp. 121–130.
- Zeger, S. L., K.-Y. Liang e P. S. Albert (1988). Models for Longitudinal Data: A Generalized Estimating Equation Approach. Biometrics, Vol. 44, pp. 1049–1060.

– Zheng, B. (2000). Summarizing the Goodness of Fit on Generalized Linear Models for Longitudinal Data. Statistics in Medicine, Vol. 19, 2000, pp. 1265–1275.