ANALISI DI SERIE STORICHE DI DATI DA SPGPS

Manuele PESENTI (*), Marco PIRAS (*), Marco ROGGERO (*)

(*) Politecnico di Torino, C.so Duca degli Abruzzi, 24 10129 - Torino
tel: 011 564 7719 / 7675 / 7630, *fax: 011 564 7699*
manuele.pesenti@polito.it, marco.piras@polito.it, roggero@atlantic.polito.it

Riassunto
A partire dagli anni novanta, l'introduzione dei sistemi di riferimento (SR) globali e di elevata precisione, ha portato ad una evoluzione dello stesso concetto di SR. L'introduzione della quarta coordinata temporale è stata conseguenza sia delle tecniche di misura di geodesia spaziale, le quali coinvolgono punti anche all'esterno della superficie terrestre (VLBI, LLR, SLR, DORIS e GPS), sia dell'elevata precisione dei risultati da queste ottenuti, incompatibile con le deformazioni che necessariamente ogni vertice della materializzazione del *datum* accumula nell'arco degli anni. Al fine di valutare spostamenti e velocità dei vertici che costituiscono un *frame*, occorre effettuare una analisi di serie storiche di soluzioni periodiche su un adeguato *set* di dati. Questa analisi ha, in ambito geodetico, un triplice interesse. Essa rappresenta un utile strumento nella definizione del SR, nell'analisi di deformazioni, nonché, nella stima della ripetibilità per il controllo di qualità delle soluzioni.
Lo studio svolto è mirato alla stima di regressioni di serie storiche, in particolare viene studiata l'individuazione di discontinuità, spesso non note a priori, e l'identificazione di *cluster* internamente coerenti rispetto alle condizioni di osservazione. Sono valutati, inoltre, i vantaggi dell'utilizzo di tecniche robuste finalizzate all'esclusione degli *outliers*.

Abstract
From the early 90s, introduction and use of high precision global reference systems, and their continuous evolution, brought to the revolution of main concept of datum *introducing the 4th temporal coordinate. This was due both to the use of space geodesy techniques (VLBI, LLR, SLR, DORIS and GPS), which involve in measurements points outside of the earth surface, and to the high precision results obtained, that are not consistent with deformations piled up during few years from vertexes of a datum materializations. With the aim of controlling deformations of the vertexes of a frame, a time series analysis of a suitable list of periodic solutions is necessary. This kind of analysis of coordinates in time has three different rules in geodesy. It is a useful tool for datum definition, but it can be also used for deformations analysis and for repeatability evaluation for solutions quality control. What we propose here is a study on techniques for zero level discontinuities detection, in order to identify clusters internally consistent respect to measurement conditions in time series. Robust techniques for outlier detection are also discussed.*

Introduzione

Nell'applicazione di tecniche di geodesia spaziale, l'analisi di serie temporali di soluzioni di coordinate è uno strumento importante sia a monte che a valle del processo di compensazione delle misure grezze. In particolare trova applicazione nelle tecniche di posizionamento satellitare, oggi di largo utilizzo, data l'elevata densità di distribuzione sulla superficie terrestre di reti di stazioni permanenti (SP) GNSS sia di ambito locale o regionale (per RTK o monitoraggio) che continentale o globale (per la materializzazione di *datum*). Di queste reti, in molti casi, sono disponibili pubblicamente, oltre ai dati di osservazione, anche le soluzioni periodiche delle coordinate dei vertici attraverso i siti

internet dei diversi enti gestori (ne sono un esempio l'IGS: ***I**nternational **G**NSS **S**ervice* e l'EPNCB: ***E**UREF **P**ermanent **N**etwork **C**entral **B**ureau*, per citare solo i principali).
I risultati delle serie periodiche subiscono, però, l'influenza di eventi di disturbo che intervengono sul rilievo della posizione dei vertici. Sono un esempio gli interventi sulla strumentazione di rilievo, quali sostituzioni *hardware*, che possono influire sulla localizzazione del centro di fase, oppure effetti imputabili a dove il punto è materializzato, come le deformazioni stagionali o le instabilità locali dei siti. Si possono, inoltre, citare anche quelle componenti correlate alle diverse possibili modellazioni delle orbite satellitari o delle componenti atmosferiche in fase di compensazione.
Per questo, sia nella fase di inquadramento che in quella di monitoraggio ed analisi dei risultati finali di posizionamento, è utile identificare e quantificare le componenti di deformazione e velocità dovute a fenomeni locali, per una corretta interpretazione. L'analisi temporale delle serie di soluzioni ricopre un ruolo importante anche nella definizione della precisione degli stessi risultati finali attraverso la valutazione della ripetibilità. Nel caso, infatti, della stima ai minimi quadrati, applicata a metodi GNSS, la precisione formale, come noto, risulta sottostimata.
Un'analisi completa di serie temporali consta di diverse fasi successive, la prima delle quali è la ricerca delle discontinuità di grado zero. Questo lavoro è dedicato in particolare all'applicazione di metodi statistici finalizzati alla ricerca in automatico proprio di quei fenomeni, ed in particolare sono prese in considerazione discontinuità a gradino di tipo *level shift*. Nonostante, infatti, il tentativo di mantenere documentati tali salti da parte degli enti gestori di SP, non sempre questo avviene con regolarità e, d'altra parte, non risulterebbe comunque nota l'entità della discontinuità introdotta. Verranno nel seguito presentate e raffrontate tre diverse metodologie basate su approcci statistici differenti al fine di valutarne reciprocamente i vantaggi e l'applicabilità.

Metodi a confronto

1. Metodo della *forward search* modificato

La metodologia qui proposta prende spunto dal metodo della *forward search* (FS), metodo in uso in vari ambiti statistici applicati e finalizzato al raggruppamento di dati in base al proprio comportamento nel tempo. La FS può essere considerata un metodo ibrido che usa in fasi consecutive l'approccio robusto LMS (*Least Median of Squares*) e quello *least squares* (LS) o minimi quadrati.
Il concetto principale del metodo consiste nel servirsi di un sottoinsieme del *dataset* completo dei dati in ingresso, considerato internamente coerente, privo, cioè, di *outlier*, da usarsi per la definizione dei parametri di un primo modello ARIMA. L'intero numero di elementi del *dataset* viene quindi ordinato in base alla sua affinità col modello, che può essere stimata in base a diversi parametri quali la distanza di Mahalanobis (Atkinson et al., 2004), la distanza di Cook o più banalmente lo scarto. A questo punto il modello lineare LS viene stimato nuovamente sulla base anche delle osservazioni (ma tipicamente solo una) il cui parametro di analisi dell'affinità è risultato minore in modulo. La procedura viene reiterata fino a che tutte le osservazioni disponibili, passo dopo passo, non sono inglobate nel *subset*. È possibile quindi raggruppare le osservazioni, individuando comportamenti "simili" degli scarti rispetto ai diversi modelli via via calcolati.
La FS, di applicazione molto generale, non considera l'ipotesi particolare di correlazione temporale dei dati, che invece, nei casi dove verificata, come quello qui studiato, può essere molto utile. È stato messo a punto un algoritmo che prendendo le mosse dalla FS considerasse, nelle sue fasi di ricerca, le ipotesi di sequenzialità temporale delle soluzioni, appunto, e di non correlazione tra gruppi (o *cluster*) di soluzioni non consecutivi. Nella procedura si rispecchiano i tre elementi fondamentali propri della FS:

1. un criterio "robusto" per la cernita di un *subset* di osservazioni privo di *outlier*;
2. un criterio di avanzamento della ricerca;
3. un insieme di metodi di monitoraggio della ricerca.

Scelta del subset iniziale

Per operare la scelta del *subset* iniziale con metodi numerici, viene applicata, su un numero di soluzioni sufficientemente piccolo, scelto arbitrariamente a priori, una prima fase di "pulitura". La fun-

zione implementata a tale scopo sfrutta la caratteristica di robustezza del metodo LMS per individuare gli *outlier* e, grazie all'elevato valore di *breakdown point* (50%), prendere a riferimento solo il *cluster* che interseca il *subset* per più della metà dei propri elementi. Secondo i metodi qui di seguito descritti, il riconoscimento degli *outlier* avviene in base ai residui previsti, che, se ricavati da un *subset* troppo selezionato, rischiano di essere sottostimati il che comporta una eccessiva sensibilità al rumore del metodo. Da queste considerazioni, e da prove pratiche, è risultato opportuno usare un test con grado di significatività pari allo 0.3%. Lo scarto (σ*) (Rousseeuw et al., 1987) rispetto al modello LMS viene definito in maniera robusta attraverso una sua stima preliminare S_0 utile per la definizione di una funzione dei pesi ω_i secondo le formule:

$$S_0 = 1.4826\left(1+\frac{5}{n-p}\right)\sqrt{med(r_i^2)} \qquad \omega_i = \begin{cases} 1 & se \quad \left|\frac{r_i}{S_0}\right| \le 2.5 \\ 0 & altrimenti \end{cases} \qquad \sigma^* = \sqrt{\frac{\sum_{i=1}^{n}\omega_i r_i^2}{\sum_{i=1}^{n}\omega_i - p}} \qquad [1]$$

nelle quali *n* è il numero di elementi considerati, r_i lo scarto dell'*i*-esimo elemento della serie, *p* il numero dei coefficienti calcolati per il modello e *med* rappresenta la funzione mediana. A questo punto la selezione degli elementi per la definizione del modello rappresentante il *cluster* può essere effettuata, secondo l'impostazione di Neyman, in base alla disuguaglianza: $\left|\frac{r_i}{\sigma^*}\right| < 3$

Monitoraggio dei parametri statistici di discriminazione della discontinuità

La discriminazione dell'appartenenza o meno di ogni elemento al *cluster* in esame avviene in base a test statistico sul rapporto tra lo scarto del valore misurato (S_r) rispetto a quello previsto (S_p) in base ad una regressione lineare (vedi *Figura 1*). In questo caso è possibile calibrare la sensibilità del test in base all'intervallo di confidenza desiderato (per es. 95%).

Se a questo punto la soluzione in esame viene considerata facente parte dello stesso *cluster* di quelle antecedenti si ricrea un modello lineare agli LS comprendendo anche la soluzione appena esaminata, e si procede ad analizzare la successiva. In questo modo, mano a mano che la procedura va avanti, l'analisi avviene in base ad un modello con sempre maggior ridondanza e ricavato un *set* controllato di dati. Nelle prove preliminari effettuate su serie sintetiche di dati, e con i valori di sensibilità dei test visti in precedenza, si è rilevato una consistente percentuale di successo, di poco inferiore al 50%, per valori di salto imposti inferiori a due volte la rumorosità. In tali casi il valore del salto è risultato ben stimato, mentre nei casi di erronea identificazione il suo valore è risultato inferiore al rumore per cui l'eventuale ricucitura della serie non ne avrebbe comunque inficiato il significato.

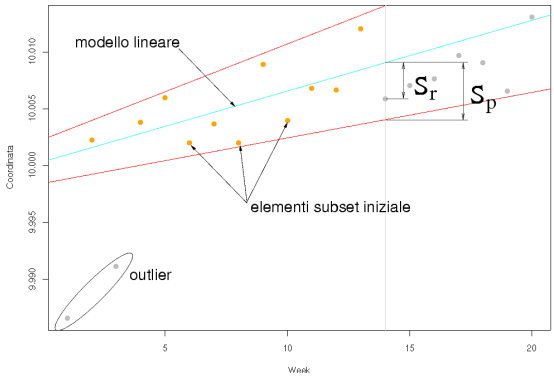

Figura 1: Dati grezzi: in arancione gli elementi del subset iniziale.

2. Metodo della finestra mobile robusta

Un altro approccio proposto, ancora di tipo sequenziale, per l'individuazione di salti non documentati nelle serie temporali è la "finestra mobile robusta". Questo metodo applica lo stimatore LMS che consente di ottenere una soluzione anche con il 50% di *outlier* presenti nel *dataset* a discapito, però, della precisione. Scelta l'ampiezza della finestra espressa in epoche (ad esempio 15-20), si effettua per ogni singolo passo una stima dei parametri con il metodo LMS. Per ogni *step*, vengono effettuate due tipologie di analisi: un test sul residuo ed un test sul *ratio* delle varianze.

Test residui

Il primo test che si effettua è un confronto tra i residui delle osservazioni stimate e di quelle reali. Ponendo un valore ε di errore d'osservazione e fissata una soglia di confronto pari a 3ε, si determinano per ogni finestra i valori che oltrepassano tale soglia, i quali vengono considerati *outlier* se non ripetuti oppure cambiamenti di stato se seguiti da fenomeni analoghi allo *step* successivo. Considerando il caso test definito in *Figura 2*, i residui sono definiti in *Figura 3*.

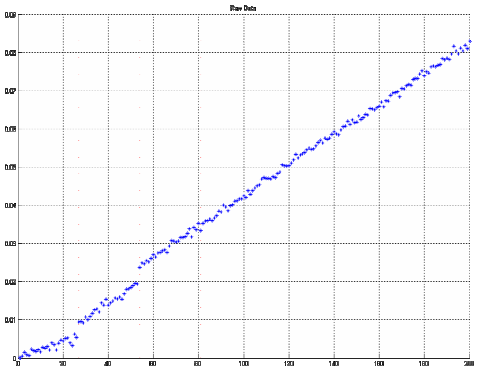

Figura 2: Dati Grezzi

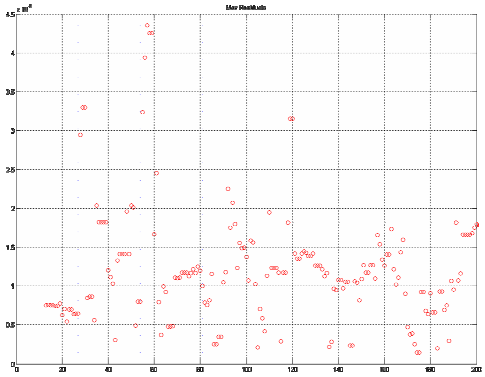

Figura 3: Residui

I valori eccedenti alla soglia sono considerati *jump*. In questo caso piccoli *jump* rischiano di non essere intercettati in quando il test non è troppo potente.

Test ratio RMSE

Congiuntamente al test sui residui LMS, si procede con un altro basato sullo stimatore LS pesato, i cui valori di RMSE vengono messi in relazione con quelli da LS normale.

Si sono considerate due diverse tipologie di *weighted* LS, derivate da due differenti modelli di pesi: modello di Huber e modello di Cauchy. I due modelli si distinguono essenzialmente per le diverse formulazioni della funzione peso e della la *tuning constant* riassunti nella tabella che segue.

Modello	Funzione peso	*Tuning constant*
Huber	$w = (\max(1, abs(r))^{-1}$	1.345
Cauchy	$w = (1 + r^2)^{-1}$	2.385

Tabella 4: Definizione dei pesi

dove il termine *r* è definito da:

$$r = \frac{v}{(\mathrm{TC} \cdot \mathrm{s} \cdot \sqrt{(1-h)})} \quad [2]$$

essendo TC la *tuning constant*, h il vettore dei leveraggi definito da LS ed $s = \frac{MAD(v)}{0.675}$,

dove *MAD* è la *median absolute deviation* e il termine a numeratore definisce la stima senza *bias* di una distribuzione normale.

Considerando la serie temporale definita in *Figura 2*, ed analizzando per ogni *step* il valore del ratio, si ottiene il risultato della *Figura 5*.

Da alcune prove condotte, si può ritenere che i valori esterni all'intervallo 0.5-1.5 sono riconducibili ad errori grossolani [caso a] o, nel caso si susseguano con la stessa intensità, alla presenza di un *jump* [caso b].

Questo tipo di test ha dimostrato di essere

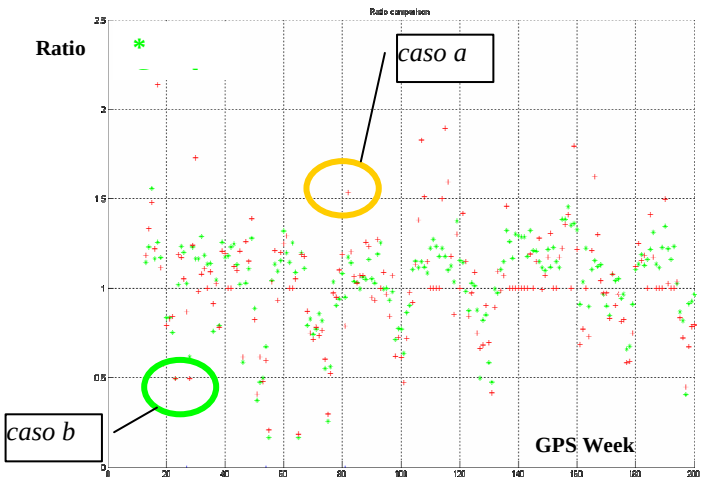

Figura 5: Test Ratio RMSE

abbastanza sensibile anche a piccole variazioni (< 3ε), consentendo, se abbinato al test sui residui, di intercettare cambiamenti anche di modeste entità.

3. Stima minimi quadrati con imposizione di vincoli dinamici

Consideriamo un sistema dinamico lineare discreto descritto da un vettore di stato finito dimensionale *x* e da un vettore di *bias b* costanti. Il sistema evolve con una dinamica nota nelle epoche *t* (*t* = *1, …, T*):

$$\begin{aligned} x_{t+1} &= T_{t+1}x_t + v_{t+1} \\ y_{t+1} &= H_{t+1}x_{t+1} + C_{t+1}b_{t+1} + \varepsilon_{t+1} \end{aligned} \qquad [3]$$

dove *y* sono le osservazioni. In relazione all'applicazione che stiamo esaminando, l'analisi di serie storiche di coordinate, è opportuno notare che sia il vettore di stato *x* che il vettore delle osservazioni *y* rappresentano coordinate. Le coordinate osservate *y* sono il prodotto dell'elaborazione di osservazioni GNSS, derivante ad esempio da una compensazione di rete, mentre le coordinate stimate *x* tengono conto di un modello funzionale e stocastico della dinamica del sistema. Tale modello tiene conto della correlazione temporale delle coordinate osservate. Il sistema conduce ad una matrice normale:

$$N = D^T W_\omega D + M^T W_\omega M \qquad [4]$$

che include il contributo della dinamica e della geometria, come descritto in (Albertella et al., 2006) e (Roggero, 2006). Il modello dinamico è descritto, epoca per epoca, dalla matrice di transizione T e dalla matrice generale D, con la matrice dei pesi W_ω. Il contributo della geometria è dato, epoca per epoca, dalla matrice disegno H e dalla matrice globale M, con la matrice di peso W_ε. Si noti come le matrici peso dipendano dalle matrici di covarianza del rumore di sistema e degli errori di misura.
Il vettore *b* può essere costante o costante a tratti. Ciò avviene nel caso in cui le osservazioni siano affette da discontinuità, come nel caso di serie temporali di coordinate in cui siano presenti eventi, documentati o no, quali la sostituzione dell'antenna oppure rapidi movimenti crostali durante un terremoto. La posizione di tali discontinuità nella serie di osservazioni è rappresentata dalla matrice *C*, i cui elementi possono assumere valore 0 o 1. Tale matrice, che collega il vettore dei *bias* alle osservazioni, può essere nota a priori se rappresenta eventi documentati, o determinata per mezzo di un qualche algoritmo se rappresenta invece eventi non documentati. Vedremo come costruire tale matrice sfruttando il vincolo stocastico sulla dinamica del sistema, determinando infine posizione ed ampiezza delle discontinuità in serie di osservazioni simulate.
Ipotizzando che il vettore delle osservazioni *y* sia distribuito normalmente con matrice di varianza-covarianza Q_y, possiamo formulare l'ipotesi nulla, in assenza di discontinuità, e un certo numero di ipotesi alternative in presenza di discontinuità:

$$H_0 : \begin{cases} E(y) = Hx \\ D(y) = Q_y \end{cases} \quad H_A : \begin{cases} E(y) = Hx + Cb \\ D(y) = Q_y \end{cases} \qquad [5]$$

La stima di queste due ipotesi fornirà anche le rispettive varianze dell'unità di peso $\hat{\sigma}_0(0)$ e $\hat{\sigma}_0(A)$.
Possiamo formulare un'ipotesi alternativa per ogni epoca di osservazione, e verificare l'adeguatezza del modello per mezzo di un test sulla variabile *ratio*, che com'è noto ha distribuzione χ^2:

$$\frac{\chi^2_{\left(\frac{\alpha}{2}\right)}}{n-r} < \frac{\hat{\sigma}_0(A)}{\hat{\sigma}_0(0)} < \frac{\chi^2_{\left(1-\frac{\alpha}{2}\right)}}{n-r} \qquad [6]$$

Se il *ratio* supera uno dei valori limite, l'ipotesi iniziale va scartata, ed accettata l'ipotesi alternativa della presenza di una discontinuità nell'epoca considerata. Tale test è eseguito per ogni epoca di osservazione, e richiede la stima di altrettanti modelli alternativi. Terminata tale fase di test siamo in grado di compilare la matrice *C*, integrandola eventualmente con informazioni note a priori (discontinuità documentate). Nota *C*, perveniamo infine alla stima di *x* e del vettore *b*, che rappresenta le ampiezze delle discontinuità.

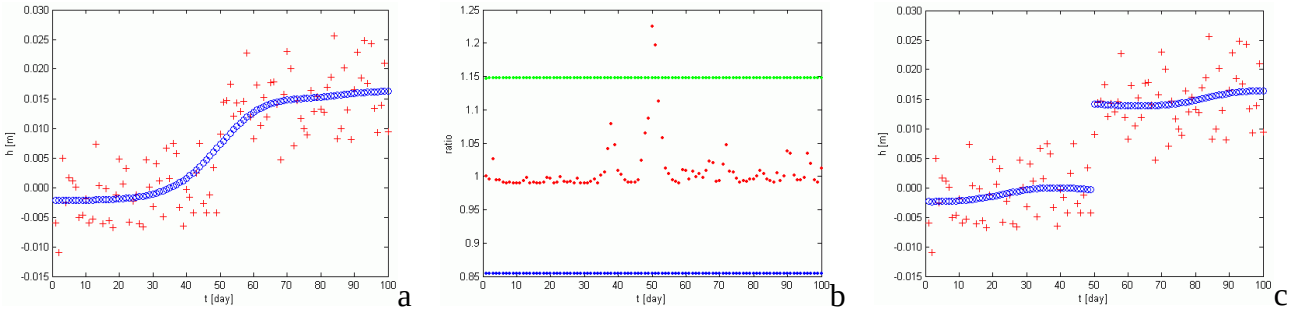

Figura 6: Serie di 100 coordinate sintetiche di una stazione GNSS permanente; σh = 0.005 m e discontinuità di 0.015 m al giorno 50. Legenda: + coordinate osservate, ○ coordinate stimate. (a) Ipotesi H0, nessuna discontinuità; le coordinate stimate sono lisciate.(b) Ratio: il limite superiore è oltrepassato in due casi contigui, la discontinuità è localizzata in corrispondenza del massimo tra i due. (c) Ipotesi HA, stima in presenza della discontinuità individuata.

Conclusioni

L'analisi delle serie temporali consente di poter individuare eventuali anomalie non documentate, che se non rimosse possono condurre ad errate interpretazioni. I metodi di analisi, come si è visto, non sono univoci, e per la loro natura (funzione obiettivo usata, fattori discriminanti, ecc.) hanno prerogative diverse. In linea di massima risultano però efficaci per intercettare discontinuità superiori a circa 1.5-2 σ_l. È in fase di conclusione uno studio comparativo dei tre metodi illustrati precedentemente per studiare in maniera completa i loro limiti di comportamento in riferimento a diverse distribuzioni di serie temporali.

Bibliografia

A. C. Atkinson, M. Riani, A. Cerioli (2004), "Exploring multivariate data with the forward search", Springer, London

T. Kailath, A. H. Sayed, B. Hassibi (2000), "Linear estimation", Prentice Hall, New Jersey.

Y. Saad (2000), "Iterative methods for sparse linear systems", Second edition with corrections.

P. Teunissen (2001), "Dynamic data processing", *Delft University Press*.

I. Colomina, M. Blàzquez (2004), "A unified approach to static and dynamic modeling in photogrammetry and remote sensing", Altan, O. (ed.), *Proceedings of the XXth ISPRS Congress*, Istanbul, 178-183.

A. Albertella, B. Betti, F. Sansò, V. Tornatore (2006), "Real time and batch navigation solutions: alternative approaches", *Bollettino SIFET*, n. 2 - 2006.

N. Perfetti, "Detection of station coordinates discontinuities within the Italian GPS Fiducial Network", *Journal of Geodesy*, Springer Berlin / Heidelberg ISSN 0949-7714 (Print) 1432-1394 (Online) – 2006.

M. Roggero, "Kinematic GPS batch processing, a constrained solution applied to antenna array", *Reports on Geodesy* n. 2 (77), pp. 235-240, ISSN 0867-3179 – 2006.

M. Roggero, "Kinematic GPS Batch Processing, improving ambiguity fixing performances", *Reports on Geodesy* n. 2 (77), pp. 227-234, ISSN 0867-3179 – 2006.

M. Roggero, "Kinematic GPS batch processing, a source for large sparse problems", Hotine Marussi Symposium on Theoretical and Computational Geodesy, Wuhan 2006.

P. J. Rousseeuw, A.M. Leroy (1987), "Robust regression and outlier detection", Wiley & sons, New York