Big Data...a few Outliers = Big Mistakes

Un nuovo processo per l'individuazione di outliers

di Maurizio Rosina

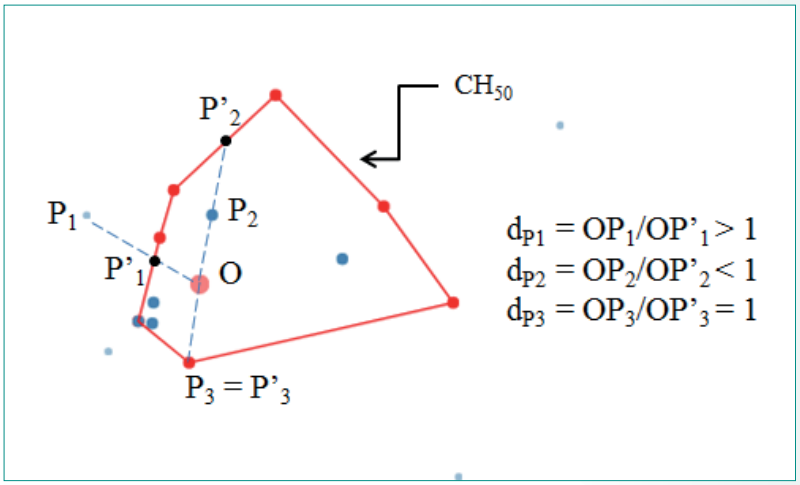

Fig. 1 – calcolo del fattore di distanza dei punti dal baricentro robusto sulla base del CH_{50} .

La tecnologia ci mette nelle condizioni di potere e dovere maneggiare grandi moli di dati. Nuvole di punti acquisite nei modi più vari e Big Data sono le parole d'ordine e le realtà con cui oggigiorno sempre più occorre misurarsi.

Nell'elaborazione dei dati sempre più spesso entrano in gioco la stima/calcolo di parametri statistici quali la media, la varianza, lo scarto quadratico medio, ecc. Ebbene è noto che bastano pochi outliers (ovvero pochi valori anomali, aberranti, chiaramente distanti dagli altri valori disponibili) per 'mettere in crisi' medie, scarti quadratici medi ed … altro, con il risultato di giungere a risultati finali definiamoli perlomeno 'fuorvianti'. La tematica dell'individuazione degli outliers assume quindi la massima importanza per poter giungere a risultati quanto più possibile corretti e significativi. Occorre quindi sempre propedeuticamente ricercare gli eventuali valori anomali - che talvolta assumono persino il ruolo del 'risultato' cercato, e ciò proprio in ragione della loro 'anomalia' che li differenzia dal resto dei dati -, e con tecniche che quanto più possibile non presuppongano una conoscenza *a priori* della tipologia di distribuzione che i dati in esame dovrebbero avere. Il nuovo approccio ideato per l'individuazione di outliers nello spazio R^2 fruisce di tecniche geometrico/statistiche largamente indipendenti dalla tipologia di distribuzione dei dati, e si articola in quattro passi metodologici. Data una nuvola di punti nello spazio R^2:

1. Individuazione dei vari cluster di punti e dei punti che non appartengono a nessuno dei cluster individuati.

Per ciascun cluster

2. Individuazione, tramite la tecnica del convex hull peeling, del particolare convex hull (CH_{50}) che al suo interno contiene non più del 50% dei punti del cluster, e calcolo su tali punti interni (che sono il 'core' del cluster) del baricentro (ora robusto) tramite una operazione di media.
3. Utilizzo di una tecnica di mapping (che realizza una nuova metrica) che porta tutti i punti che giacciono sul CH_{50} a trovarsi ad un fattore di distanza pari ad uno dal baricentro, che è come dire che il CH_{50} viene ad assume la forma di un cerchio con centro nel baricentro e raggio pari ad uno. Tale tecnica, illustrata più in dettaglio nel seguito, è applicata a tutti i punti del

cluster ed ai punti che sono risultati non appartenere ad alcuno dei cluster individuati. Con questa tecnica tutti i punti strettamente contenuti in CH_{50} avranno un fattore di distanza dal baricentro minore di uno e tutti i punti esterni al CH_{50} avranno un fattore di distanza dal baricentro maggiore di uno. A seguito di questa tecnica si potrà nel seguito operare su tali valori di distanze (ovvero su sequenze di valori - dati univariati - nello spazio R^1) e non più su coordinate nello spazio R^2.

4. Utilizzo, sulle distanze calcolate nel passo precedente, della diseguaglianza di Chebychev (valida per una qualsiasi tipologia di distribuzione univariata di valori). La distribuzione di Chebychev garantisce che per una distribuzione qualsiasi di valori, una volta calcolata la sua media (μ), il suo scarto quadratico medio (σ) e fissata una costante k >0, al massimo lo $[(1/k^2)*100]$% dei valori potranno risultare esterni all'intervallo μ-kσ, μ+kσ. Ciò permette, su base statistica, di definire in modo 'fine' come outlier un qualsiasi punto la cui distanza dal baricentro ricada all'esterno dell'intervallo μ-k, μ+kσ Molto spesso nell'utilizzo della diseguaglianza di Chebyschev piuttosto che fissare la costante k si preferisce fissare un valore di probabilità (p), in quanto per una distribuzione unimodale di valori tra k e p sussiste la relazione $p = 1/k^2$, quindi fissato p è immediato risalire al relativo $k = \sqrt{(1/p)}$. La ricerca degli outliers viene quindi condotta, per ciascun cluster, individuando come outliers i punti, sia del cluster che non appartenenti a nessun cluster, le cui distanze dal baricentro del cluster siano esterne all'intervallo sopra definito.

Tramite i quattro passi metodologici sopra sommariamente descritti si giunge, senza alcuna ipotesi preventiva sulla tipologia di distribuzione dei dati, a poter individuare la presenza di eventuali outliers rispetto ai vari cluster individuati.
Inoltre, è di tutta evidenza che l'approccio proposto è teoricamente facilmente espandibile a dati nello spazio R^3, con i vari convex hull che potrebbero assumere la struttura di politopi di minima chiusura convessa di punti nello spazio R^3.

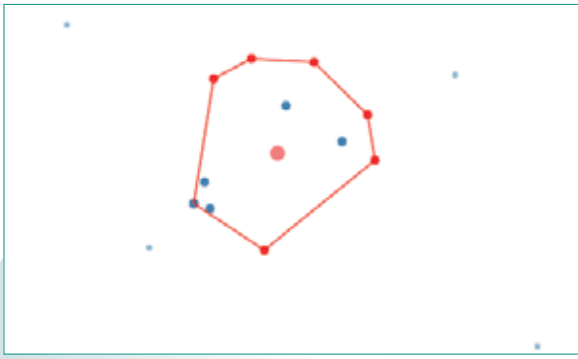

Fig. 2 – il mapping dei punti sulla base dei nuovi fattori di distanza calcolati - si noti in particolare come i punti del CH_{50} cui viene attribuito il nuovo fattore di distanza giacciano su di un cerchio ideale (di raggio 1) dal baricentro.

Il dettaglio delle operazioni

L'approccio perseguito è altamente modulare, ed è quindi opportuno fornire qualche dettaglio circa le operazioni condotte nei passi sopra elencati. Il passo 1 non impone alcun specifico metodo nella individuazione dei cluster, tanto che, se ritenuto opportuno, tale passo può persino essere saltato, vedendo la nuvola dei punti in esame come un unico cluster, su cui operare con i passi successivi. Inoltre, nel caso di analisi condotte su dati originali univariati, già il solo passo 4, saltando tutti i precedenti, risulterebbe teoricamente sufficiente all'individuazione di eventuali outliers.
In merito al passo 2 il baricentro calcolato sui punti strettamente contenuti nel CH_{50} risulta essere particolarmente robusto, ed è assimilabile all'analogo calcolo spesso condotto su dati appartenenti al secondo e terzo quartile di un boxplot.
La tecnica utilizzata nel passo 3 è particolarmente interessante, in quanto permette di tenere conto della 'forma' assunta dai punti del cluster nella successiva valutazione/individuazione degli outliers, che opera sulla base di un particolare 'fattore di distan-

Figura 3 – il campione dei dati - 3243 coordinate relative a localizzazioni di POI presenti nelle province di Viterbo e Latina.

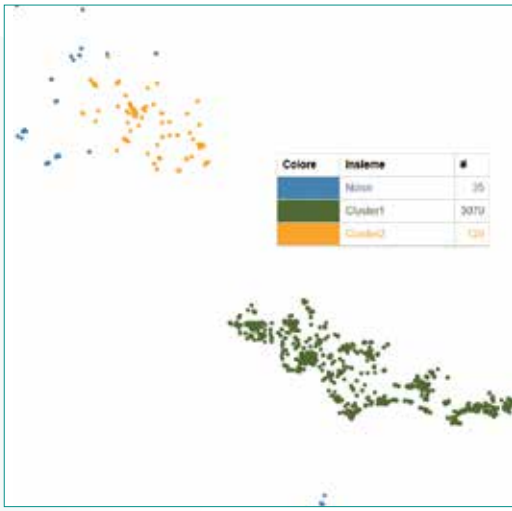

Fig. 4 – i due cluster individuati (algoritmo Dbscan) ed i punti (noise) non assegnabili dall'algoritmo ad alcuno dei due cluster.

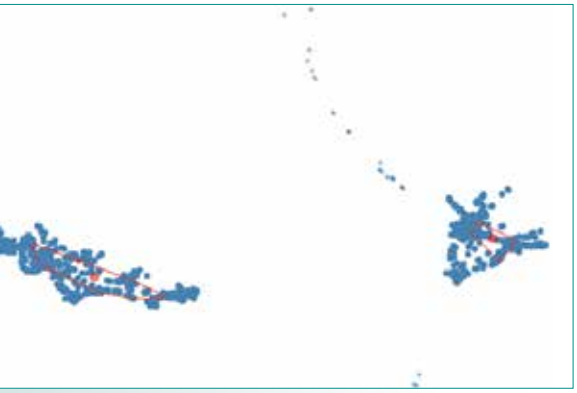

Fig. 5 – primo cluster – a sinistra il cluster con evidenziato il suo CH_{50} ed il baricentro ricavato dai punti strettamente contenuti nel CH_{50}. A destra l'immagine del mapping attivato sui punti del cluster e su quelli che non appartengono a nessun cluster. Le distanze dal baricentro dei punti della figura a destra saranno l'oggetto dell'analisi volta all'individuazione di outliers che verrà condotta nella fase successiva tramite l'utilizzo della diseguaglianza di Chebyshev.

za' di ciascun punto rispetto al baricentro (robusto) del cluster. Nello stimare se un punto appartiene o no ad un determinato cluster è pratica generale assumere che più un punto è vicino al baricentro (robusto) del cluster, più è verosimile che il punto appartenga al cluster. Bisogna però anche tenere in considerazione se l'insieme dei punti del cluster è distribuito simmetricamente o asimmetricamente rispetto al baricentro, per poter decidere se la valutazione di una distanza dal baricentro è significativa per l'individuazione di un outlier. Occorre, quindi, tenere conto della 'forma' del cluster nel calcolo del 'fattore di distanza' del generico punto dal baricentro (robusto). La tecnica adottata per il calcolo di un fattore di distanza che tenga in conto tale 'forma' è particolarmente semplice. Detto O il baricentro (robusto) del cluster, per il generico punto P si calcola il punto di intersezione P' con il CH_{50} generato dalla semiretta con origine O e passante per P, quindi si calcola il rapporto dp=OP/OP', che assume il significato di *fattore di distanza* del punto P dal baricentro O. Risulta di tutta evidenza che qualsiasi punto P che giaccia sul CH_{50} avrà un fattore di distanza dp = 1, ovvero un fattore di distanza unitaria dal baricentro; viceversa qualsiasi punto P strettamente contenuto entro il CH_{50} avrà un fattore di distanza dp < 1, e qualsiasi punto esterno al CH_{50} avrà un fattore di distanza dp >1 (vedi figura 1).

E' come se ogni punto venisse rimappato, con il nuovo fattore di distanza, sulla semiretta che lo congiunge al baricentro robusto (vedi figura 2). Quindi tale tecnica definisce una metrica sull'insieme dei punti, ed il fattore di distanza dal baricentro permette di poter tenere conto della 'forma' del cluster, che si assume sia definita dalla forma del $CH_{50.}$ Un po' quanto è ottenibile, ma in modo assai più complesso, tramite la distanza di Mahalanobis, che però opera correzioni basandosi esclusivamente su forme strettamente ellissoidali.

Nel passo 4 si è voluta seguire la tecnica di utilizzare, su ciascun cluster, in sequenza due volte la disuguaglianza di Chebyshev con parametri diversi. Per ciascun cluster inizialmente vengono calcolate media μ e scarto quadratico medio σ di tutti i suoi valori (valori che sono le distanze dei punti del cluster dal baricentro, ottenute nel passo precedente) e viene applicata la diseguaglianza di Chebyshev con un fattore k piuttosto basso (oppure, data la relazione k = che sussiste tra $k = \sqrt{(1/p)}$ e la probabilità p, imponendo un valore di p piuttosto alto), che si traduce nel selezionare, come valori 'core/fondamentali' sicuramente appartenenti al cluster, solo quelli molto prossimi alla media (ovvero valori a distanza di pochi kσ rispetto alla media μ). Su tali valori 'core/fondamentali' si calcolano nuovamente media μ_1 e scarto quadratico medio σ_1 (che ora si ritengono parametri molto rappresentativi e robusti) e si effettua la effettiva ricerca degli outliers, ancora tramite la diseguaglianza di Chebyshev in cui si utilizzano i nuovi e robusti valori μ_1 e σ_1, ed si applica con un fattore k più alto del precedente (o, il che è dire lo stesso, un fattore p più piccolo del precedente), ciò che si traduce nell'individuare come outliers solo valori molto distanti dalla media μ_1, ovvero valori che 'sicuramente' non appartengono al cluster. La ricerca degli outliers viene quindi effettuata sui tutti i valori del cluster e sui punti che sono risultati non appartenere ad alcun cluster, per questi ultimi calcolandone preventivamente il mapping rispetto al baricentro del cluster e le relative distanze dal baricentro. Modulando opportunamente i valori con cui attivare in sequenza le due diseguaglianze di Chebyshev si giunge ad una individuazione quanto si vuole 'fine' degli outiers, ottenuta in base a parametri statistici dichiarabili ed ad una metodologia indipendente dalla tipologia di distribuzione dei dati.

Fig. 6 – secondo cluster - a sinistra il cluster con evidenziato il suo CH_{50} ed il baricentro ricavato dai punti strettamente contenuti nel CH_{50}. A destra l'immagine del mapping attivato sui punti del cluster e su quelli che non appartengono a nessun cluster. Le distanze dal baricentro dei punti della figura a destra saranno l'oggetto dell'analisi volta all'individuazione di outliers che verrà condotta nella fase successiva tramite l'utilizzo della diseguaglianza di Chebyshev.

Un esempio su dati reali

Nel seguito viene presentato un esempio su di un campione di dati reali, relativo a 3243 coordinate relative a localizzazioni di POI presenti nelle province di Viterbo e Latina. Nella figura 3 viene presentato il campione dei dati da elaborare. La figura 4 propone i 2 cluster individuati (tramite il classico algoritmo DBscan) ed i punti (noise) che non risultano assegnabili dall'algoritmo a nessuno dei due cluster. Le figure 5 e 6 presentano, per ciascuno dei due cluster:, il CH_{50} ed il baricentro ricavato dai punti strettamente contenuti nel $CH_{50,}$ quindi la successiva fase di mapping, attivata sia sui punti del cluster che su quelli che non appartengono a nessun cluster. Tale mapping ha l'effetto di portare tutti i punti che giacciono sul CH_{50} ad un fattore di distanza pari ad uno dal baricentro, tutti i punti strettamente contenuti entro il CH_{50} ad un fattore di distanza dal baricentro minore di uno, ed infine tutti i punti esterni al CH_{50} ad un fattore di distanza maggiore di uno. Infine nelle figure e 7 e 8 sono riportati gli outliers individuati tramite la diseguaglianza di Chebychev applicata ai due cluster ed i punti noise fissando valori diversi per il parametro di probabilità.

Gli outlier proposti nelle seguenti figure 7 e 8 sono individuati tramite uno specifico processo che utilizza nell'analisi due volte la disuguaglianza di Chebyshev su ciascuno dei cluster individuati. Per ogni cluster individuato una prima volta la disuguaglianza di Chebyshev viene utilizzata per individuare/selezionare i valori 'core/fondamentali' del cluster, sui quali calcolare dei nuovi valori di media μ_1 e scarto quadratico medio σ_1 (che ottenuti in tal modo si ritengono parametri estremamente 'rappresentativi' del cluster e 'robusti'). Quindi la disuguaglianza di Chebyshev viene utilizzata una seconda volta, con i valori di media μ1 e scarto quadratico medio σ precedentemente calcolati, per l'effettiva individuazione degli outliers. In particolare, per ottenere quanto illustrato nella figura 5, nel primo utilizzo della diseguaglianza di Chebyshev, per ricavare i valori 'core/fondamentali' si è fissato quale valore di probabilità p = 0.3 (che corrisponde a fissare un valore di k =1.8), valore che assicura che almeno il 70% dei dati del cluster giacciono entro l'intervallo $[\mu-1.8\sigma..\mu+1.8\sigma]$, in cui μ e σ sono la media e lo scarto quadratico medio calcolati dei dati (i valori delle distanze dal baricentro) del cluster. Quindi si sono individuati i valori del cluster che ricadono entro tale intervallo e su di essi si è calcolata una nuova media μ_1 ed un nuovo scarto quadratico medio σ_1 (che ora si ritengono parametri estremamente 'rappresentativi' del cluster e molto 'robusti') e si è fissato quale nuovo valore di probabilità p =0.005, valore che corrisponde al fissare un k=14, e che assicura che al massimo il cinque per mille dei valori del cluster potrebbero giacere esternamente all'intervallo $[\mu_1-14\sigma_1..\mu_1+14\sigma_1]$, La ricerca degli outliers è quindi stata effettuata individuando i punti le cui distanze sono esterne a tale intervallo, e tale analisi è stata condotta per tutte le distanze dei punti del cluster e sulle distanze dei punti che non appartenevano ad alcun cluster (i punti noise), per questi ultimi calcolandone preventivamente il mapping rispetto al baricentro del cluster e le relative distanze dal baricentro. Tale processo, come detto, è condotto su ciascuno dei cluster individuati. Per ottenere quanto illustrato nella figura 6 si è si è fissato, analogamente al caso precedente, nel primo utilizzo della diseguaglianza di Chebyshev un valore di probabilità p = 0.3, mentre nel secondo utilizzo della diseguaglianza si è fissato un valore p =0.01, assai più lasco del precedente. In tal modo si è ottenuto che al massimo l'uno per cento dei valori del cluster poteva giacere esternamente all'intervallo $[\mu_1-10\sigma_1..\mu_1+10\sigma_1]$, e tale rilassamento nelle condizioni di verifica ha portato all'incremento degli outliers individuati.

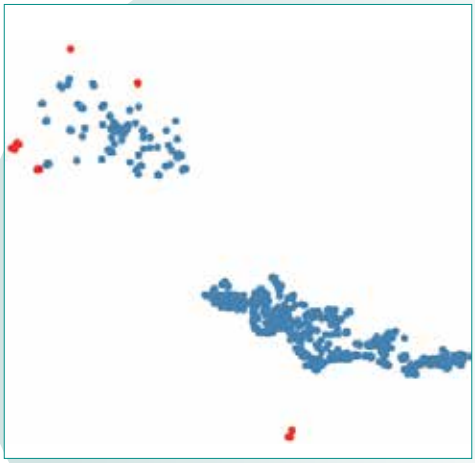

Fig. 7 – Gli agglomerati di punti in rosso corrispondono a 17 outliers individuati sul campione dei dati tramite l'utilizzo della diseguaglianza di Cherbychev applicata ai due cluster ed ai dati che non appartengono a nessun cluster.

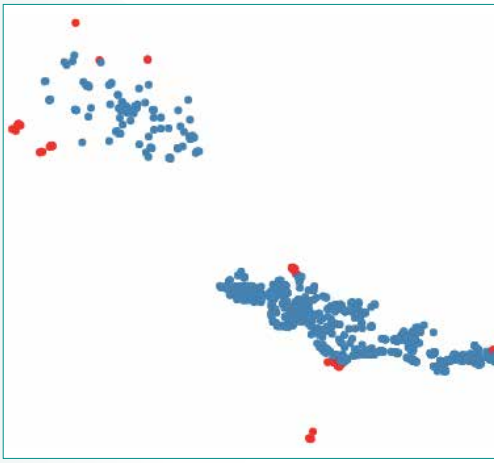

Fig. 8 – Gli agglomerati di punti in rosso corrispondono a 100 outliers individuati sul campione dei dati tramite l'utilizzo della diseguaglianza di Cherbychev applicata ai due cluster ed ai dati che non appartengono a nessun cluster.

Conclusioni

Molti si occupano di ottenere risultati elaborando dati e Big Data, molti meno si preoccupano di verificare se nei dati da elaborare sono presenti valori anomali ed aberranti (outliers). Se presenti anche in minime quantità gli outliers possono rendere assai poco consistenti i risultati delle elaborazioni. La ricerca e l'individuazione di outliers è quindi un passo fondamentale, generalmente propedeutico alle elaborazioni volte ad ottenere risultati consistenti. Il nuovo approccio ideato per la individuazione di outliers nello spazio R^2 fruisce di tecniche geometrico/statistiche largamente indipendenti dalla tipologia di distribuzione dei dati, e poggia su quattro pilastri metodologici: il clustering, la tecnica del convex hull peeling, una specifica metrica e la diseguaglianza di Chebyshev, che è valida per una qualsiasi tipologia di distribuzione univariata di valori. La modularità e la generalità dell'approccio, accoppiate alla ricerca ed alla individuazione di outliers in base a parametri strettamente statistici, fanno dell'approccio presentato un utile e quotidiano strumento per chi debba elaborare dati bivariati per gli scopi più vari, con la sicurezza di poter preventivamente verificare la eventuale presenza di outliers sulla base di specifici intervalli di confidenza.

Ringraziamenti

Un sentito ringraziamento va al collega dott. Andrea De Lullo, che ha implementato l'intero processo algoritmico e ne ha incrementato la flessibilità d'uso introducendo per l'utente la possibilità di scegliere tra più iniziali algoritmi di clustering.

BIBLIOGRAFIA

Amidan B. G., Ferryman T. A., Cooley S. K. (2005) *Data Outlier Detection using the Chebyshev Theorem*, IEEE Aerospace Conference Proceedings

Porzio G. C. & G. Ragozini (2000) Peeling multivariate data sets: a new approach, *Quaderni di Statistica*, Vol. 2

Ester M., Kriegel H-P., Sander J., Xu X. (1996) A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining.*

Riani M. & S. Zani (1998) Generalized Distance Measures for Asymmetric Multivariate Distributions, in A*dvances in Data Science and Classification: Proceedings of the 6th Conference of the International Federation of Classification Societies (IFCS-98),* Università "La Sapienza", Rome, 21–24 July, 503-508, Springer

Savage R., (1961) Probability Inequalities of the Tchebycheff Type, *Journal of Research of the National Bureau of Standards*, B. Mathematics and Mathematical Physics, Vol. 65B, No.3

Zani S., Riani M., Corbellini A. (1998), *Robust bivariate boxplots and multiple outlier detection,* Computational Statistics & Data Analysis, Elsevier

PAROLE CHIAVE

Outliers; convex hull peeling; clustering; diseguaglianza di Chebychev; scarto quadratico medio

ABSTRACT

The search and identification of outliers is a fundamental step, generally preparatory to the elaborations aimed at obtaining consistent results. The new approach devised for the identification of outliers in space R2 benefits from geometric / statistical techniques largely independent from the type of data distribution, and is based on four methodological pillars: clustering, the convex hull peeling technique, a specific metric and Chebyshev's inequality, which is valid for any type of univariate distribution of values. The modularity and the generality of the approach, coupled to the research and identification of outliers based on strictly statistical parameters, make the approach presented a useful and daily tool for those who need to process bivariate data with the security of being able to previously identify outliers.

AUTORE

Maurizio Rosina
mrosina@sogei.it
RLD – Ricerca e Laboratorio Digitale – Società Generale d'Informatica

STONEX

S800A

Oltre l'immaginazione

Ricevitore GNSS con 394 canali e alte prestazioni

atlas®

- **aRTK**, è in grado di continuare a generare posizioni precise fino a 20 minuti in caso di perdita del segnale RTK
- **SureFix**, fornisce posizioni RTK ad alta fedeltà anche in condizioni avverse
- **ATLAS**, servizio di correzione globale GNSS per un posizionamento di precisione in tutto il mondo
  - Non è richiesta alcuna stazione base o network RTK
  - Tre diversi livelli di correzione a seconda della precisione richiesta

suite cube

www.stonex.it