

Tratto da Sis-Magazine

<http://www.sis-statistica.it/magazine>

R, un comune {open source} o una nuova frontiera per il software statistico?

- Articoli -

Data di pubblicazione : mercoledì 18 novembre 2009

Sis-Magazine

Il 7 gennaio 2009 l'edizione *online* del New York Times ha pubblicato un [articolo](#) di Stuart Isett in cui il software open source R viene descritto come prodotto che sta diventando lingua franca sia per l'accademia che per le aziende. L'articolo ha suscitato una notevole sorpresa ovunque: il fatto che R stesse diventando uno dei principali software di riferimento per gli statistici era noto da tempo, ma non era certo facile supporre che il fenomeno potesse riguardare anche gli ambienti aziendali.

In realtà nell'articolo si parla di aziende di grandi dimensioni (Google, Pfizer, Merck, Bank of America, InterContinental Hotel Group, Shell, ecc.) ma non conosciamo la reale portata del fenomeno (le uniche informazioni che circolano, riportate da [Vance, 2009](#), parlano di tentativi di stima del numero di utenti di R che variano, però, tra 250.000 e 2 milioni!). Tuttavia i termini con cui l'autore dell'articolo si è espresso sono inequivocabili e il fatto ha sorpreso anche gli ideatori di R, Ross Ihaka e Robert Gentleman, come pure John Chambers, uno degli autori di S (da cui R ha tratto ispirazione e con cui mantiene buona compatibilità).

L'attualità e la rilevanza dell'argomento meritano sicuramente attenzione, pertanto nel seguito si tenterà un'analisi dei fatti e del relativo contesto.

Origini e caratteristiche di R

Le origini di R risalgono al 1993, quando Ross Ihaka e Robert Gentleman iniziano alcuni esperimenti presso la University of Auckland, New Zealand, finalizzati alla realizzazione di un software per l'insegnamento della statistica in laboratorio (Ihaka, 1998). Questi esperimenti si sono protratti negli anni successivi fino a portare, nel giugno 1995, ad un prodotto software di un certo interesse, che gli autori rendono pubblico e disponibile a tutti via FTP, nei termini della *Free Software Foundation's GNU general license*. Le sperimentazioni di Ihaka e Gentleman sono passate per momenti diversi fino a convergere verso un ambiente di programmazione che utilizzava una sintassi ispirata a S, un linguaggio, quest'ultimo, che stava delineandosi come una delle proposte per la statistica più versatili e avanzate dal punto di vista informatico.

S, a sua volta, era nato presso i Bell Laboratories (AT&T) da alcune idee di John Chambers, messe a punto insieme a Rick Becker, Doug Dunn, Paul Tukey, e Graham Wilkinson. Lo sviluppo di S inizia nel 1976 e continua fino al 1998, ma rimane a lungo un prototipo ad uso strettamente interno. Solo nel 1993 viene presa la decisione di consentirne la commercializzazione, con il nome S-plus, affidando l'esclusiva alla società Statistical Science (StatSci), successivamente incorporata da MathSoft e passata quindi ad Insightful Corporation, che nel 2004 acquista anche tutti i diritti.

Il 1995 segna la data di nascita di R come prodotto *open source* perché alcuni utenti cominciano a collaborare con gli autori per segnalare *bug* e suggerire migliorie. Ciò induce Ihaka e Gentleman a creare una *mailing list* in cui poter discutere pubblicamente dei vari problemi e degli ulteriori sviluppi.

Nel 1997, visto il crescente interesse per questo software, si forma un *core group* (Doug Bates, Peter Dalgaard, Robert Gentleman, Kurt Hornik, Ross Ihaka, Friedrich Leisch, Thomas Lumley, Martin Mächler, Paul Murrell, Heiner Schwarte, and Luke Tierney) che, su base volontaria, si incarica di apportare correzioni e miglioramenti al codice sorgente. Il *core group* si arricchisce rapidamente anche di altri collaboratori, tra cui lo stesso John Chambers, ed è a questo [gruppo allargato](#) che si deve la crescita continua di R, nonché la realizzazione di versioni per la maggior parte di piattaforme, sia hardware che software (in origine erano disponibili solo versioni per ambienti Unix).

Attualmente R è un prodotto molto affidabile, stabile e in continua evoluzione: nell'anno 2008 sono state rilasciate 6 diverse versioni (dalla 2.6.2 del mese di febbraio alla 2.8.1 del mese di dicembre) e 3 sono state quelle rilasciate nei primi otto mesi del 2009.

È noto che i prodotti *open source*, a differenza di quelli commerciali, solitamente non dispongono di una documentazione propria accurata ed estesa.

R non si discosta da questa ottica, tuttavia anche per questo aspetto si è assistito ad un fatto abbastanza inusuale, ma molto significativo. A partire dal 2000 sono apparsi nel mercato librario internazionale diversi testi di statistica basati su R, ed il loro numero è cresciuto di anno in anno. È interessante, a questo proposito, visitare la sezione [Books Related to R](#) del sito web ufficiale (AA.VV., 2009d): da essa si desume che dal 2005 in poi i volumi pubblicati sono oltre 65, di cui 20 nel 2008 e 22 nel corso dei primi 8 mesi del 2009.

A questo proposito va anche segnalato che dal 2001 è iniziata la pubblicazione online di una rubrica tecnico-scientifica, [R News](#) (AA.VV., 2008), che poi, dal 2009, è diventata [The R Journal](#) (AA.VV., 2009c), ovvero una rivista scientifica online con regolare referaggio.

Ruolo e diffusione di R

R, come anche S-plus, è da sempre considerato un software per addetti ai lavori che richiede una preparazione statistica specifica e qualche conoscenza dei concetti di base della programmazione ad oggetti. Nei corsi universitari l'introduzione a R non è mai associata a corsi di statistica di base (soprattutto quando sono corsi di servizio inseriti nell'ambito di curricula di altro tipo) e comunque l'insegnamento di questo linguaggio trova spazio solo negli indirizzi di studio più specificamente orientati all'analisi dei dati.

Queste considerazioni sono facilmente riscontrabili nella maggior parte dei paesi e indurrebbero pensare che R abbia una utilizzazione di *élite* e quindi piuttosto limitata. Peraltro, non è un mistero il fatto che spesso le aziende, in sede di valutazione dei candidati, richiedono (o comunque preferiscono) profili che includono conoscenze specifiche sull'uso dei prodotti software commerciali più affermati. Evidentemente tali prodotti appaiono ai loro occhi più affidabili e completi, senza considerare che una migrazione da R verso altri ambienti è sempre facile e rapida, ma non vale certo il viceversa.

In base a queste considerazioni, l'articolo di Isett sembra introdurre novità inattese che non sono in grado di giustificarsi da sole. Visto però che tali novità fanno riferimento a quanto sta accadendo negli USA, è lecito domandarsi che cosa stia cambiando in quel Paese, considerando che certe tendenze prima o poi si diffondono anche altrove.

Uno sguardo alla situazione americana

All'inizio del mese di gennaio 2009 nel sito web [CareerCast: your job search portal](#), un portale dedicato ad analisi e informazioni sul mercato del lavoro in America, appare la classifica dei [10 Best Jobs You Can Get Today](#) con un commento di Tony Lee: i primi tre posti della classifica sono occupati, nell'ordine, dai matematici, dagli attuari e dagli statistici (Lee, 2009).

Nel [numero di marzo](#) di [AMSTAT News](#) (il notiziario dei membri dell' [ASA](#), *American Statistical Association*) compare un breve articolo dal titolo *Statistician: a Sexy Job* (ASA 2009, p. 17) che sostanzialmente riporta il punto di vista di Hal Varian secondo cui la professione dello statistico è *part of the reconfiguring of the business industry's future*, per concludere infine dicendo che *the sexy jobs in the next ten years will be statisticians* (Varian 2009).

All'inizio di agosto, poi, è il *New York Times* a tornare sull'argomento con un articolo di Steve Lohr dal titolo forte e inequivocabile: [For Today's Graduate, Just One Word: Statistics](#) (Lohr, 2009).

Caratteristica rilevante di queste voci è il fatto che esse vengono da figure di notevole rilievo che collaborano con aziende di grande prestigio. Le motivazioni riportate sono diverse, ma riconducibili ad un solo fatto: si sta finalmente prendendo atto dell'importanza delle informazioni che possono essere acquisite mediante l'analisi dei dati. Certo, negli ambienti aziendali si cerca spesso di evitare la parola *statistica*, un termine sempre meno amato, a favore di un vocabolario più accattivante, come *Data Mining* e processi di *Business Intelligence*, tuttavia sembra che la consapevolezza delle potenzialità di certe metodologie ed i relativi vantaggi siano finalmente filtrati in quegli ambienti, almeno a certi livelli, anche se le recenti vicende nell'ambito finanziario hanno fatto nascere diversi interrogativi.

Fermi restando autorevolezza e prestigio dei commentatori, non sappiamo però quanto il fenomeno sia realmente esteso e quali tipologie di aziende siano interessate. Sono comunque importanti le notizie di fonte [ASA](#): le conferenze annuali dell'Associazione recentemente hanno fatto registrare una crescita costante del numero dei partecipanti più giovani, e il fatto diventa molto interessante se si considera che tra i partecipanti si contano anche diversi manager aziendali.

Accanto a queste notizie, non completamente riscontrabili, esistono comunque anche indicatori che sembrano convalidare le aspettative di sviluppo del settore e provengono dal mercato finanziario. Negli ultimissimi anni si è osservato un forte e crescente interesse verso le aziende produttrici di software statistico (in particolare di *Data Mining*) e recentemente, luglio 2009, IBM ha acquistato SPSS. Ma il mercato sembra più che mai in movimento e gli osservatori parlano sempre più spesso di nuovi possibili accordi e/o acquisizioni: aziende come SAS, Oracle, Insightful, KSN sembrano essere i sorvegliati speciali.

Vale la pena notare che R non è completamente estraneo a questi interessi, seppure in forma diversa: nelle valutazioni degli osservatori, infatti, talvolta si parla anche di questo prodotto, vista la rapida crescita e il notevole riscontro che comincia ad avere in vari settori.

Nel prendere atto di questo scenario viene spontaneo porsi domande anche sulla situazione Italiana.

L'articolo di Lohr dello scorso agosto è stato ripreso pochi giorni dopo dal *Sole 24 ore* con un articolo di Eliana Di Caro dal titolo *Tutti pazzi per gli statistici* (Di Caro, 2009), ma la sensazione che si riceve in alcuni punti di quell'articolo (sono forse prevenuto?) è che l'autore stenti un po' a credere alla reale portata dei fatti menzionati e alla loro ripetibilità nel nostro Paese.

In realtà, anche le aziende italiane hanno iniziato a comprendere il valore delle informazioni che possono essere ottenute con l'analisi dei dati, ma il fenomeno appare abbastanza circoscritto, almeno per ora. Inoltre, quando nasce l'esigenza di conoscere meglio la realtà generalmente si preferisce chiedere supporto all'esterno, interpellando aziende specializzate, e non sembra che si senta la necessità di creare strutture interne, e sviluppare così una cultura statistica aziendale. Intanto, però, in Italia sono stati chiusi un po' ovunque corsi di laurea e facoltà di statistica per carenza di studenti.

È lecito allora domandarsi *quando* la moda dell'analisi dei dati sbarcherà completamente sulle nostre coste e *come*

si potranno mettere in campo le forze necessarie al momento giusto. Fortunatamente c'è sempre un *gap* temporale tra quanto accade oltre oceano e la successiva ricaduta sul nostro Paese. C'è solo da sperare che ci sia anche il tempo necessario per attrezzarsi adeguatamente.

Considerazioni conclusive

Gli elementi che emergono dallo scenario americano aiutano sicuramente a spiegare il fenomeno della inaspettata diffusione di R e consentono di avanzare qualche ipotesi. Si è visto che il crescente interesse delle aziende per i processi di *Business Intelligence* ha determinato un'alta richiesta di statistici, e questi generalmente sono anche in possesso di PhD. Ovviamente, queste figure hanno portato con sé la capacità di utilizzare un'ampia gamma di supporti software avanzati; è ragionevole quindi supporre che R, trattandosi di software gratuito, sia entrato senza difficoltà nei rispettivi posti di lavoro, facilitando tra l'altro il processo di integrazione e interazione con altri ambienti software preesistenti. È ragionevole anche supporre che questo rappresenta solo l'inizio di un processo di diffusione ancora più esteso e destinato a propagarsi anche al di fuori degli Stati Uniti.

A conclusione di queste riflessioni e ad ulteriore conferma del ruolo assunto da R, va segnalato che recentemente questo prodotto ha riscosso attenzioni anche al di fuori del suo terreno accademico tradizionale. Alla fine del 2007, infatti, è nata [REvolution Computing](#), una società che ha come obiettivo lo sviluppo di estensioni e ottimizzazioni di R; in particolare, obiettivo dell'azienda è la commercializzazione di REvolution R Enterprise, un prodotto che si basa su moduli aggiuntivi (*ParallelR*) che implementano algoritmi per il calcolo parallelo su computer dotati di più processori e/o di più nuclei (*multicore*).

Per saperne di più

AA.VV. (2008), [R News](#)

AA.VV. (2009a), [The R Project for Statistical Computing](#)

AA.VV. (2009b), [CRAN](#)

AA.VV. (2009c), [The R Journal](#)

AA.VV. (2009d), [Books Related to R](#)

ASA (2009), Statistician: A Sexy Job . *AMSTAT NEWS*, Issue # 381, p. 17

Chambers J. (2000), [Stages in the Evolution of S](#)

Chambers M.J. (2008), *Software for Data Analysis: Programming with R*. Springer

Di Caro E. (2009), Tutti pazzi per gli statistici . *Il Sole 24 ore*, 7 agosto 2009 pag. 8

Ihaka R. (1998), [R : Past and Future History](#)

Iacus S. (2006), [R: un ambiente statistico in continua evoluzione](#) . *Statistica & Società*, anno IV, n. 3, pp. 19-23

Isett S. (2009), [Data Analysts Captivated by R's Power](#) . *The New York Times*,

Lee T. (2009), [The 10 Best Jobs You Can Get Today](#) . Career Cast

Lohr S. (2009), [For Today's Graduate, Just One Word: Statistics](#) . *The New York Times*

Vance A. (2009), [R You Ready for R?](#) . *The New York Times*

Varian H. (2009), [How the Web Challenges Managers](#) . *The McKinsey Quarterly*, January

Wilson D. (2008), "[The rise of R](#)". *VHAYU blog*,