La relazione sulla gestione delle società italiane quotate sul mercato regolamentato[1]

Maria Spano, Nicole Triunfo

1Università FEDERICO II di Napoli – maria.spano@unina.it; nicole.triunfo@unina.it

Abstract

This paper has been developed in the frame of the European project BLUE-ETS (Enterprise and Trade Statistics), in the work-package devoted to propose new tools for collecting and analyzing data. Aim of this paper is to show the possibility to extract and analyze data in order to produce official statistics by mining into the management commentaries attached to the annual report. Each year listed companies on Italian market have to produce these informative documents on their operations and situation.

Furthermore we propose to prove the interdependence between performance indexes, calculated by financial statement data, and the clearness of the language used for writing of the «letter to the shareholders». To do this we introduce, in the quantitative variables set, the readability index, usually used in the financial content analysis. The tool, that we have chosen, is Canonical Correspondence Analysis, proposed by ter Braak to identify relationships between species and the environment from the data on the composition of communities and associated habitat measurements.

We suggest, therefore, an innovative use of this technique when analyzing two sets of data, a text, and a set of quantitative variables.

Riassunto

Questo lavoro nasce nell'ambito del progetto Europeo BLUE-ETS, nel work-package dedicato alla ricerca di nuovi modi per raccogliere ed analizzare dati. Il nostro obiettivo è mostrare, utilizzando come fonte amministrativa una base dati documentale (la relazione sulla gestione allegata al bilancio d'esercizio delle società italiane quotate sul mercato regolamentato), la possibilità di estrarre ed analizzare dati utili alla produzione di statistiche ufficiali in ambito economico. In particolare, ci proponiamo di dimostrare l'interdipendenza esistente tra gli indici di performance calcolati con dati provenienti dal bilancio e la chiarezza del linguaggio utilizzato per la stesura della «lettera agli azionisti». Per questo abbiamo introdotto tra le variabili quantitative un indice di leggibilità, seguendo proposte della letteratura della financial analysis. Lo strumento metodologico che abbiamo prescelto è una tecnica sviluppata in ambito ecologico, l' Analisi delle Corrispondenze Canoniche, proposta da ter Braak per individuare le relazioni tra le specie e l'ambiente a partire dai dati sulla composizione della comunità e dalle misurazioni associate all'habitat. Proponiamo, quindi, un innovativo utilizzo di questa tecnica nell'ambito dell'analisi di due insiemi di variabili, uno testuale, l'altro quantitativo.

Keywords: Canonical correspondence analysis, Information extraction, Administrative data.

1 Questo lavaro è cofinanziato dal progetto Europeo BLUE-ETS. La redazione materiale dei paragrafi 1, 3, 6 è a cura di Maria Spano, i paragrafi 2, 4, 5 sono a cura di Nicole Triunfo. Gli autori condividono la responsablilità per il contenuto dell'intera pubblicazione.

1. Introduzione

Nello scenario economico mondiale, e più da vicino in quello Europeo, la statistica ufficiale riveste un ruolo di fondamentale importanza; questo principio compare nello stesso trattato istitutivo della Comunità Europea all'articolo 211.

La produzione di statistiche ufficiali di elevata qualità e attendibilità rende necessario controbilanciare le necessità di dati, da un lato, e l'onere gravante sui rispondenti, dall'altro. Il principio 9 del codice delle statistiche europee sancisce che le autorità statistiche devono stabilire un programma per la riduzione nel tempo dell'onere per i rispondenti.

Lungo questa direttrice l'Istituto Nazionale di Statistica ha avviato già da tempo un programma di adeguamento a questo principio. Ne sono un esempio l'archivio delle imprese attive (ASIA), che viene utilizzato dal 2004 nella fase di identificazione della popolazione di riferimento delle indagini; e le metodologie di rilevazione miste, attraverso le quali l'Istat, ove possibile, integra i dati provenienti da rilevazioni dirette (questionario) con dati provenienti da fonti informative disponibili (es. fonti amministrative).

In questo scenario, il progetto BLUE-ETS (Enterprise and Trade Statistics) cofinanziato dalla Commissione Europea si pone gli obiettivi di:

- far progredire le conoscenze statistiche e metodologiche;
- ridurre l'onere statistico dei rispondenti (imprese);
- migliorare la qualità delle statistiche in ambito economico.

Il nostro contributo al progetto ha come obiettivo «New ways of collecting and analysing data». Questo lavoro nasce, con lo scopo di raccogliere dati utili per la produzione di statistiche ufficiali in ambito economico, nonché per tutti i soggetti di interesse, utilizzando come fonte amministrativa una base dati documentale. Inoltre ci proponiamo di dimostrare l'interdipendenza esistente tra gli indici di performance delle società quotate sul mercato italiano e la chiarezza del linguaggio utilizzato per la stesura di documenti amministrativi e contabili.

La base documentale utilizzata è «la lettera agli azionisti» contenuta nella relazione sulla gestione; ne saranno definiti contenuti e peculiarità nel prossimo paragrafo.

Successivamente presenteremo dettagliatamente la tecnica di analisi utilizzata, l'Analisi delle Corrispondenze Canoniche. Si tratta di un metodo di analisi multidimensionale dei dati, che ha riscosso un grande successo nella comunità ecologica, finalizzato all'individuazione delle relazioni che intercorrono tra le specie e le caratteristiche dell'ambiente in cui vivono.

Nel quarto paragrafo verranno illustrate le opportune assunzioni e le dovute considerazioni che ci hanno permesso di utilizzare questa tecnica per raggiungere i nostri obiettivi di analisi. Il quinto è interamente dedicato alla presentazione dei risultati. L'ultimo paragrafo riguarda gli sviluppi metodologici futuri e le possibilità che l'analisi delle corrispondenze canoniche offre per applicazioni riguardanti i dati testuali.

2. La relazione sulla gestione

L ‘articolo 2428 del codice civile dispone che: *«il bilancio d’esercizio deve essere corredato da una relazione degli amministratori contenente un’analisi fedele, equilibrata ed esauriente della situazione della società»*.

La relazione sulla gestione, quale documento informativo sull’andamento della gestione e sulla situazione della società, offre lo spunto per un’analisi finalizzata all’estrazione di informazioni utili ad arricchire la produzione di statistiche ufficiali in ambito economico, nonché ad ampliare l’informativa d’interesse degli stakeholders. Del resto, la tempistica annuale della sua stesura e la normativa che la disciplina inducono a pensare che nel documento stesso siano trattate in maniera esaustiva tutte le tematiche che hanno condizionato l’andamento e le strategie di gestione avvenute nell’esercizio a cui essa si riferisce. La relazione sulla gestione rappresenta una fonte «amministrativa», disponibile on-line (nel sito di Borsa Italiana), potenzialmente utile alla produzione di statistiche ufficiali.

Questo documento, decisamente ricco dal punto di vista dei contenuti, presenta problemi di «forma». La normativa sopra citata detta i principi che gli amministratori devono rispettare nella stesura, ma lascia ampia discrezionalità agli stessi per la struttura che deve assumere. Per questo motivo, come prima analisi, abbiamo deciso di analizzare un’unica sezione del documento, la «lettera agli azionisti», comune a tutte le società campionate: Effegi, Azimut, Ceramiche Ricchetti, Autostrade Torino, Mediaset, Mediacontech, Esprinet, Isagro, Centrale del latte di Torino, Kme, Elica, Acque Potabili, Recordati, Gas Plus, Juventus, Mondadori, Montefibre, Meridiana Fly, Prima Industrie, Enel, Carraro, Kinexia, S.E.I., Rcf.

3. L’analisi delle corrispondenze canoniche

In ecologia gli studi sulla composizione delle comunità ecologiche, sono sviluppate attraverso l’utilizzo di una tecnica di analisi, nota come: Analisi delle corrispondenze canoniche. I motivi che ci hanno indotto, per questo studio, all’utilizzo di tale tecnica, sono da ricercarsi nelle similitudini che abbiamo riscontrato tra le matrici lessicali e le matrici riguardanti la composizione delle specie nei siti. Infatti, sia i dati ecologici che i dati testuali hanno la peculiarità di generare matrici tipicamente sparse. Inoltre, così come le comunità ecologiche sono osservate attraverso lo studio delle variabili che le caratterizzano e delle specie che le abitano, la chiarezza dei documenti contabili è valutata attraverso lo studio di indicatori di performance e il linguaggio utilizzato dalle aziende per la loro stesura.

Lo studio delle comunità ecologiche è basato generalmente sull’analisi di due matrici di dati, una che contiene informazioni sulle composizioni delle specie nei siti (ad esempio, l’abbondanza, la copertura delle specie) e un’altra contenente le principali caratteristiche dell’habitat in tali siti (un “sito” è un’unità campionaria base, separata nello spazio o nel tempo da altri siti, ad esempio: un mucchio di legna, un campione di plankton, una trappola), informazioni queste che influenzano la distribuzione delle specie. Questo tipo di analisi è caratterizzato da un duplice obiettivo, da un lato l’individuazione di modelli di distribuzione delle specie e l’ordinamento dei siti compatibili con un dato gradiente, e dall’altro lo studio del rapporto tra questi risultati e le variabili ambientali misurate. Queste tabelle contengono una grande quantità di informazioni, parte delle quale risulta ridondante. Per sintetizzare l’informazione ed individuare la struttura

latente di tali tabelle, è possibile utilizzare tecniche multivariate, che permettono l'organizzazione dei siti lungo gli assi, in base ai dati riguardanti la composizione delle specie. L'analisi delle corrispondenze (Benzécri 1973) è un esempio di queste tecniche di analisi multidimensionale dei dati e può essere considerata come un metodo di ordinamento non vincolato (ACC ter Braak 1986). Gli assi fattoriali individuati con un'analisi delle corrispondenze classica sono solitamente interpretati con l'aiuto di conoscenze esterne o effettuando un'analisi di regressione multipla, o attraverso il calcolo dei coefficienti di correlazione tra gli stessi assi e le variabili ambientali, che non concorrono di fatto alla loro determinazione. Questo approccio organizzato in due fasi (individuazione degli assi e interpretazione degli stessi), in cui vengono dedotti gradienti ambientali dai dati ecologici, è conosciuto come analisi indiretta del gradiente (Whittaker 1967).

Imponendo la limitazione che gli assi fattoriali siano una combinazione lineare delle variabili ambientali, è possibile far sì che le caratteristiche dell'habitat, osservate in ciascun sito, svolgano un ruolo attivo nell'analisi. Il metodo che ha riscosso maggior successo nell'ambito degli studi ecologici, collocandosi nel contesto dell'analisi diretta del gradiente, è l'Analisi delle Corrispondenze Canoniche (ACC ter Braak 1986). Questa tecnica di ordinamento vincolato permette di legare direttamente le variazioni nella comunità alle variazioni ambientali. Le applicazioni dimostrano infatti che l'ACC può essere utilizzata sia per indagare circa le relazioni tra specie e ambiente, sia per rispondere a domande più specifiche circa la risposta delle specie alle variabili ambientali. Come suggerisce il nome, l'ACC è un'estensione dell'Analisi delle Corrispondenze, che consente di visualizzare non solo la distribuzione delle specie nei siti esaminati, ma anche le principali caratteristiche delle specie lungo le variabili ambientali.

Pertanto, trattandosi di un'analisi delle corrispondenze vincolata al sottospazio generato dalle variabili ambientali in cui i siti e le specie sono proiettati, il numero massimo di dimensioni che possono essere rappresentate è pari al massimo al numero di variabili ambientali che intervengono nell'analisi, siano esse quantitative e/o nominali. Considerati ad esempio n siti, se il numero di variabili aumenta l'analisi delle corrispondenze risulta sempre meno vincolata, fino al caso limite in cui il numero di variabili $p \geq n-1$ e l'ACC non è altro che una AC.

3.1. Metodologia e Algoritmo

Consideriamo uno studio condotto su n siti in cui si quantifica la frequenza o la presenza/assenza (presenza=1, assenza=0) di m specie e i valori di q variabili ambientali ($q<n$). Sia y_{ik} la frequenza o la presenza-assenza della specie k nel sito i e sia z_{ij} il valore della variabile ambientale j misurata nell'i-esimo sito.

Il primo passo di un'analisi indiretta del gradiente è quello di riassumere la maggior parte della variabilità delle specie tramite l'ordinamento. Partendo dal presupposto che il rapporto tra i dati delle specie e le variabili ambientali segua una curva gaussiana di risposta, Gauch *et al.* (1974) hanno proposto una tecnica chiamata Ordinamento Gaussiano. Pertanto il modello di risposta per le specie è rappresentato dalla funzione campanulare:

$$E(y_{ik}) = c_k \exp[1/2(x_i - u_k)^2 / t_k^2] \quad (1)$$

Dove $E(y_{ik})$ rappresenta il valore atteso di y_{ik} al sito i, la cui coordinata (score) sull'asse di ordinamento è x_i. I parametri per le k specie sono: c_k, massimo della curva di risposta delle

specie; u_k, la moda, ossia il valore di x per cui si ottiene il massimo e t_k, la tolleranza, una misura di ricchezza ambientale.

Il passo successivo è quello di effettuare un'analisi di regressione multipla, che metta in relazione gli stessi assi con le variabili ambientali:

$$x_i = b_0 + \sum_{j=1}^{q} b_j z_{ij} \quad (2)$$

Dove b_0 è l'intercetta, b_j è il coefficiente di regressione della j-esima variabile ambientale e x_i è la coordinata (score) sull'asse di ordinamento di y_{ik} al sito i. Si noti che le coordinate sull'asse di ordinamento sono ottenute nella prima fase a partire dalla matrice contenenti i dati circa la composizione delle specie nei siti; i coefficienti di regressione b_j sono stimati successivamente, mantenendo fissati i valori x_i.

Pertanto, le specie sono indirettamente legate alle variabili ambientali, tramite gli assi di ordinamento. Sebbene questa tecnica combinata in due passi, denominata da ter Braak (1985) Ordinamento Canonico Gaussiano, sia statisticamente rigorosa risulta dal punto di vista computazionale molto onerosa. È per questa ragione che ter Braak, avendo dimostrato che l'analisi delle corrispondenze approssima la soluzione di massima verosimiglianza dell'Ordinamento Gaussiano, introduce l'analisi delle corrispondenze canoniche, come approssimazione euristica dell'Ordinamento Canonico Gaussiano.

Le considerazioni che portano a tale approssimazione si concretizzano nelle formule di transizione dell'ACC (ter Braak, 1986):

$$\lambda u_k = \sum_{i=1}^{n} y_k x_i / y_{.k} \quad (3)$$

$$\lambda u_k = \sum_{i=1}^{n} y_k x_i / y_{.k} \quad (4)$$

$$b = (Z'RZ)^{-1} Z'Rx^* \quad (5)$$

$$x = Zb \quad (6)$$

Dove $y_{.k}$ e $y_{i.}$ sono rispettivamente i marginali di colonna e di riga della matrice riguardante la composizione delle specie nei siti, R è una matrice diagonale con elemento generico $y_{i.}$ di dimensioni $n \times n$; $Z=\{z_{ij}\}$ è una matrice di dimensioni $n \times (q+1)$ contenente i valori delle variabili ambientali e una colonna di 1; b, x , x^* sono tre vettori colonna: $b=(b0,b1,.....,bq)'$, $x=(x1,.....xn)'$ e $x^*=(x1^*,.....xn^*)'$. Le formule di transizione definiscono un problema vettoriale analogo a quello nell'analisi delle corrispondenze in cui λ rappresenta l'autovalore.

Il problema illustrato può essere risolto utilizzando l'algoritmo iterativo seguente:

- Step1: Attribuire arbitrariamente degli scores iniziali ai siti;
- Step2: Calcolare gli scores delle specie come medie pesate degli scores dei siti (Eq.3 conλ=1);

- Step3: Calcolare i nuovi scores dei siti x_i^* come medie pesate degli scores delle specie (Eq.4);
- Step4: Stimare i coefficienti di una regressione multipla pesata degli scores dei siti sulle variabili ambientali (Eq.5), in cui i pesi sono i totali marginali dei siti $y_{i.}$;
- Step5: Calcolare i nuovi scores dei siti tramite l'Eq.6. I nuovi scores sono infatti i valori predetti della regressione dello step precedente;
- Step6: Standardizzare i nuovi scores: $\sum_i y_{i.} x_i = 0$ e $\sum_i y_{i.} x_i^2 = 1$
- Step7: Arrestare l'algoritmo ottenuta la convergenza, ad esempio, quando i nuovi scores dei siti sono sufficientemente vicini a quelli della precedente iterazione; altrimenti procedere con lo Step2.

L'algoritmo è sostanzialmente analogo a quello di un'analisi delle corrispondenze, con l'aggiunta dei passi 4 e 5, che di fatto vincolano gli score dei siti. Il secondo e i successivi assi della CCA sono anch'essi combinazione lineare delle variabili ambientali che massimizzano la dispersione delle specie, ma sono soggetti al vincolo di essere non correlati (ortogonali) con i precedenti assi.

I coefficienti di regressione finali sono chiamati coefficienti canonici e il coefficiente di correlazione multipla dell'ultima regressione è definito come correlazione specie-ambiente e misura quanta parte della variabilità nella composizione della comunità può essere spiegata dalle variabili ambientali. Guardando ai segni e ai valori dei coefficienti canonici possiamo stabilire l'importanza di ogni variabile ambientale nel predire la composizione della comunità. Se le variabili considerate sono fortemente correlate tra loro (multicollinearità), ad esempio perché il numero di variabili si avvicina al numero di siti presi in esame è difficile separare gli effetti di differenti variabili ambientali sulla composizione della comunità, di conseguenza i coefficienti canonici sono instabili. È importante notare che, a differenza di quanto accade per le variabili ambientali, il numero di specie può essere superiore al numero di siti.

3.2. La rappresentazione grafica

La rappresentazione grafica di un'Analisi delle Corrispondenze Canonica è un tri-plot e consente di visualizzare congiuntamente i siti, le specie e le variabili ambientali. I siti e i punti specie sul grafico possono essere interpretati come in un'analisi delle corrispondenze classica. Le variabili ambientali sono rappresentate attraverso delle frecce. In senso lato, la freccia per una variabile ambientale punta nella direzione di massima variabilità per quella stessa variabile e la sua lunghezza è direttamente proporzionale alla percentuale di variazione in quella direzione. Variabili con frecce più lunghe sono maggiormente correlate con gli assi, rispetto a quelle con frecce più corte, quindi più strettamente legate al modello di variazione della comunità. Una regola per l'interpretazione di questo grafico è quindi la seguente: ogni freccia, che rappresenta una variabile ambientale, determina una direzione o un "asse" nel diagramma, su cui i punti specie possono essere proiettati. L'ordine dei punti proiettati corrisponde approssimativamente alla posizione delle medie pesate delle specie rispetto a quella variabile ambientale.

4. Dati

Per il raggiungimento degli obiettivi precedentemente definiti, proponiamo un innovativo utilizzo di questa tecnica, che ci consente di analizzare congiuntamente variabili quantitative da un lato, e testuali dall'altro. Nei paragrafi successivi sono presentati i due insiemi di variabili considerate.

4.1. Matrice lessicale

La base documentale è stata costituita a fronte di un campionamento casuale effettuato sulla totalità delle società quotate sul mercato Italiano (406), dal quale sono state estratte 24 società italiane quotate sul mercato regolamentato. Come precedentemente specificato, hanno concorso alla costruzione della base oggetto di studio solo le «lettere agli azionisti» contenute nella relazione sulla gestione allegata al bilancio d'esercizio chiusosi il 31 dicembre 2009.

Il corpus analizzato, è costituito, a seguito della normalizzazione effettuata (con il software TalTac 2.0), da 40.347 occorrenze, espresso in 5.028 forme grafiche diverse.

La lettera agli azionisti si presenta come un documento informale, attraverso il quale, il presidente di una società quotata sintetizza, agli azionisti di minoranza (investitori), l'andamento della gestione e i risultati economico-finanziari conseguiti nel corso dell'anno. A seguito di una prima analisi lessicale, si evince con chiarezza che il linguaggio utilizzato per la sua stesura si caratterizza per la presenza di un elevato numero di forme grafiche tipiche di un lessico economico finanziario, difficilmente identificabili con le funzioni di base del software utilizzato. Per questa ragione, abbiamo usufruito di una risorsa esogena, nello specifico il glossario Italiano-Inglese costruito dall'esperienza di Pricewaterhouse Coopers SPA, una delle «BIG FOUR» nell'ambito della revisione contabile, al fine di lessicalizzare le forme grafiche e le polirematiche proprie del linguaggio contabile. Dal confronto del corpus con il lessico di riferimento, siamo riusciti ad identificare 355 forme testuali.

Al fine di approfondire lo studio delle forme grafiche rilevanti, si è scelto di procedere con l'analisi testuale calcolando l'indice TF-IDF (Salton,1989), il quale descrive il peso attribuito ad ogni forma grafica in base alla sua frequenza e alla sua distribuzione all'interno del corpus. Tale indice risulta quindi di supporto alla scelta delle forme grafiche da "studiare" poiché permette di individuare le forme che più di altre presenti nel corpus sono in grado di discriminare le «lettere agli azionisti» delle società campionate.

Dalla procedura svolta, siamo pervenuti alla matrice lessicale di dimensioni 24x79.

4.2. Indicatori di "performance"

Considerando che ci troviamo in una fase iniziale dello studio, abbiamo utilizzato indicatori generici dei differenti aspetti societari.

Le prime due variabili considerate sono due indici di natura contabile, il *ROI* e l'*EBITDA*. Il primo, Return On Investments, indica la redditività e l'efficienza economica della gestione caratteristica a prescindere dalle fonti utilizzate; questo indice esprime quanto rende il capitale investito in quella società. Il *ROI* si calcola dal rapporto tra il risultato operativo e il capitale investito netto operativo. L'*EBITDA*, anche chiamato margine operativo lordo, è un indicatore di redditività che evidenzia il reddito di un'azienda basato solo sulla sua gestione caratteristica,

al lordo, quindi, di interessi, tasse, deprezzamento di beni e ammortamenti. Questo indicatore risulta utile al fine di comparare i risultati di diverse aziende che operano in uno stesso settore, inoltre essendo il suo valore molto simile ai flussi di cassa prodotti da un'azienda, fornisce l'indicazione più significativa al fine di valutarne il valore.

All'opposto, abbiamo considerato 2 variabili che rappresentano l'aspetto «qualitativo» della società. Il primo indicatore «*% di amministratori indipendenti*» è stato costruito rapportando il numero di amministratori indipendenti al totale degli amministratori che costituiscono il consiglio di amministrazione. Questo indicatore di governance, garantisce la trasparenza gestionale e la tutela per i diritti degli azionisti di minoranza e degli altri portatori d'interessi. La variabile categorica «*società di revisione*» garantisce l'attendibilità delle poste di bilancio. Abbiamo assegnato valore 1 alle società che si avvalgono, come società di revisione, di una delle «BIG FOUR» (in ambito di revisione la Pricewaterhouse, KPMG, Ernest&Young, Deloitte&Touche sono considerate le migliori società di revisione), e zero per tutte le altre.

L'ultima variabile considerata è un indice di leggibilità, che esprime la chiarezza e la facile comprensione di un testo. In questo lavoro abbiamo utilizzato l'indice GULPEASE, proposto per la lingua italiana. Esso è stato definito nel 1988 presso l'Istituto di Filosofia dell'Università degli Studi di Roma «La Sapienza», dal gruppo universitario linguistico pedagogico (GULP de Mauro). Si tratta di un'attenta e ponderata revisione di indici precedenti, quale quello elaborato negli anni quaranta del secolo scorso da Rudolf Flesch per l'*American English*, e quello di *Gunning* conosciuto come *Gunning Fog Index* (Gunning R., 1952),proposti nell'analisi del contenuto in ambito di *financial analysts*(Lehavy R. *et al.* 2011).

L'indice Gulpease si calcola applicando la seguente formula:

$$89 - \left(\frac{Lp}{10}\right) + \left(3 \times Fr\right)$$

Dove:

Lp= (numero delle lettere del testo 100)/numero delle parole del testo

Fr = (numero delle frasi del testo 100)/numero delle parole del testo.

I valori che si ottengono sono compresi tra 0 e 100. I lettori che hanno un'istruzione elementare leggono facilmente i testi che presentano un indice superiore a 80; i lettori che hanno un' istruzione media leggono facilmente i testi che presentano un indice superiore a 60; in ultimo, i lettori che hanno un' istruzione superiore leggono facilmente i testi che presentano un indice superiore a 40.

5. Presentazione dei risultati

Per le similitudini già esplicitate nel paragrafo 3, i siti sono assimilabili alle 24 società campionate, le specie sono le 79 parole selezionate in seguito alle operazioni di pretrattamento e le variabili ambientali saranno i 5 indicatori di "*performance*", illustrati nel paragrafo precedente.

L'analisi delle corrispondenze canoniche, condotta con il software XLSTAT 2011, ha portato all'identificazione di due dimensioni rappresentate rispettivamente sul primo e sul secondo asse del grafico in Fig.1 e che complessivamente spiegano il 58,63% della variabilità totale.

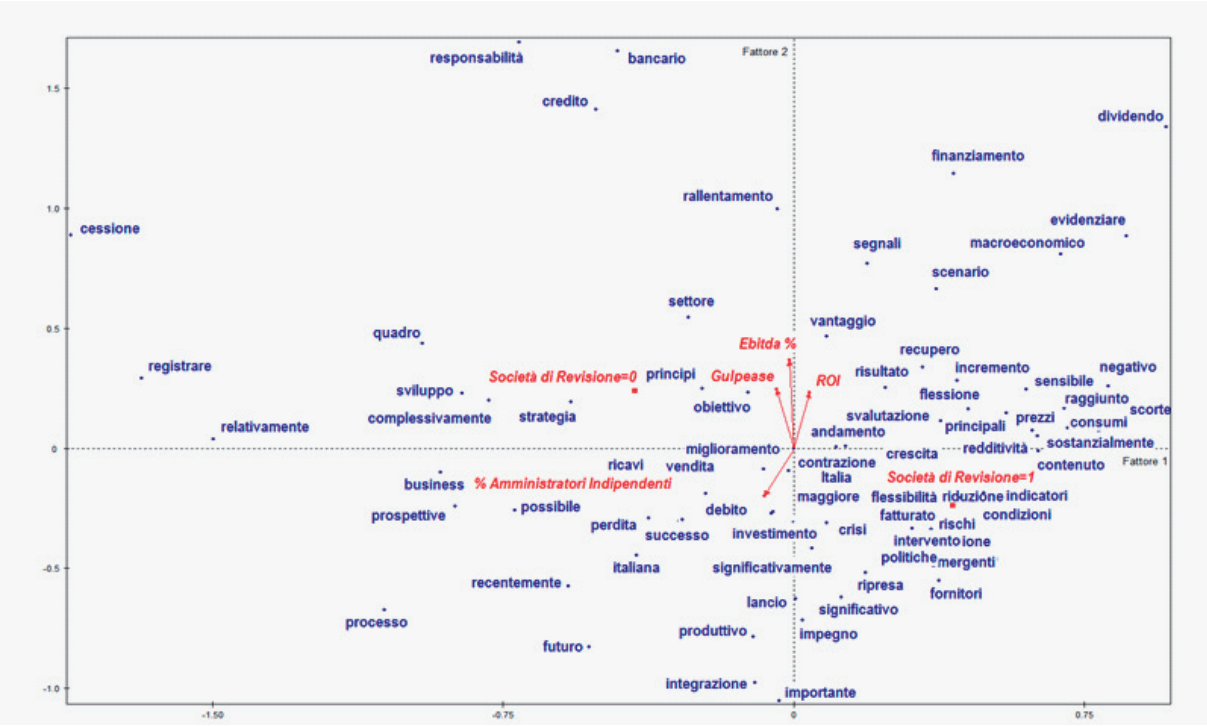

Figura 1: Rappresentazione congiunta delle parole e delle variabili di «performance»

Sebbene, come in precedenza sottolineato, la peculiarità della tecnica è la possibilità di rappresentare simultaneamente in un tri-plot le parole, i documenti e le variabili di "performance", abbiamo preferito presentare 2 diagrammi separati per una questione di chiarezza e per permettere di comprendere il grafico più agevolmente.

La dimensione rappresentata sul primo asse nello spazio generato dalle variabili di "*performance*", identifica le differenti tematiche trattate e i differenti termini utilizzati per comunicare agli stakeholders l'andamento globale della gestione, risulta evidentemente condizionata dalla società di revisione di cui le differenti aziende si avvalgono. I termini che influiscono nell'attribuzione di senso al primo semiasse negativo e che quindi si trovano localizzati nella parte sinistra del grafico sono: <relativamente>, <complessivamente>, <quadro>, <settore>, <sviluppo>, <registrare>. I termini invece che si trovano localizzati nella parte destra del grafico e che contribuiscono ad etichettare il primo semiasse positivo sono: <rischi>, <flessibilità>, <scorte>, <fornitori>, <indicatori>. Dal primo asse possiamo quindi notare che le aziende che si avvalgono, come società di revisione, di una delle «BIG FOUR» argomentano l'andamento della gestione toccando tematiche difficili. Ne sono un esempio i rischi a cui l'azienda è esposta, gli indicatori di performance raggiunti, etc.

In questo passaggio possiamo affermare che l'informazione a più elevata trasparenza e chiarezza è assimilabile alla revisione contabile effettuata dalle società definite «BIG FOUR».

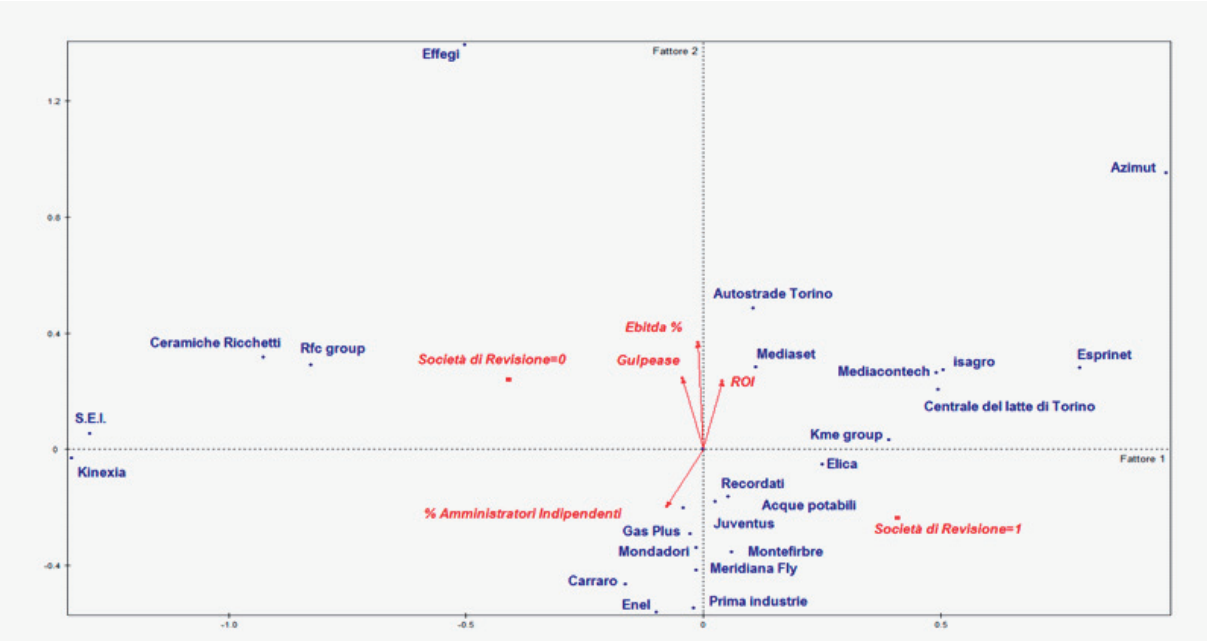

Figura 2: Rappresentazionecongiunta delle società e delle variabili di «performance»

Il secondo asse descrive, lungo un continuum, le performance aziendali. Sul primo semi asse positivo sono rappresentate le aziende che nell'esercizio considerato hanno raggiunto performance positive, e tutto ciò si riflette ancora una volta sulla chiarezza e la trasparenza del linguaggio utilizzato; infatti ad alti valori di performance, corrisponde un valore dell'indice di leggibilità prossimo a 40. La parte bassa del grafico è caratterizzata da una forte correlazione con la percentuale di amministratori indipendenti. Un valore alto di questo indicatore garantisce la tutela degli azionisti di minoranza e dei soggetti interessati. Se però andiamo a guardare i termini localizzati in questa parte del grafico, tra i quali troviamo: <diminuzione>, <indebitamento>, <intervento>, <perdita> notiamo che ad un buon indicatore di governance non sempre corrispondono buone performance. La spiegazione di ciò, costituisce oggetto di un aperto dibattito in ambito di corporate governance, nel quale si discute che l'indipendenza degli amministratori a volte si tramuta in eterogeneità dell' organo di amministrazione (poca comunicazione, interazione), il tutto si può concretizzare in basse performance.

6. Conclusioni e sviluppi futuri

L'approccio metodologico adottato in questo lavoro, è caratterizzato da grandi potenzialità per ciò che concerne l'applicazione ai dati testuali. Tuttavia, la considerazione più naturale è che effettuare un'analisi delle corrispondenze su dati testuali comporta dei problemi derivanti dalla metrica del chi-quadrato.. La metrica del chi-quadrato tende ad enfatizzare l'importanza di termini rari, poiché l'inverso del marginale di un termine con bassa frequenza tende ad esplodere.

Per risolvere questo inconveniente, il metodo da utilizzare è l'analisi non simmetrica delle corrispondenze, in cui alla metrica del chi-quadrato si sostituisce la metrica euclidea. I migliori risultati ottenuti applicando un'analisi delle corrispondenze non simmetrica ai dati testuali

(Balbi 1995), ci conducono all'idea di sviluppare un metodo che metta in relazione l'ACC e l'ANSC, sulla scorta del contributo di Willems e Galindo Villardòn (2008), che propongono una tecnica denominata analisi delle corrispondenze canoniche non simmetriche (ACNC) applicata a dati ecologici.

Bibliografia

Airoldi G., Forestieri G.(1998). Corporate Governance - Analisi e prospettive del caso italiano. Milano. ETAS Libri.

Balbi, S. (1995). Non symmetrical correspondence analysis of textual data and confidence regions for graphical forms.In Bolasco S. *et al.* (eds.).Actesdes 3es Journéesinternationalesd'Analysestatistique des DonnéesTextuelles. CISU, Roma. vol.(II): 5-12.

Benzécri, J.P. (1973). L'AnalysedesDonnées: Tome I: La Taxonomie. Tome 2:L'analyse desCorrespondance. Paris. Ed.Dunod.

Bolasco, S.(1999). Analisi multidimensionale dei dati. Metodi, strategie e criteri d'interpretazione. Roma. Carocci.

Bolasco, S., Pavone, P. (2010). Automatic dictionary and rule-based systems for extracting intext. In F. Palumbo, C.N. Lauro, and M.J. Greenacre, . Editors. Data analysis and classification. London. Springer. 189-198.

Hugh G. Gauch, Jr., Gene B. Chase and Robert H. Whittaker(1974). Ordination of vegetation samples by Gaussian species distributions. Ecology, vol.(55):1382-1390.

Jensen, M.C. (2000) A theory of the firm. Cambridge. Harvard University Press.

Lucisano, P., Piemontese, M.E. (1988). Gulpease: una formula per la predizione della difficoltà dei testi in lingua italiana. Scuola e città, vol.(XXXIX): 110-124.

Salton, G. (1989). Automatic text processing: The transformation, analysis and retrieval of information by computer. Boston (MA): Addison-Wesley.

TerBraak, C.J.F. (1985). Correspondence analysis of incidence and abundance data: properties in terms of a unimodal response model. Biometrics, vol.(41):859-873.

TerBraak, C.J.F. (1986). Canonical correspondence analysis: a new eigenvector technique for a multivariate direct gradient analysis. Ecology, vol.(67):1167-1179.

Whittaker, R.H. (1967). Gradient analysis of vegetation. Biological Reviews, vol.(49): 207-264.

Willems, P. M., Galindo Villardón, M.P. (2008). Canonical non-symmetrical correspondence analysis: an alternative in constrained ordination.Statistics and Operations Research Transactions. vol.(32): 93-111.

ReuvenLehavy, Feng Li, Kenneth Merkley (2011). The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts. American Accounting Association, vol (86): 1087-1115.