Università degli Studi di Napoli
Federico II

Produzione di informazione statistica ufficiale: il ruolo dei dati testuali

Nicole Triunfo

Tesi di Dottorato
in Statistica

XXV Ciclo

Dipartimento
di Matematica e Statistica
Università degli Studi di Napoli "Federico II"
via Cintia, Monte Sant'Angelo – 80126 Napoli

Produzione di informazione statistica ufficiale: il ruolo dei dati testuali

Napoli, 1 Aprile 2013

III

Ringraziamenti

Grazie di cuore a tutte le persone che hanno condiviso con me questo percorso professionale e di crescita umana.

Nicole

Introduzione

La crisi economica degli ultimi anni ha generato anche nell'ambito della statistica ufficiale un nuovo modo di pensare *'up-to-date'*, cioè ha evidenziato la necessità di essere in grado di reagire tempestivamente ed in modo adeguato ai cambiamenti e agli eventi inattesi. Il contesto descritto ha sottolineato le carenze del sistema statistico esistente, ponendo quest'ultimo di fronte a nuove scelte, per le quali non è facile trovare una soluzione univoca, che riesca a soddisfare contemporaneamente tutti i soggetti interessati. Gli utenti: le imprese, le persone fisiche e soprattutto le istituzioni governative richiedono dati sempre più dettagliati, tempestivi e di buona qualità. Contestualmente i fornitori di dati, in particolare le imprese, richiedono di diminuire il carico amministrativo e statistico gravante su di loro.
L'incessante sviluppo tecnologico e la crescente diffusione di dispositivi collegabili alla rete Internet sta creando una nuova miniera informativa utile per la produzione di informazioni. Le nuove tecnologie di comunicazione offrono opportunità di raccolta di dati semplificate che dovrebbero ridurre l'onere per le imprese e migliorare la qualità delle informazioni statistiche. La creazione della raccolta elettronica dei dati, i sistemi di elaborazione, l'uso dei dati forniti dalle imprese nei loro bilanci annuali, i dati dei social network e la creazione di tassonomie XBRL integrate rappresentano

le *'new sources'* utilizzabili al fine di semplificare il trasferimento dei dati dalle aziende agli istituti nazionali di statistica e per rendere il processo di produzione statistica più efficiente. L'uso di queste fonti rappresenta una grande opportunità per gli istituti nazionali di statistica ancora non sufficientemente sfruttata a causa dei problemi connessi alla raccolta dei dati. Per contribuire al raggiungimento di questo obiettivo questo lavoro di tesi propone strumenti di text mining utili a facilitare il maggiore uso dei documenti espressi in linguaggio naturale.
In particolare è stato proposto l'uso dell'analisi delle corrispondenze lessicali unitamente alla *network analysis* per la costruzione di risorse statistico linguistiche. Inoltre è stata proposta una strategia di *text classification*, per la costruzione di strumenti di interrogazione di testi: le *query* testuali. In ultimo, è stata proposto l'uso di un metodo fattoriale vincolato (analisi delle corrispondenze canoniche), per una analisi congiunta di variabili quantitative e testuali. Questo strumento consente di arricchire e comprendere i dati numerici con l'ausilio di dati testuali (parole).
A titolo di esempio sono presentate alcune applicazioni a dati reali.
Nello specifico la tesi si articola in sei capitoli.
Nel primo capitolo vengono definiti i contorni della statistica ufficiale. In particolare viene presentato il quadro normativo (europeo e nazionale) che disciplina questa attività. Inoltre si descrive il contesto economico ed evolutivo nel quale la statistica ufficiale si trova ad operare. Si pongono quindi in rassegna critica i tradizionali obiettivi e le tradizionali procedure di raccolta utilizzate per la produzione di statistiche ufficiali, partendo dalla relazione scritta dal Presidente dell'Istat Enrico Giovannini, in occasione della X conferenza nazionale della statistica: 'La statistica 2.0: the next level'.
Nel secondo capitolo si descrivono, in maniera dettagliata, le fonti della

statistica ufficiale: le fonti primarie (indagini campionarie e censuarie) e le fonti secondarie (i dati raccolti non a fini statistici, di cui sono un esempio i dati amministrativi). Si illustrano le principali caratteristiche, mettendo in risalto i vantaggi e gli svantaggi derivanti dal loro utilizzo.
Nel Terzo capitolo viene presentata una rassegna esaustiva in ambito europeo dell'uso di 'new sources' per la produzione di statistiche ufficiali.
Nel quarto capitolo si presenta una rassegna delle tecniche classiche di analisi dei dati testuali, con una particolare attenzione alla procedura di trasformazione del testo in dato strutturato.
Il quinto capitolo contiene la componente maggiormente innovativa del lavoro, da un punto di vista propriamente metodologico. Le opportunità offerte dal ricorso a basi di dati non strutturate per la costruzione di informazione statistica, illustrata nei capitoli precedenti, si realizzano qui, attraverso la proposta di alcune procedure utili alla produzione di statistiche ufficiali partendo da basi documentarie. Da prima si affronta il problema della costruzione di risorse statistico linguistiche che consentono, in una successiva fase di strutturazione della base di dati (e, quindi, nella fase di analisi statistica), di tener conto di informazioni di contesto ignorate dagli strumenti classici di analisi dei dati testuali. Si propone una strategia basata sull'utilizzo congiunto di strumenti propri dell'analisi delle corrispondenze lessicali e della *network text analysis*, così da annotare il testo con metainformazioni, utili per la selezione dei termini rilevanti, e, in definitiva, per l'identificazione del contenuto del testo oggetto di analisi. Il problema della *high dimensionality*, caratteristica delle basi di dati documentarie, viene affrontato, in un ambito più legato all'*information retrieval*, attraverso la proposta di una strategia di *text classification* finalizzata alla costruzione di strumenti di interrogazione di testo più efficienti, perché riferite a porzioni

di corpus ritenute rilevanti sulla base delle relazioni fra termini identificate all'interno di un training set di documenti e, successivamente validate. In ultimo, si affronta un problema di grande rilievo al fine di produzione di statistica da fonti secondarie, quando si dispone di informazioni sia numeriche che testuali. In questo ambito, si propone un metodo di analisi fattoriale che analizza congiuntamente variabili numeriche (siano esse continue o categoriche) e testuali, al fine di costruire un'informazione statistica sulla base di informazioni numeriche, ma con l'ausilio di informazioni testuali. Il contesto metodologico di riferimento è quello dell'analisi delle corrispondenze canoniche.

I metodi proposti nel precedente capitolo sono stati applicati alla relazione sulla gestione delle società italiane quotate sul mercato regolamentato. Questo documento rappresenta un esempio calzante, di una risorsa secondaria (dato amministrativo), ma sottoforma di testo (nuova fonte). Nel sesto capitolo vengono presentati i risultati ottenuti.

Capitolo 1

La nuova frontiera della statistica ufficiale: produzione di informazione

La funzione statistica è un servizio pubblico dedito a fornire un quadro informativo della realtà in cui opera. Secondo i principi fondamentali della statistica ufficiale, adottati dalla commissione economica per l'Europa delle Nazioni Unite, *le statistiche ufficiali costituiscono un elemento indispensabile nel sistema informativo di una società democratica.* È importante sottolineare che il concetto di statistica pubblica, seppur correlato, non coincide con quello di statistica ufficiale:

- *il concetto di statistica pubblica fa riferimento all'insieme delle persone, dei dipendenti e delle attività coinvolte nel processo di produzione del servizio statistico;*
- *il concetto di statistica ufficiale fa invece riferimento al mandato istituzionale fornito dalla legge, che qualifica come ufficiale le produzioni*

di statistiche di alcuni enti;

dunque il carattere di ufficialità è subordinato alle disposizioni normative.

1.1 Attuale contesto normativo

In Italia, sono definite ufficiali (*D.lgs 322/89*) le statistiche prodotte dall' ISTAT (Istituto Centrale di Statistica, istituito con *legge 9 luglio 1926 n. 1162*) e dagli organismi appartenenti al Sistema Statistico Nazionale (SISTAN), cioè sia quelle incluse nel Programma Statistico Nazionale (Psn), sia quelle al di fuori del Psn, purchè validate da un ente del SISTAN. Il carattere di ufficialità oltre ad essere garantito dal mandato istituzionale degli enti produttori, è garantito dall'ampia normativa esistente in materia di raccolta, trattamento, analisi e diffusione di statistiche ufficiali. Dalla data di costituzione del sistema statistico nazionale sono stati emanati numerosi provvedimenti normativi, che hanno, in tempi diversi ma soprattutto a diversi livelli (europeo, nazionale e locale), continuamente rimodulato i profili sia strutturali che funzionali del sistema normativo della statistica. Per un riordino sistematico dell'intero quadro normativo, l'Italia, pioniera in questa iniziativa, ha designato il *'Codice Italiano della Statistica Ufficiale'* (*Direttiva n.10, Gazzetta Ufficiale n.240 del 2010*). Il codice Italiano della statistica ufficiale è fortemente ispirato al codice europeo delle statistiche (*European Statistics Code of Practice*, 2005) e rappresenta per l' intero sistema statistico nazionale il testo unico a sostegno di una cultura comune per la qualità del servizio. Ai fini della misurazione della qualità delle statistiche, l'Istat ha adottato la definizione della qualità rilasciata da Eurostat (Istituto Statistico dell'Unione Europea) nel 2003 (ESS Working Group '*Assessment of Quality in Statistic*') secondo cui la qualità viene

definita, anche in questo ambito, come '*il complesso delle caratteristiche di un prodotto o di un servizio che gli conferiscono la capacità di soddisfare i bisogni impliciti o espressi*'. I principi europei, ripresi dal codice italiano riguardano tre ambiti: il contesto istituzionale, i processi statistici e la produzione statistica. I fattori istituzionali e organizzativi influiscono in modo rilevante sull' efficienza e sulla credibilità di un autorità statistica che sviluppa, produce e diffonde statistiche. In questo ambito gli aspetti da considerare sono l'indipendenza professionale, il mandato per la rilevazione di dati, l'adeguatezza delle risorse, l'impegno a favore della qualità, la riservatezza statistica, l'imparzialità e l'obiettività. Gli *standard*, le linee guida e le buone pratiche europee e internazionali sono pienamente rispettate nei processi utilizzati dalle autorità statistiche per organizzare, rilevare, elaborare e diffondere le stesse. La credibilità delle statistiche risulta rafforzata da una reputazione di efficienza e di buona gestione. In tal senso i processi statistici devono basarsi su una solida metodologia, procedure appropriate, un onere non eccessivo sui rispondenti e un buon rapporto costi/efficacia. La produzione statistica rispetta gli standard di qualità europei e risponde alle esigenze di tutti gli utilizzatori (istituzioni europee, amministrazioni pubbliche, imprese e cittadini). Le variabili da considerare nella valutazione della qualità delle statistiche sono il grado di pertinenza, l'accuratezza e l'attendibilità, la tempestività, la coerenza, la comparabilità tra le diverse aree geografiche e la facilità di acceso per gli utilizzatori.

1.2 L'Istat e il Sistema Statistico Europeo

Fin dai primi giorni di vita della Comunità Europea ci si è reso conto che la pianificazione e l'attuazione delle politiche comunitarie dovevano basarsi

su statistiche affidabili e comparabili. Così è stato costruito il Sistema Statistico Europeo (ESS), con l'obiettivo di fornire statistiche affidabili, che seguano criteri e definizioni comuni, in modo da garantire la comparabilità delle informazioni tra i diversi paesi della UE (attualmente regolato dalla Legge statistica europea, approvata nel 2009 con il *Regolamento (CE) n. 223/2009* del Parlamento europeo e del Consiglio). L'ESS rappresenta il partenariato tra l'autorità statistica comunitaria, ovvero la commissione e gli istituti nazionali di statistica (INS), nonché le altre autorità nazionali responsabili in ciascuno Stato membro per lo sviluppo, la produzione e la diffusione di statistiche europee.

L'ESS coordina e pianifica l'intera attività statistica attraverso la compilazione del programma statistico europeo, fornendo un quadro completo delle azioni e degli obiettivi previsti per il periodo di riferimento (non superiore ai 5 anni). Anche se la pianificazione delle attività viene effettuata congiuntamente dagli Istituti nazionali di statistica e da Eurostat, la produzione di statistiche armonizzate nazionali fa capo alle autorità dei singoli stati membri (mentre Eurostat raccoglie i dati forniti dagli Stati, li analizza e li diffonde a livello europeo). In riferimento a questo principio, ogni Stato membro, in conformità con gli obiettivi definiti nel programma statistico europeo, compila il Programma Statistico Nazionale (PSN). In Italia l'organo preposto alla compilazione del PSN è il comitato di indirizzo e coordinamento dell'informazione statistica (Comstat), organo dell'Istat, che ne predispone le linee guida.

Le esigenze che scaturiscono dagli obiettivi di armonizzazione e integrazione dell'informazione statistica a sostegno delle politiche dell'Unione Europea richiedono una sistematica presenza dell'Istat, come degli Istituti nazionali di statistica degli altri stati membri.

Il contributo dell'Istat si concretizza in una crescente partecipazione alle attività dei comitati, gruppi di lavoro, *task force* che si riuniscono presso il Consiglio dell'Unione Europea in vista dell'adozione di atti legislativi; inoltre partecipa a gruppi di interesse in specifiche aree, istituiti presso Eurostat, che rappresentano un *forum* nel quale vengono discusse le varie tematiche in ambito statistico. Essi contribuiscono a garantire il necessario coordinamento e la condivisione degli aspetti strettamente tecnici e a definire linee e indirizzi strategici e di programmazione che confluiscono più specificatamente nel Programma Statistico Comunitario Pluriennale e nel Programma Statistico Annuale.
L'Istituto partecipa come diretto interlocutore nei processi decisionali relativi alla definizione e all'implementazione di metodologie e *standard* per l'armonizzazione delle statistiche europee e come fornitore delle informazioni necessarie alla produzione della statistica europea anche in relazione al *Regolamento (CE) n. 223/2009* del Consiglio dell'Unione Europea. L'Istat partecipa inoltre ai progetti europei di Ricerca e Sviluppo, in forte crescita sotto la spinta del messaggio europeista, nella convinzione che le attitudini intellettuali dei ricercatori europei, sviluppata in contesti ambientali, culturali e istituzionali diversi, costituiscano un patrimonio di grandi potenzialità da sfruttare come un bene comune.

1.3 La Statistica 2.0: verso un nuovo sistema di statistiche ufficiali

Il Novecento è stato definito come 'il secolo della statistica'. Questa definizione, seppur azzardata, trova le sue ragioni nello scenario evolutivo che ha caratterizzato il secolo scorso.

Lo sviluppo delle tecnologie di informazione e comunicazione hanno consentito, ad un numero sempre più elevato di soggetti, di diffondere statistiche provenienti da fonti amministrative e da rilevazioni campionarie. Sul *web* è oggi possibile accedere in tempo reale ad una massa sconfinata di dati; la televisione quotidianamente diffonde statistiche sui temi più disparati; i giornali pubblicano grafici e indicatori semplici o complessi per informare i propri lettori sull'andamento economico nazionale ed europeo. Ma al crescere delle fonti disponibili e delle dichiarazioni pubbliche, cresce la confusione degli analisti e della pubblica opinione, i quali si trovano spesso a consultare dati contraddittori sugli stessi fenomeni.
Allo stesso tempo, la crisi economica, la globalizzazione dei mercati, l'abbattimento delle frontiere e l'ambizioso tentativo di condivisione di una prospettiva politica planetaria (Dichiarazione del Millennio delle Nazioni Unite) hanno generato un esponenziale incremento delle esigenze informative.
Se dunque, questo è il mondo in cui viviamo, è necessario riflettere se e in che modo, la statistica ufficiale riesce a garantire la qualità del prodotto offerto e a froteggiare le esigenze informative sopra indicate. Come per un'impresa, parlare di valutazione del prodotto offerto significa identificare il valore che lo stesso produce nel consumatore. L'analisi del valore dell'attività statistica è stata affrontata dal presidente dell'Istat Enrico Giovannini, in occasione della decima conferenza nazionale della statistica; Giovannini afferma che dalla lettura di un grafico o di un dato statistico l'utente dovrebbe accrescere la sua conoscenza [1]. Secondo questa prospettiva, è necessario passare da una concezione di produzione della statistica ufficiale basata sul numero di microdati prodotti o di volumi stampati, ad una basata sull'aumento di conoscenza della realtà nella popolazione. Infatti

mettere al centro della valutazione del servizio l'utente e il processo cognitivo che egli compie per trasformare dati in informazioni comprensibili e poi in conoscenza cambia completamente la prospettiva. Dalla 'Statistica 1.0' bisogna passare alla 'Statistica 2.0'.
Il passaggio da compiere non è né facile né automatico. Passare ad un livello superiore del servizio statistico vuol dire non interrompere la *production chain* al momento della diffusione dell'informazione, ma proseguire, curandosi di come quest'ultima sia trasferita all'utente finale dai media, di quanto gli utenti si fidino di quelle informazioni (e quindi dell'istituzione che le produce), nonché della loro capacità di trasformare i dati in conoscenza. Le leve di miglioramento da apportare affinché si passi al *next level* riguardano diversi ambiti: il quadro istituzionale, la tecnologia, la cultura e le risorse umane.
Per discutere del quadro istituzionale della statistica ufficiale italiana è necessario, in primo luogo, riflettere su quello europeo. Un generico sistema di *governance* statistica, che determina indipendenza e qualità dell'informazione prodotta è basato su sei elementi chiave [1]:

1. il meccanismo di nomina del presidente e dei *manager* dell'INS (e degli altri enti pubblici);

2. il sistema di finanziamento per la determinazione delle risorse disponibili per la statistica pubblica;

3. gli strumenti con i quali viene definita la 'domanda' di informazione statistica;

4. le regole attraverso le quali vengono stabilite le classificazioni, i concetti e le definizioni utilizzate per la produzione delle statistiche e le

metodologie di raccolta e di elaborazione dei dati ;

5. le procedure attraverso cui le informazioni statistiche vengono diffuse al pubblico;

6. le regole con cui l'istituto di statistica accede alle informazioni detenute da altre enti (tipicamente quelle amministrative), protegge i dati raccolti a fini statistici e li rende disponibili per fini di ricerca.

Per affrontare i problemi sopra citati bisognerebbe creare un nuovo sistema statistico, analogo a quello delle banche centrali. Si ritiene che l'Eurostat dovrebbe diventare un'istituto autonomo con un *budget* definito dal Parlamento Europeo, gli INS dovrebbero avere lo stesso status delle banche centrali nazionali, ed insieme l'Eurostat e gli INS dovrebbero essere dotati di potere regolamentare in materia statistica. Una tale proposta, richiede una modifica dei trattati e quindi è perseguibile solo nel medio lungo termine.
Accanto ad un quadro giuridico volto a rafforzare l'autonomia delle istituzioni statistiche, la tecnologia rappresenta il secondo ingrediente utile per il passaggio al *next level*. Inutile sottolineare quanto lo sviluppo delle tecnologie abbiano già contribuito a migliorare il servizio statistico, in tutti le fasi del suo processo produttivo, ma è evidente che la diffusione dei nuovi *mobile devices* e lo sviluppo dei motori di ricerca e del linguaggio XML, rivoluzioneranno interamente le modilità di raccolta, diffusione e comunicazione dei dati. In particolare, lo sviluppo del linguaggio XML ha permesso di designare il cosidetto '*web semantico*', termine (coniato da Tim Berners-Lee) con il quale si intende il nuovo ambiente *web*, nel quale i documenti pubblicati sono associati ad informazioni e metadati che ne specifichino il contesto semantico. Con tale sistema saranno possibili ricerche molto più

evolute delle attuali. Discorso analogo vale per l'informazione resa disponibile con l'uso dei *mobile devices*. Grafici dinamici, mappe interattive e video aumenteranno la visibilità delle informazioni statistiche aumentando la conoscenza negli utenti. L'impatto delle nuove tecnologie non si fermerà, naturalmente, agli aspetti di comunicazione e diffusione, ma consentiranno di costruire sistemi informativi complessi in ambienti aperti, quindi senza la necessità di piattaforme *hardware* e *software*.
Le nuove sfide da affrontare per un'impresa, pubblica o privata che sia, non possono essere superate senza il contributo del capitale umano e di un suo profondo cambiamento culturale. Per passare alla 'Statistica 2.0' gli INS devono, 'da produttori di informazione' trasformarsi in 'generatori di conoscenza'; ciò significa non limitare il proprio campo di attività, essere aperti a misurare fenomeni emergenti importanti per la società, guardare alle nuove generazioni e al loro modo di apprendere e porre le esigenze informative dell'utente al centro dell'attività statistica.

1.4 Dal dato all'informazione

Per comprendere chiaramente il nuovo orientamento della statistica ufficiale è indispensabile definire alcuni concetti importanti: dato, informazione e conoscenza. Da una semplice ricerca sul *web*, è possibile consultare le diverse definizioni di dato:

- un dato è il risultato di una misurazione su una unità appartenente ad una collettività (Istat);

- un dato è un simbolo (Russell Ackoff);

- un dato è tipicamente un valore che supporta i processi operativi aziendali (Mark Cudmore);

- un dato è un elemento determinato utilizzabile per eseguire una ricerca, per compiere un'elaborazione, per esprimere un giudizio (Dizionario Italiano Hoepli);

Seppur queste definizioni siano differenti nella forma, in ognuna di loro il concetto di dato viene associato ad un elemento statico, che necessita di essere contestualizzato, per essere capito.
Nell'ambito della statistica ufficiale un dato rappresenta il prodotto grezzo che viene raccolto, trasferito e analizzato per la produzione delle statistiche. Queste ultime rappresentano il prodotto finito dell'attività statistica e solo nel momento in cui vengo diffuse diventano informazioni utili per gli utenti che le consultano. Quindi il processo di trasformazione del dato in informazione avviene quando l'utente è in grado di utilizzare i dati per soddisfare le proprie esigenze informative. Quindi l'informazione è conoscenza? Con Einstein riteniamo che la conoscenza è qualcosa di diverso dalla semplice informazione; quest'ultima esiste indipendentemente da chi la possa utilizzare, e quindi può in qualche modo essere preservata su un qualche tipo di supporto (cartaceo, informatico, etc), al contrario la conoscenza esiste solo se 'c'è una mente in grado di possederla' (Wikipedia). Quindi è possibile definire la conoscenza come il processo cognitivo attraverso il quale l'utilizzatore collega le informazioni alla propria esperienza personale e le immagazzina nella propria mente, arricchendo il suo sapere.
Precedentemente abbiamo affermato che il valore del servizio statistico 2.0 è la quantità di conoscenza prodotta. In virtù delle considerazioni trattate in questo paragrafo, per accrescere la conoscenza nei fruitori del servizio,

la statistica ufficiale deve migliorare gli strumenti di comunicazione e diffusione delle statistiche, rendendo la loro consultazione semplice e chiara. Quindi, insieme alle tabelle *standard*, dovranno essere utilizzati qualunque formato (grafici, testi) che sia in grado di facilitare la consultazione dei dati.

1.5 La comunicazione con le imprese

Gli utilizzatori del servizio statistico sono molteplici ed eterogenei tra loro. Le imprese rappresentano una delle categorie di maggiore interesse, in quanto sono, contestualmente, i principali fornitori di dati ed utilizzatori delle statistiche in ambito economico(progetti: blue-ets, ESSnet AdminData). Allo stato attuale le imprese, in qualità di fornitori, sono continuamente 'disturbate' con richieste di questionari per le indagini statistiche attive, ma non sempre riescono ad utilizzare i dati prodotti dagli INS per conoscere la realtà nella quale operano. La difficile fruibilità delle statistiche viene generata dall'offerta smisurata di fonti disponibili, di dati molto aggregati che sono di difficile consultazione. Questo genera due problemi: aumenta la percezione dell'onere statistico gravante sui rispondenti e non fornisce strumenti conoscitivi idonei alle definizione di opportune strategie di imprese. Sull'onere statistico gravante sui rispondenti, è riposta una grande attenzione da parte delle autorità statistiche europee, che hanno incluso tra i principi dell'*European Statistics Code of Practice*, in relazione alla efficienza di un processo statistico, il rispetto di un carico non eccessivo sui rispondenti. Essendo questa la principale causa delle mancate risposte e essendo quindi pregiudizievole della qualità delle statistiche, rappresenta una delle principali sfide che la statistica ufficiale deve affrontare. Nell'ambito del progetto europeo BLUE-ETS (*work package* 3) è stata condatta un'indagi-

ne sulla comunicazione esistente tra gli INS e le imprese perche' considerata una delle leve di miglioramento per ridurre il response burden. I risultati ottenuti hanno rilevato che gli INS utilizzano come principale strumento di comunicazione e diffusione dei dati il loro sito *web*; ma spesso i dati sono accompagnati da una documentazione (meta dati) lunga e complessa, che non consente nemmeno ai principali motori di ricerca (es. Google) una loro facile leggibilità, e considerando che la maggioranza delle persone non va oltre la prima pagina, il ruolo della statistica ufficiale nel mondo dell'informazione diventa marginale. Dal punto di vista delle imprese, le statistiche sono considerate affidabili (percezione della fonte 'ufficiale') e quindi potenzialmente utilizzabili nel contesto aziendale, ma quasi mai riescono a trovare le informazioni rilevanti (ad esclusione dei prezzi, indici di prezzi, indicatori di settori e dei generici indicatori economici). Questo genera una percezione negativa dell'utilità del servizio statistico e quindi la motivazione per i rispondenti a partecipare alle indagini è legata esclusivamente al loro obbligo legale [2]. Tutto questo, chiaramente, influisce negativamente sull'aumento di conoscenza negli utenti, che in questo caso sono rappresentati dalle imprese.

La statistica ufficiale deve quindi sviluppare nuovi strumenti utili a ridurre il gap esistente tra l'onere statistico e la fruibilità del servizio.

Capitolo 2

Le fonti della statistica ufficiale

Quando ci si accinge ad effettuare uno studio di qualsiasi tipo e su qualsiasi aspetto della società e della realtà di un territorio, demografico, economico e sociale, la prima cosa da fare è individuare, raccogliere ed organizzare tutte quante le possibili informazioni esistenti relative al fenomeno oggetto di studio. Definito l'obiettivo, la prima cosa da fare è individuare la fonte dei dati. Per fonte dei dati possiamo intendere: l'ente o l'organizzazione che produce ed è in possesso del dato; la rilevazione e l'ente da cui proviene il dato (raccolta diretta per l'indagine ad hoc, oppure elaborazioni di dati contenuti in archivi amministrativi); la pubblicazione su carta, *cd rom*, o sito *web* su cui è rilasciato il dato (ad esempio l'annuario statistico italiano o il data base contenente i dati). Il compito di un istituto statistico nazionale è quello di produrre e diffondere statistiche di qualità capaci di descrivere le condizioni demografiche, sociali, economiche e ambientali del paese e rilevare i cambiamenti che avvengono in esso, con il vincolo del più rigoro

rispetto della *privacy*.
In questo capitolo si descriveranno i due tipi di fonti statistiche: primarie e secondarie, evidenziando le loro peculiarità.

2.1 Le fonti primarie

Un istituto statistico per produrre statistiche deve innazittutto identificare le fonti di raccolta dati. Sono definite primarie, le fonti che forniscono dati a fini statistici. La fonte primaria disponibile per un istituto statistico è l'indagine. L'indagine è un processo di raccolta dei dati, che nasce esclusivamente per finalità statistiche. Con l'utilizzo di questo strumento è possibile reperire esattamente le informazioni di cui si necessita. Si può dire, quindi, che è la fonte più importante per un istituto statistico nazionale, per lo studio di un fenomeno oggetto di interesse. Esistono due tipi di indagini: censuaria e campionaria, ampiamente descritte di seguito.

2.1.1 Le metodologie d'indagine

Un sondaggio è qualsiasi attività che raccoglie informazioni in modo organizzato e metodico sulle caratteristiche d'interesse di alcune o tutte le unità di una popolazione [3]. Di solito un'indagine inizia dall'esigenza di informazioni di un ente pubblico o di un suo reparto, per la quale non esistono dati disponibili, oppure risultano essere insufficienti. Per svolgere un'indagine che sia in grado di fornire informazioni accurate e significative è necessario seguire una procedura minuziosa e precisa. In particolare, percorrendo i seguenti *step* [4]:

- definizione degli obiettivi;

- selezione di un piano d'indagine;
- determinazione del piano di campionamento;
- progettazione del questionario;
- raccolta dei dati;
- acquisizione e codica dei dati;
- editing e imputazione;
- stima;
- analisi dei dati;
- diffusione dei dati;
- documentazione.

Di seguito è riportata una breve descrizione di ogni passo. La definizione degli obiettivi rappresenta un passo fondamentale per lo svolgimento dell'intero processo, in quanto specifica le definizioni operative da utilizzare e il piano di analisi. Fissati gli obiettivi, si costruisce il piano d'indagine [5], che ha la forma di una lista, nella quale sono incluse tutte le unità considerate interessanti. Il passo successivo prevede la determinazione del piano di campionamento. Esistono due tipi di indagine: per campione e indagine censuarie. Nel secondo caso vengono raccolti i dati di tutte le unità appartenenti alla popolazione di interesse, al contrario, nel primo caso vengono raccolti solo i dati di alcune unità considerate rappresentative dell'intera popolazione. Le principali variabili da considerare per la scelta di un tipo d'indagine o l'altro, sono i costi i tempi e l'entità del fenomeno oggetto di

studio [6]. Per la determinazione del campione esistono due metodi di campionamento probabilistico e non probabilistico [7]. L'ufficio statistico ai fini di un corretto svolgimento dell'indagine deve quantificare l'entità dell'errore. Per il metodo di campionamento probabilistico esistono diverse metodi di stima, la misura più comunemente utilizzata è la varianza [8]. Con l'utilizzo di questo metodo possono essere prodotte stime più attendibili sull'errore di campionamento e quindi è possibile fare deduzioni più attendibili sulla popolazione. Oltre all'errore di campionamento un'indagine è anche soggetta ad una varietà di errori non legati al processo di campionamento. Comunemente questi errori sono chiamati extra-campionari, cioè gli errori derivanti dall'intero processo d'indagine. Questo tipo di errori sono presenti in entrambi i tipi d'indagine (per campione, censimento), e possono essere classificati in due gruppi: errori casuali e errori sistematici. Gli errori casuali non seguono nessuna distribuzione e al crescere delle dimensioni del campione i loro effetti tendono ad annullarsi. Gli errori sistematici, invece tendono ad andare verso la stessa direzione producendo una distorsione sui risultati finali, a differenza della precedente categoria al crescere della dimensione la distorsione non si riduce. Questo tipo di errori è molto difficile da identificare e quindi da misurare, infatti desta la maggiore preoccupazione dell'analista. Gli errori extra-campionari derivano principalmente dalle seguenti fonti: copertura (omissioni, inclusioni errate, duplicazioni e errori di classificazione delle unità nel telaio sondaggio), misura (differenza tra la risposta registrata e il valore vero, causata da domande mal formulate, linguaggio tecnico o da una formazione inadeguata dell'intervistatore), mancata risposta (parziale o totale) e dall'elaborazione (errore di codifica) [9]. Un'altra delle fasi cruciali di una indagine è la progettazione del questionario. Un questionario mal progettato può con-

dizionare l'intera indagine, producendo dati di bassa qualità. I problemi legati alla loro progettazione sono: decidere quali domande porre, quali parole utilizzare al fine di garantire una loro chiara leggibilità e comprensione, e come disporre le domande per ottenere le informazioni richieste. Mentre, ci sono principi ben consolidati per la costruzione di un questionario, una sua buona lavorazione rimane un'arte che richiede ingegno ed esperienza. Infatti se i dati richiesti non sono strutturati in modo da essere analizzati con strumenti di analisi, anche da un buon campione possono uscire risultati negativi.

Quando il questionario è pronto, si passa alla fase di raccolta delle informazioni. Per una panoramica completa dei metodi di raccolta dei dati consultare Cox [10]. Una volta che i dati sono stati raccolti si passa alla fase di codifica. Questa fase è lunga e dispendiosa, ma importante, in quanto trattandosi per lo più di attività manuale, i codificatori possono commetere molti errori, che chiaramente compromettono la qualità del dato finale. È buona pratica la prevenzione degli errori nelle fasi iniziali e l'uso di metodi di controllo e monitoraggio degli stessi. Per *editing* si intendono le operazioni di controllo al fine di rilevare i dati mancanti, dati incoerenti e dati non validi. Questa operazione può essere effettuata dall'intervistatore sul campo oppure da un un *software* specifico. Alcuni errori (la maggioranza) vengono corretti e modificati con i metodi richiamati, ma è impossibile auspicare la loro totale correzione; quindi per gestire i casi rimanenti si utilizza l'imputazione. L'imputazione è un processo grazie al quale ai dati mancanti, non validi e incoerenti viene imputato un valore sostitutivo. I metodi di imputazione possono essere raggruppati in due categorie: stocastici o deterministici. Quest'ultimi imputano valori uguali a tutte le unità aventi le stesse caratteristiche; invece i metodi stocastici, anche ad unità aventi

le stesse caratteristiche, possono imputare valori differenti. Questo gruppo di metodi si caratterizzano per la presenza di una componente aleatoria, detta anche residuo, corrispondente ad uno schema probabilistico associato al particolare metodo d'imputazione prescelto. I metodi di imputazione più utilizzati sono: il metodo dedutttivo, *cold deck*, il vicino più prossimo e il valore medio [11]. Il metodo deduttivo è un metodo in base al quale un valore mancante o incoerente pu'o essere dedotto con certezza. Ad esempio, in una somma di quattro elementi con un totale previsto pari a 100, e solo due di questi si presentano con valori pari 40 e 60, e gli altri due elementi sono lasciati in bianco, si pu'o dedurre che il loro valore è pari a zero. Il metodo *cold deck* invece prevede che il valore mancante viene imputato con un valore proveniente da una fonte esterna. Al contrario di ci'o che avviene con l'uso di questo metodo, con il metodo del vicino più prossimo il valore da imputare proviene da un donatore appartenente alla base dati. È chiaro che le caratteristiche del donatore devono essere molto simili a quelle del non rispondente. In ultimo, con il metodo del valore medio, al valore mancante o incoerente, viene imputato il valore medio della classe a cui appartiene l'unità. Questo chiaramente potrebbe alterare le relazionali esistenti nei dati, quindi è necessario scegliere una metodologia di imputazione appropriata. Una volta che i dati sono stati raccolti e codificati si puó procedere con la fase di stima. La stima è il mezzo con cui l'ufficio statistico, sulla base delle informazioni raccolte sul campione, ottiene valori sulla popolazione di riferimento. A questo punto è possibile effettuare la vera e propria analisi dei dati. L'analisi dei dati può essere limitata ai soli dati delle indagini o può confrontare le stime dell'indagine con i risultati ottenuti da altre indagini. I risultati spesso si presentano sottoforma di tabelle, grafici, e misure riassuntive diverse (distribuzioni di frequenze, medie,

etc), utili a sintetizzarli. Inoltre, vengono utilizzate tecniche di inferenza statistica, come per esempio il modello di regressione, analisi della varianza, il *test* del chi-quadrato, per verificare ipotesi o studiare le relazioni esistenti tra le variabili. Quando i risultati dell'indagine sono stati ottenuti, devono essere rilasciati agli utenti, attraverso i vari mezzi di comunicazione, come per esempio, un comunicato stampa, una intervista televisiva o radiofonica, un file di micro dati, ma sempre nell'assoluto rispetto della *privacy* degli intervistati. Come già precedentemente affrontato, la comunicazione dei dati è un tema molto delicato. Per gli utenti dovrebbe essere facile trovare i dati, e una volta trovati dovrebbero essere in grado di comprenderli, utilizzarli ed interpretarli correttamente, infatti è buona norma che i risultati siano accompagnati da una documentazione chiara ed esaustiva utile a fornire agli utenti un contesto per un utilizzo consapevole dei dati. Se ció non avviene, l'indagine perde la sua utilità.

2.1.2 La qualità dei dati

La produzione di un dato statistico si compone di una serie di attività o fasi eseguite in sequenza. In ogni fase vi sono molte potenziali fonti di errori che influiscono negativamente sulla qualità del dato finale. Gli errori nei quali si può incorrere, sia nel caso di una rilevazione per campione sia di un censimento, sono di diversa natura e origine, ma principalmente riconducibili a problemi di mancata osservazione delle unità e a problemi di misurazione. Questi ultimi sono collegabili agli strumenti di rilevazione, alle tecniche di raccolta delle informazioni, agli intervistati e agli intervistatori, quindi denominati non campionari. Al contrario, gli errori derivanti dalla mancata misurazione sono detti errori campionari.
Un dato statistico è di buona qualità se il suo valore è vicino (non distorto)

alla realtà, e quindi se si commettono il minor numero di errori durante il suo processo di produzione. La qualità di un dato è esprimibile attraverso quattro dimensioni: rilevanza, accuratezza, tempestività e accessibilità. La rilevanza si può definire come la capacità dell'informazione di soddisfare le esigenze conoscitive degli utenti. È interessata a controllare se le informazioni disponibili fanno luce sui problemi di maggiore importanza degli utenti. Mentre l'accuratezza delle informazioni statistiche è il grado di corrispondenza tra la stima ottenuta dall'indagine e il vero (ma ignoto) valore della caratteristica in oggetto nella popolazione obiettivo. È descritta in termini di errori nelle stime statistiche ed è tradizionalmente scomposta in componenti di distorsione (errore sistematico) e di varianza (errore casuale). La tempestività delle statistiche è la capacità di produrre i risultati in termini ravvicinati rispetto all'esecuzione dell'indagine. In altre parole, è l'intervallo di tempo che intercorre tra il momento della diffusione dell'informazione prodotta e l'epoca di riferimento della stessa. Questa dimensione è strettamente connessa alla necessità di disporre di dati aggiornati e quindi va valutata con riferimento al fenomeno osservato ed alle esigenze degli utilizzatori. In ultimo, l'accessibilità delle informazioni statistiche si riferisce alla semplicità per l'utente di reperire, acquisire e comprendere l'informazione disponibile in relazione alle proprie finalità. Queste caratteristiche sono influenzate dai mezzi di diffusione dei risultati ottenuti.

A queste dimensioni, si aggiungono altre caratteristiche della qualità che ne aumentano l'importanza, che sono la confrontabilità, che punta a confronti attendibili di statistiche accessibili attraverso lo spazio, tra domini tematici e nel tempo; e la coerenza che implica relazioni chiare e semplici tra basi di dati.

Infine, in termini di versatilità in un campo di contesti e situazioni di utiliz-

zo dei dati, si deve aggiungere un'altra dimensione della qualità di un dato: la completezza. Quest'ultima garantisce la esaustività di un dato rispetto al fenomeno oggetto di analisi.

2.2 Le fonti secondarie

Gli organismi statistici di tutto il mondo sono da sempre chiamati a migliorare l'efficienza del loro processo di produzione. Allo stesso tempo vi sono crescenti richieste politiche affinchè si riduca l'onere statistico gravante sui rispondenti. Questo è, in particolare, il caso delle imprese, dove molti governi vedono la riduzione della burocrazia come un elemento fondamentale per sostenere e promuovere lo sviluppo economico. Alla luce di queste pressioni gli statistici sono stati costretti a vedere se i dati utilizzabili esistono già in natura. Quindi oltre alle fonti primarie, si considerano delle fonti di dati alternative, chiamate fonti seconderie.
Le fonti secondarie nascono dalla presenze di molte organizzazioni non statistiche che raccolgono dati, in diverse forme, e su differenti temi, utili alla produzione statistica. In letteratura, sono state proposte diverse definizioni di fonti secondarie, qui ci limitiamo a ripotarne solo 2. La prima, considerata più stringente, identifica le fonti secondarie come dati amministrativi [12], successivamente, nel 1996, una *task force* di Eurostat ha definito le fonti secondarie come 'il complesso dei dati raccolti per fini non statistici, quindi includendo anche i dati in possesso di organizzazioni private. D'ora in avanti faremo riferimento alla definizione più ampia di fonti secondarie.

2.3 I dati amministrativi

I vincoli di bilancio e le preoccupazioni circa il peso statistico gravante sui rispondenti, hanno portato gli uffici di statistica ad esaminare metodi alternativi per ottenere i dati. Una delle alternative è l'uso dei dati amministrativi (per maggiori dettagli si veda il progetto *ESSnet AdminData*, nato per incentivare e diffondere l'uso dei dati amministrativi per la produzione di statistiche economiche). Questa pratica si è sviluppata grazie allo sviluppo dell'informatica nella pubblica amministrazione (PA), che ha reso disponibile una grande quantità di dati su imprese, istituzioni e individui, quali la fiscalità, la previdenza sociale, l'assicurazione sanitaria, occupazione, assegni familiari, etc.
L'utilità dei dati amministrativi a fini statistici dipende dai loro concetti, dalle loro definizioni, dalla loro copertura, dalla qualità con cui sono stati riportati i dati, dalla loro tempestività e disponibilità. Questi fattori possono variare a seconda del tipo di fonte amministrativa considerata. Prima di decidere di utilizzare i dati amministrativi, dovrebbero essere valutati attentamente le seguenti condizioni [13]:

- La tempestività. Un sondaggio che utilizza solo dati amministrativi pu'o essere in grado di produrre risultati in modo più tempestivo di un'indagine campionaria. Vicerversa, l'uso dei dati amministrativi pu'o essere più lenta se i dati devono essere raccolti da diverse giurisdizioni di governo, e quindi al momento del recepimento la fase di elaborazione potrebbe essere lunga e complessa.

- I costi. Molte delle operazioni da svolgere per un'indagine possono essere eliminate con l'uso dei dati amministrativi (per esempio, la fase

di raccolta), riducendo così i costi.

- Response burden. L'uso dei dati amministrativi (totale) riduce a zero l'onere di risposta gravante sui rispondenti.

- La Copertura. La popolazione *target* viene definita dalla struttura dei dati amministrativi, che spesso non coincide con le esigenze statistiche, e quindi produce una copertura non totale delle unità d'interesse a fini statistici.

- Il contenuto. I dati amministrativi nascono per soddisfare esigenze informative della pubblica amministrazione, quindi non tutti i temi di interesse statistico sono coperti da questi.

I dati amministrativi possono essere utilizzati in modo e per scopi differenti. I processi produttivi che utilizzano i dati amministrativi possono essere raggruppati in 3 tipologie: produzione diretta da una fonte amministrativa, produzione diretta attraverso l'integrazione di più fonti amministrative e l'utilizzo di queste ultime a supporto delle indagini statistiche (censuarie e campionarie).
Nel primo caso è necessario che si verifichino alcune condizioni: che la variabile di interesse sia oggetto di 'osservazione' dell'ente amministrativo, che la qualità dei dati sia adeguata allo scopo statistico e che la popolazione di riferimento sia uguale oppure contenuta nella popolazione definita dalla norma generatrice dell'archivio. Se queste condizioni non sono verificate si deve procedere alla trasformazione delle informazioni originarie in dati statistici. Nel secondo caso, dove per realizzare un dato statistico è necessario procedere all'integrazione di più archivi amministrativi, diventano di

fondamentale importanza i protocolli esistenti per la creazione dei registri statistici (che rappresentano la base informativa del processo produttivo degli INS). Nell'ultimo caso i dati amministrativi possono essere utilizzati durante tutto il processo di produzione standard; in particolare, per: la definizione della strategia campionaria, disegno di campionamento, controllo e correzione dei dati d'indagine, analisi di copertura, etc. L'introduzione dei dati amministrativi nella filiera produttiva, ha generato l'attivazione di nuove linee di produzione modificando i tradizionali processi produttivi.

2.3.1 I metodi per l'uso dei dati amministrativi

Ogni ente pubblico che raccoglie informazioni potenzialmente utili anche a fini statistici, gestisce tali informazioni con regole coerenti alle proprie finalità istituzionali. Per un corretto uso dei dati amministrativi, nel contesto statistico, è quindi necessario che ci sia un modello di acquisizione e di trattamento dei dati armonizzato alle finalità di tale contesto. I due principali fattori che ostacolano gli INS alla piena attuazione di tale principio sono: la gestione dei dati amministrativi *(ex-post)* acquisiti con definizioni, classificazioni e modalità di trattamento, determinati dall'ente fornitore, e la mancanza di protocolli che consentano di intervenire *(ex-ante)* nella modalità di acquisizione di tali dati [14]. Per descrivere le problematiche connesse alla gestione dei dati amministrativi *ex-post*, è necessario fare una distinzione tra il loro uso diretto o il loro uso indiretto (utilizzati per l'integrazione di altre fonti).
Se i dati amministrativi vengono utilizzati per integrare quelli provenienti da altre fonti, l'organizzazione statistica deve trovare un modo di collegamento di tali dati [15]. Concretamente, questo vuol dire collegare *(record-*

linkage) dati provenienti da diverse fonti sulla base di caratteristiche comuni presenti in esse, questo tipicamente assume la forma di corrispondenza *(matching)*. Se dunque, tutte le caratteristiche presentano un numero di identificazione comune, il processo può essere indicato come corrispondenza esatta. Ma anche quando c'è corrispondenza esatta, non è consigliabile procedere alla loro fusione, in quanto basta la presenza di un errore in uno dei file, per compromettere la qualità dei dati. Per ovviare a questi problemi si possono utilizzare cifre di controllo.
In alternativa al metodo dell' esatta corrispondenza, è possibile utilizzare una corrispondenza probabilistica. Questo metodo seleziona solo alcune variabili, considerate maggiormente distintive, tra le quali il nome e cognome, indirizzo e data di nascita, per identificare la corrispondenza esistente tra le diverse forme. Pertanto, la scelta accurata di 'corrispondenti chiavi', tenendo conto del concetto di potere distintivo, può avere un impatto significativo sul successo di un tale processo. La corriposndenza può essere effetuata manualmente *(matching manuale)* oppure con strumenti automatici *(matching automatico)*. L'uso del primo strumento permettere di effettuare una procedura molto accurata, ma a costi elevatissimi; la seconda, utilizzando uno strumento automatico, riduce al minimo l'intervento umano, ma di contro ha una intelligenza limitata. La soluzione migliore e quindi quella di utilizzare uno strumento automatico per trovare le corrispondenze ovvie, e di sottoporre le stesse al controllo umano.
Riguardo la capacità di intervento *ex-ante*, degli istituti statistici, sulla 'costruzione' dei dati amministrativi, si è solo agli inizi. Questo ostacolo è intrinsecamente legato alle diverse culture e normative vigenti nei diversi paesi europei. Possono essere identificati due punti di debolezza: il diritto degli INS di accedere ai dati amministrativi è stabilito solo in linea di princi-

pio, e quindi devono essere rilasciate singole autorizzazioni per ogni singolo caso; inoltre non è stabilito chiaramente un dazio corrispondente per i titolari di dati amministrativi, e quindi, i termini della fornitura devono essere negoziati di volta in volta. Ma il punto focale è che la legge non conferisce nessun potere agli INS di coordinamento con le pubbliche amministrazioni sulla progettazione, per fini comuni, dei documenti amministravi, dei sistemi informativi, per l'adozione di definizioni e classificazioni condivise. Di contro, nella maggior parte dei paesi, la legge identifica l'INS, come coordinatore di un più ampio sistema statistico nazionale, nel quale solitamente sono inclusi tutti i principali detentori di dati amministrativi. Nonostante ciò, il coinvolgimento degli INS nella progettazione o revisione dei sistemi informativi amministrativi, resta totalmente marginale [16]. Verso questo orientamento la Francia e i paesi nordici rappresentano chiaramente l'avanguardia europea, utilizzando le indagini solo, ove necessario, per integrare i dati amministrativi. Un miglioramento della cooperazione tra gli enti pubblici e gli INS nella situazione attuale, sembra essere la strada principale per aumentare l'uso dei dati amministrativi e ridurre il *response burden*.

2.3.2 I registri statistici

Tipicamente un registro è una sorta di elenco di unità strutturato, che contiene un certo numero di attributi per ciascuna di tali unità, che ha un regolare meccanismo di aggiornamento. In questo modo, molti file di dati amministrativi possono essere considerati registri. Un registro statistico è un registro che viene costruito e mantenuto a fini statistici, secondo concetti e definizioni statistiche, e sotto il controllo di statistici. Quindi i registri amministrativi possono essere utilizzati come fonte utile alla costruzione di un registro statistico. Ma se le statistiche sono prodotte direttamente da

un'unica fonte amministrativa, questa fonte non può essere considerta un registro statistico. Un registro statistico svolge tipicamente la funzione di strumento di coordinamento dei dati, integrando dati provenienti da fonti diverse, statistici e amministrativi. Inoltre, questi possono fornire la base per collegare dati provenienti da diverse fonti nel corso del tempo, consentendo lo sviluppo di analisi longitudinali. I registri possono essere costruiti collegando i *record* attraverso l'uso di identificatori comuni, oppure usando le tecniche di *matching* sopra descritte.
Vi sono molti modi in cui i registri amministrativi possono essere utilizzati per la produzione statististica, ma poichè le fonti amministrative disponibili variano da un paese all'altro, non sono stati definiti degli *standard* internazionali. Di seguito si mostrano degli esempi, di come alcuni ricercatori hanno utilizzato i dati amministrativi per la costruzione e l'aggiornamento dei registri statistici.

I registri come combinazione di più fonti La figura 2.1 rappresenta il modello semplificato delle fonti utilizzate per la costruzione e l'aggiornamento del registro delle imprese nel regno unito. Questo modello pone al centro il registro statistico, come strumento per unire e riconciliare i dati provenienti da fonti statistiche (indagini) e da tutte le altre fonti disponibili (informazioni su imprese, i dati provenienti dal sistema di informazione geografica e i dati provenienti da registri satellite).

Registri centralizzati I registri centralizzati sono solitamente utilizzati per migliorare l'efficienza di governo, in quanto forniscono una singola interfaccia per le agenzie governative riducendo le duplicazioni, e quindi l'onere

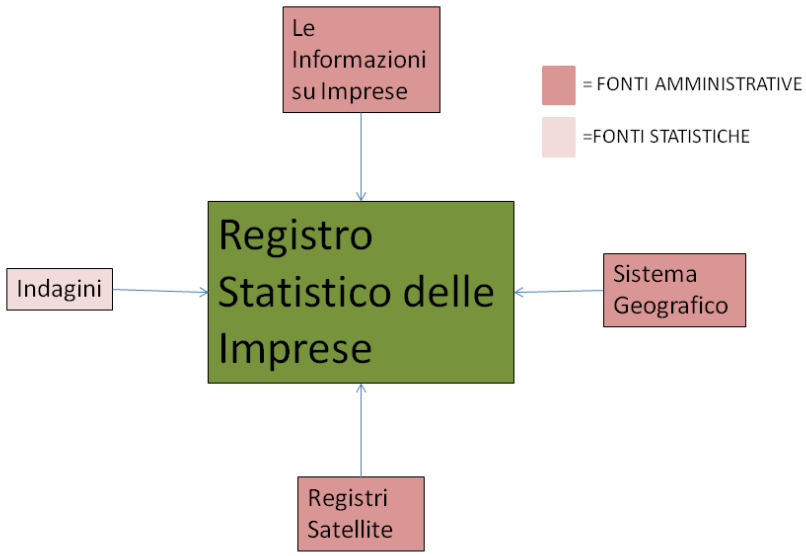


Figura 2.1: Il modello Inglese.

delle procedure amministrative. Ad esempio, una persona o una impresa, che cambiano il loro indirizzo, devono comunicare una sola volta la nuova informazione, e questa poi sarà condivisa tra tutte le agenzie competenti. Questo tipo di registro amministrativo produce un immenso beneficio alla produzione statistica, in quanto rimuove gran parte della procedura di matching e riconciliazione dei dati. Per massimizzare i benifici, sarebbe necessario che gli INS partecipassero allo sviluppo e alla gestione dei registri amministrativi, per garantirsi che le informazioni raccolte rispondano

alle loro esigenze informative. Un buon esempio di dove questo approccio ha funzionato è il registro delle imprese australiano, che è stato sviluppato dall'agenzia delle entrate per l'amministrazione australiana, ma ha stretta collaborazione con l'ufficio statistico australiano. Infatti lì molte delle statistiche sulle imprese vengono prodotte semplicemente con i dati contenuti nel registro delle imprese, e si ricorre alle indagini solo per lo studio di realtà più complesse, come le grandi imprese.

Condivisione dei dati *hub* Una variante del modello del registro centralizzato amministrativo, è il concetto di una condivisione dei dati *hub*. In questo modello l'ente centrale non è un modello vero e proprio, ma è uno strumento che ricerca la congruenza tra i dati detenuti da diverse agenzie. Il suo scopo principale è quello di fornire una porta attraverso la quale i dati provenienti da diverse organizzazioni possono essere condivisi all'interno del settore delle amministrazioni pubbliche. Questo approccio non è stato ancora implementato ma resta un'opzione valida per la condivisione dei dati amministrativi.

Registri satellite Un'approccio piuttosto diverso dai precedenti, è quello di organizzare i dati amministrativi in specifiche fonti legate ad un registro statistico. Se queste specifiche fonti soddisfano determinati criteri possono essere definite 'registri satellite'. I registri satellite sono dei registri disponibili per l'intero sistema statistico, che contengono informazioni sulle unità e le variabili d'interesse, e soddisfano le seguenti condizioni:

- non sono parte integrante di un registro statistico, ma sono collegate ad esso;

- contengono meno informazioni del registro statistico, ma nel loro contesto informativo possono avere una copertura più estesa di unità o di variabili;
- contengono una o più variabili che non sono contenute nel registro statistico.

I registri satellite sono quindi, strumenti per l'integrazione dei dati amministrativi che risultano essere rilevanti solo per un sotto-insieme di unità di un registro statistico. Essi possono essere costruiti utilizzando fonti amministrative, indagini statistiche o una combinazione di entrambi. Per garantire che i registri satellite siano sufficientemente coerenti con i registri statistici, pu'o essere utile prendere in considerazione ulteriori criteri, ad esempio identificatori di unità, definizioni e classificazioni comuni.

2.3.3 Qualità dei dati amministrativi

Le preoccupazioni per la qualità dei dati amministrativi sono spesso uno dei principali ostacoli al loro maggiore utilizzo a fini statistici. Gli INS possono utilizzare i dati amministrativi per vari scopi, quindi per valutare la qualità di questa fonte è necessario stabilire prima la sua destinazione d'uso. Per esempio, una fonte amministrativa pu'o essere considerata inappropriata a fornire dati sulla variabile principale del fenomeno oggetto di studio, ma la stessa fonte potrebbe essere utile a fornire informazioni ausiliarie. Quindi per valutare la qualità dei dati amministrativi, è necessario porsi nell'ottica che i dati amministrativi sono i fattori d'ingresso in un processo di produzione, statistica.
Per valutare la qualità dei dati amministrativi in ingresso, bisogna considerare due domini di qualità: i metadati e il dominio dei dati. Per la

valutazione dei metadati, in olanda, sono stati sviluppati 31 indicatori di qualità, racchiusi in una lista, definita di controllo [17]. Invece, per la valutazione della qualità del dominio dei dati è anzitutto necessario definire le dimensioni che esprimono la qualità di un dato. È stato condotto uno studio da Daas e Osten su questo tema, e sono arrivati a definire quattro dimensioni di qualità: accuratezza, completezza, coerenza e dimensione temporale. Per accurattezza si intende la misura in cui i dati sono corretti, affidabili e certificati. Possono compromettere la validità di questo principio, chiavi di identificazione errate, records inaffidabili, procedura di registrazione errata, etc. Per completezza si intende il grado in cui una fonte comprende i dati che descrivono il fenomeno oggetto di studio. Possono causare incompletezza di una fonte i dati mancanti, ridondanza di records e l'overcoverage (la presenza di unità non appartenenti alla popolazione bersaglio). Inoltre, per coerenza si intende il grado di corrispondenza tra i dati e le unità contenuti nella fonte. L'ultima dimensione della qualità di una fonte è rappresentata dalla dimensione temporale. Cioè bisogna verificare se c'è un davario troppo ampio tra il periodo di raccolta dei dati e la consegna della fonte, perch'e questo potrebbe rendere le statistiche 'vecchie' e quindi non in grado di catturare i cambiamenti della realtà.

2.3.4 Innovazioni organizzative

L'uso sistematico di informazioni amministrative ha modificato in modo considerevole l'organizzazione dei sistemi statistici nazionali, in quanto è nata l'esigenza di gestire il rapporto con gli enti produttori e i flussi informativi. Per soddisfare l'esigenza suddetta, all'interno degli istituti nazionali di statistica, è stata istituita una specifica direzione centrale, la quale rappresenta il punto di riferimento di tutte le attività connesse agli archi-

vi amministrativi. Questo ufficio raccorda le diverse unità di produzione, garantendo l'adozione di standard comuni di trattamento dei dati amministrativi e un loro uso condiviso nel rispetto delle disposizioni in materia di privacy. La razionalizzazione dei processi di acquisizione, trattamento e utilizzazione degli archivi amministrativi ha inoltre generato l'esigenza di un'organizzazione ad-hoc della meta-informazione. La meta-informazione è di fondamentale importanza, in quanto fornisce agli utenti la conoscenza sulla portata informativa delle fonti amministrative e sulle caratteristiche dei processi adottati per la loro validazione statistica.

2.4 Miscela di fonti

Per rendere più efficiente il processo di produzione statistica, è possibile utilizzare una miscela di fonti, cioè utilizzando sia dati provenienti da fonti primarie che dati provenienti da fonti secondarie. Questo tipo di approccio (misto) consente di ridurre al minimo le problematiche connesse all'utilizzo di una o dell'altra fonte, sfruttando i dati di buona qualità già esistenti (dati amministrativi) ed integrandoli con i dati raccolti tramite indagini (Campionaria o censuaria). Per questo tipo di procedura sono stati sviluppati metodi *ad-hoc*, conosciuti in letteratura come metodi misti.
Il primo metodo noto come '*split population approach*', suddivide, in base allo scopo della ricerca, la popolazione in due o più parti. I dati amministrativi vengono utilizzati per le unità in cui tali dati sono di sufficiente qualità, mentre per le restanti unità i dati sono raccolti tramite indagine. Questo è il tipico scenario delle statistiche sulle imprese, le informazione delle imprese che presentano una struttura semplice sono ottenute dalle loro dichiarazioni fiscali, mentre sono utilizzate le indagini per raccogliere le

informazioni delle imprese più grandi. Questo approccio trova la sua giustificazione nell'importanza di alcune unità rispetto ad altre, infatti sono proprio le imprese più grandi ad avere un maggiore impatto sulla qualità dei dati, e quindi sono quelle per le quali è più importante disporre di dati accurati.
Il secondo metodo noto come '*split data approach*', al contrario del primo, divide i dati e non la popolazione. Ad esempio, la popolazione considerata potrebbe essere composta da tutte le persono che vivono in un determinato paese, e la richiesta di dati potrebbe essere il solito set di variabili necessarie per un censimento della popolazione. I dati amministrativi sono utilizzati per fornire alcune delle variabili, ma non per tutta la popolazione. Questo tipo di approccio, quindi, non riduce il numero dei questionari ma riduce il volume dei dati richiesti. Questo metodo viene solitamente utilizzato per la raccolta di dati complessi, i censimenti ne rappresentano il principale esempio.

Capitolo 3

L'uso di nuove risorse: metodologie e pratiche esistenti

3.1 Le nuove fonti di dati per la statistica ufficiale

La crisi economica degli ultimi anni ha generato un nuovo modo di pensare *'up-to-date'*, cioè essere in grado di reagire tempestivamente ed in modo adeguato ai cambiamenti e agli eventi inattesi. Il contesto descritto ha sottolineato le carenze del sistema statistico esistente, ponendo quest'ultimo di fronte a nuove sfide, per le quali non è facile trovare una soluzione univoca che soddisfi contemporaneamente tutti i soggetti interessati. Gli utenti: le imprese, le persone fisiche e soprattutto le istituzioni governative richiedono dati sempre più dettagliati, tempestivi e di buona qualità. Contestualmente i fornitori di dati, in particolare le imprese, richiedono di diminuire il carico amministrativo e statistico gravante su di loro. L'incessante sviluppo tecnologico e la crescente diffusione di dispositivi collegabili

alla rete Internet sta creando una nuova miniera informativa utile per la produzione di informazioni. Le nuove tecnologie di comunicazione offrono opportunità di raccolta di dati semplificate che dovrebbero ridurre l'onere per le imprese e migliorare la qualità delle informazioni statistiche. La creazione della raccolta elettronica dei dati, i sistemi di elaborazione, l'uso dei dati forniti dalle imprese nei loro bilanci annuali, i dati dei social network e la creazione di tassonomie XBRL integrate rappresentano le *'new sources'* utilizzabili al fine di semplificare il trasferimento dei dati dalle aziende agli istituti nazionali di statistica e per rendere i prodotti statistici più efficienti. L'uso di queste fonti rappresenta una grande opportunità per gli INS ancora non sufficientemente sfruttata a causa delle problematiche connesse alla raccolta dei dati.

3.2 Le esperienze europee

Lo scopo di questo capitolo è quello di fornire alcuni esempi di strumenti più sofisticati che gradualmente penetrano nella raccolta di dati statistici. Le ragioni principali sono la riduzione per gli oneri degli intervistati, la riduzione dei costi e il miglioramento della qualità delle statistiche ufficiali. Inoltre, attraverso procedure di*data mining* e di aggregazione (ad esempio aggregando i dati grezzi delle imprese in dati utili per le statistiche ufficiali) si generano anche economie nella fase di raccolta dei dati [18]. Nel campo della raccolta dei dati provenienti da altre fonti il *mining* di dati testuali (noto in letteratura come *text mining*) viene utilizzato al fine di rivelare dati e informazioni utili. Questo è uno dei principali temi affrontati nel progetto europeo BLUE-ETS [19]. In questo campo molti altri modi, come *web mining* e dispositivi mobili di posizionamento stanno lentamente

penetrando nelle statistiche ufficiali.

3.2.1 *Open-ended questions*

Nella metodologia di indagine, l'uso di molte domande a risposta chiusa e un paio di domande a risposta aperta è molto frequente. Il motivo principale è dovuto alle difficoltà di analisi di queste ultime. Recentemente l'informazione potenziale che questo tipo di domande fornisce ha guidato l'interesse della ricerca nel campo dell'analisi dei dati testuali (ADT). In letteratura esistono numerosi sviluppi dei metodi esplorativi per l'analisi dei dati testuali [20]. L'idea è analizzare una particolare tabella (tabella lessicale) ottenuta attraverso operazioni di pre-elaborazione del testo. In questa sezione si illustrano alcune delle esperienze sviluppate dall'Istituto statistico italiano nel settore delle indagini basate su domande aperte.

Introduzione Nel 2012, l'Istat ed il CNEL (Consiglio Nazionale Economia e Lavoro) hanno sviluppato un'iniziativa congiunta per misurare il benessere della società italiana. L'iniziativa Cnel-Istat pone l'Italia nel gruppo dei paesi (Francia, Germania, Regno Unito, Stati Uniti, Australia, Irlanda, Messico, Svizzera, Olanda) che hanno recentemente deciso di misurare il benessere e la qualità della vita della società. La premessa di fondo dell'approccio seguito in Italia è che il concetto di benessere equo e sostenibile (Bes) è strettamente legato a tempi, luoghi e culture e quindi non può essere definito univocamente, ma solo attraverso una strategia condivisa che coinvolga la società civile nella definizione delle dimensioni che costituiscono i fondamenti del benessere. Pertanto, sono stati individuate, principalmente, 12 dimensioni del benessere: ambiente, salute, benessere economico, istruzione, formazione, lavoro e conciliazione dei tempi di vi-

ta, relazioni sociali, sicurezza, benessere soggettivo, paesaggio e patrimonio culturale, ricerca e innovazione, qualità dei servizi, politica e istituzioni [21].

La raccolta dei dati Sulle 12 dimensioni individuate (o domini) si è aperta una fase di consultazione pubblica per esperti e singoli cittadini, costituita da un questionario on line, attraverso cui è possibile fornire le proprie opinioni sull'utilità di misurare il benessere e sulla scelta delle dimensioni che lo determinano, lasciando la possibilità ai rispondenti di segnalare dimensioni aggiuntive. Le risposte aperte sono state analizzate con tecniche di analisi testuale dei dati. Attraverso l'analisi di alcune parole o combinazioni di queste (segmenti ripetuti e linguaggio peculiare) e le loro frequenze, si è ottenuta una rappresentazione sintetica delle tematiche suggerite dai rispondenti.

I risultati Dalle analisi testuali effettuate, sono emerse per l'Italia, in termini di qualità della vita elementi peculiari che la differenziano nelle risposte ottenute, infatti, con frequenza piuttosto elevata, compaiono termini che rimandano all'importanza del patrimonio artistico, storico, culturale e paesaggistico. Questo caratterizza in positivo il paese, insieme alla qualità del cibo, alle miti condizioni climatiche, alla qualità delle relazioni sociali e alla disponibilità di un sistema di welfare universale. I termini come corruzione, burocrazia, privilegi, disorganizzazione, individualismo, degrado rimandano alla segnalazione di quello che invece ci caratterizza in negativo rispetto agli altri paesi, come la sfiducia nelle istituzioni, lo scarso investimento in ricerca e istruzione, la mancata tutela dei giovani e l'inadeguato

sviluppo di una società multiculturale. Inoltre sono emerse interessanti differenze di genere: le donne parlano di equità, giustizia sociale, benessere psicologico, al contrario, gli uomini, parlano di politica, salute fisica e soprattutto di benessere economico.

La classificazione delle professioni: l'esperienza Italiana

Introduzione L'indagine sulle professioni, svolta nel biennio 2006-2007, è finalizzata alla costruzione di un sistema descrittivo delle professioni presenti nel mercato del lavoro italiano. Nel corso dell'indagine, condotta dall'Istat, dal Ministero del lavoro e dall'Isfol, sono state adottate delle strategie di trattamento delle risposte aperte per individuare e codificare le specifiche attività di lavoro. In particolare, il testo è stato sottoposto ad alcune tecniche di analisi testuale per standardizzare le risposte e produrre una lista delle attività dettagliate. Tuttavia, per ottenere un elenco di attività standard per ciascuna unità professionale è stato necessario utilizzare le tecniche di analisi testuale in modo innovativo, riconducendo le risposte ad un sistema di categorie definito. A tal proposito è stato adoperato il *software* Taltac, considerando, come unità di analisi, l'intero frammento di testo, all'interno del quale è possibile ricercare alcune combinazioni di parole cui associare un'etichetta. La categorizzazione ottenuta in automatico è stata successivamente validata mediante un'analisi puntuale del risultato, finalizzata sia a completare il lavoro per i frammenti non classificati, sia a controllare l'aderenza dell'etichetta apposta in automatico al testo originario e al contesto delle unità professionali [22].

La raccolta dei dati L'indagine ha riguardato circa 16000 lavoratori, rappresentativi delle 800 unità professionali (UP) oggetto dell'indagine. Nel questionario sono state richieste le principali attività svolte nell'ambito della propria professione. Le informazioni testuali sono state trattate con l'obiettivo di produrre una lista delle attività dettagliate per ciascuna UP, standardizzando le risposte registrate dai rilevatori, ma conservando per quanto possibile la specificità e la variabilità del testo originario. Inizialmente, è stata effettuata l'analisi delle corrispondenze sulla distribuzione del vocabolario per Grandi Gruppi professionali, mostrando, dal punto di vista del linguaggio usato per descrivere il lavoro, la sostanziale omogeneità interna dei raggruppamenti professionali. In seguito, è stata utilizzata la funzione di 'Ricerca entità', presente nel *software* Taltac 2.0 che, inserendo una nuova variabile nel file di partenza, ha permesso di ricercare ed etichettare nell'intero frammento di testo gli insiemi di parole necessari a definire le attività. La ricerca sul testo delle informazioni utili per alimentare questa nuova variabile avviene ricorrendo alla costruzione di *query* complesse che consentono di individuare le diverse espressioni che possono essere ricondotte ad un unico concetto. Attraverso l'analisi delle concordanze delle parole chiavi del testo, è stata effettuata un'analisi del contenuto, sia per ottenere un sistema di categorie, sia per individuare ulteriori attività da ricercare ed etichettare. Successivamente, si è provveduto alla ricostruzione del testo con l'aggiunta della nuova variabile che riporta, quindi, le etichette definite in fase di categorizzazione.

I risultati Per assicurare la possibilità di confronto e individuare attività trasversali, è stata realizzata una fase di validazione manuale in modo

da controllare l'uniformità delle etichette nel caso di attività comuni a più di una UP. L'inevitabile variabilità introdotta dai codificatori ha dimostrato l'utilità di un approccio semi-automatico, che consente di ridurre o almeno di controllare la variabilità delle decisioni prese in fase di codifica. Nonostante la fase di validazione manuale del risultato sia risultata piuttosto onerosa, la possibilità di categorizzare in automatico una parte dei frammenti costituisce un risultato molto interessante, in quanto ha ridotto i tempi di esecuzione del lavoro. Inoltre ha garantito un buon livello di standardizzazione del risultato, ha permesso di individuare le attività trasversali a diverse Up e ha costruito un criterio guida per la fase di codifica manuale. Infine, la disponibilità del dizionario delle attività, oltre a facilitare e velocizzare il lavoro, ha limitato la soggettività della codifica da parte di differenti codificatori.

L'uso del tempo: l'esperienza Italiana

Introduzione L'indagine Multiscopo 'Uso del tempo', svolta dall'ISTAT nel biennio 2002-03 [59], descrive le attività abituali delle famiglie italiane. Il principale strumento di rilevazione utilizzato in questa indagine è il diario giornaliero, che costituisce una fonte ricca e dettagliata di informazioni sull'organizzazione dei tempi individuali espresse in linguaggio naturale. La presenza di informazioni testuali necessita di una complessa fase di codifica; infatti, il linguaggio dei codici è un linguaggio rigido, non in grado di gestire le ambiguità o fornire interpretazioni. Le esperienze negative della codifica manuale e del *software* di codifica automatica, hanno permesso l'individuazione di strategie alternative ed innovative rispetto al

passato; con l'utilizzo del *software* Blaise si è sviluppato un sistema di codifica assistita, così da ridurre la discrezionalità dei codificatori ed il rischio di commettere errori. In questo modo si sono contenuti i tempi di realizzazione del lavoro ed è stato possibile un efficace monitoraggio in corso di raccolta, ottenendo un notevole miglioramento della qualità dei dati. Tale strategia, pur fornendo uno strumento elettronico che facilita, velocizza e supporta l'attività umana, lascia comunque al centro dell'attività la capacità critica del codificatore [23].

La raccolta dei dati Attraverso la compilazione del diario è possibile conoscere il modo in cui ciascun rispondente ripartisce la propria giornata (considerando intervalli di 10 minuti) tra le varie attività, gli spostamenti, i luoghi frequentati e le persone con cui ha trascorso del tempo. Gli strumenti di rilevazione utilizzati nell'indagine sono quattro: un questionario per intervista individuale, un questionario per intervista familiare, un diario settimanale e il diario giornaliero, destinato specificamente a rilevare l'uso del tempo quotidiano. La codifica delle informazioni testuali, avvenuta mediante il *software* Blaise, considera due tipologie di ricerca, l'albero dei codici e il dizionario elettronico, utilizzati singolarmente e/o congiuntamente. Il dizionario da fornire al *software* per effettuare la ricerca del codice corretto si costruisce con la registrazione delle informazioni testuali (attività principale, attività contemporanea) dei diari ed è composto da un archivio di frasi-attività. Operativamente, è stata progettata una maschera per la codifica che ha consentito al codificatore di avere a disposizione tutte le informazioni di contesto necessarie allo svolgimento corretto del lavoro, ovvero le caratteristiche socio-anagrafiche sulla persona (la professione,

l'età, il sesso ecc.), alcune informazioni sulle caratteristiche della giornata (data, giorno della settimana), la composizione della famiglia, nonché la sequenza delle attività da registrare. Dal dizionario si evince il codice in base alla similitudine (a gruppi di tre lettere) della frase registrata con quelle presenti nel vocabolario appositamente costruito. Se la frase è presente ed univocamente associata ad un codice, il codificatore passa all'attività successiva, se invece il *software* suggerisce più alternative, il codificatore deve interpretare il contesto per selezionare il codice corretto. Nel caso in cui la frase registrata non trovasse una corrispondenza plausibile con nessun'altra presente nel dizionario, il codificatore avvia la ricerca sull'albero dei codici, strutturato secondo la tradizionale forma gerarchica. Grazie alle numerose informazioni testuali derivanti dagli archivi, è possibile un'analisi testuale delle molteplici espressioni che descrivono le condizioni di contesto o uno stato d'animo, per identificare, poi, specifici gruppi sociali.

I risultati Un importante vantaggio di questa indagine è dovuto alla disponibilità dell'intero archivio testuale dei dati che consente di effettuare analisi innovative (l'applicazione di tecniche di *text mining* e di approfonditi studi qualitativi), ma soprattutto permette di individuare le principali tipologie del *gap* comunicativo tra ricercatori e rispondenti, analizzato sul piano quantitativo. Il *gap* è dovuto principalmente al fatto che i rispondenti esprimono spesso quello che pensano ci si aspetti da loro, facendone una sintesi personale, seguendo la loro percezione e utilizzando il linguaggio che ritengono più appropriato. Si cerca, quindi, di ridurre quanto più possibile il *gap* comunicativo tra intervistati e intervistatori utilizzando un nuovo strumento di raccolta dei dati. Per quanto concerne l'utilizzo della codifica

assistita, i principali vantaggi che ne sono derivati riguardano la riduzione del numero di codificatori, caratterizzati da un processo di formazione più efficace; la supervisione nella codifica di attività di difficile interpretazione; l'individuazione tempestiva di errori sistematici e la conseguente adozione dei correttivi più appropriati. Uno svantaggio del *software* Blaise è che non consente di calcolare in automatico le statistiche sulle diverse strategie di codifica adottate dagli operatori, ovvero non si riesce a stabilire in quali casi il processo di codifica è avvenuto tramite il ricorso al dizionario e in quali tramite l'albero gerarchico della classificazione.

3.2.2 *World Wide Web*

Il continuo sviluppo della tecnologia e la crescente popolarità di dispositivi collegabili a *internet* crea una nuova miniera informativa utile per la produzione di informazioni. Gli istituti nazionali di statistica stanno gradualmente rivalutando le tradizionali fonti di raccolta dei dati, a favore di altre fonti più efficaci. Questa pratica, anche se condivisa dal sistema statistico europeo, è operativamente ancora in fase di sperimentazione. Di seguito mostriamo l'esperienza internazionale in questo settore.

L'indice dei prezzi al consumo: l'esperienza Olandese

Introduzione Ogni giorno su internet vengono effettuate milioni di transazioni per l'acquisto dei prodotti più diversi, biglietti aerei, pernottamenti in hotel, alimenti, abbigliamento etc. Le transazioni avvenute con successo registrano dati utili all'estrazione di informazioni sui prezzi. L'i-

stituto nazionale statistico dei Paesi Bassi utilizza questa fonte per la determinazione dell'indice dei prezzi al consumo (IPC) di biglietti aerei, libri, CD e DVD [24].

La raccolta dei dati Esistono diverse modalità di raccolta automatica da pagine web. Per la determinazione dell'IPC l'istituto statistico Olandese ha utilizzato un linguaggio di *scripting*, come phyton e perl, ed il *software Djuggler* (commerciale). La scelta di questi strumenti è stata determinata da un'accurata analisi costi/benefici derivanti dal loro utilizzo. La principale criticità dei linguaggi di *scripting* è la loro inabilità a raccogliere dati incorporati su pagine *web* dinamiche, con immagini e/o animazioni (ad esempio *'flash'*). Questo limite seppur rilevante non ha condizionato la procedura di raccolta dati che al contrario sarebbe stata compromessa, ad esempio con l'uso dei motori di ricerca (essenzialmente progettati per indicizzare, piuttosto che raccogliere). I dati sono stati raccolti automaticamente (in notturno, per ridurre il carico sul sito *web*) da quattro siti di compagnie aeree per un periodo di dieci mesi.

I risultati Nel periodo di raccolta automatica dei dati si sono riscontrate alcune problematiche che hanno condizionato tale fase: il tempo di risposta del sito *web* (ritardi nelle richieste) e le modifiche occasionali dei siti *web*. Il primo problema deve essere risolto in fase di progettazione dello *script*, invece le problematiche connesse al secondo evento, essendo imprevedibili, sono state risolte in corso di raccolta. Nello specifico tre dei quattro siti utilizzati sono stati modificati durante il periodo di raccolta automatica, quindi la procedura si è interrotta nelle ore impiegate per la

riprogrammazione del nuovo *script* (i tempi dipendono dalla complessità del sito *web* e dalle modifiche apportate al sito). Lo *script* sviluppato per la compagnia aerea che non ha modificato il proprio sito durante il periodo di raccolta ha funzionato regolarmente. La procedura proposta, nonostante abbia presentato i limiti sopra esposti, è stata utile alla determinazione del IPC e si considera una grande opportunità per la determinazioni dei prezzi di prodotti che si svilupperanno in nuovi mercati.

I messaggi pubblicati su *twitter*: l'esperienza Olandese

Introduzione Un numero di persone sempre maggiore pubblica quotidianamente le loro opinioni, i loro stati d'animo e le loro emozioni sui social network. Questa abitudine, diventata ormai un *life style*, da un punto di vista statistico rappresenta una nuova fonte di dati utili a misurare fenomeni sociali. L'istituto statistico Olandese ha condotto uno studio per l'individuazione dei temi trattati su *twitter* [25]. *Twitter* è un servizio di micro *blogging* che consente l'invio di brevi messaggi di testo, di 140 caratteri di lunghezza, chiamati '*tweets*'. I *tweets* pubblicati in bacheca sono consultabili da tutti gli utenti del sito. Sono utenti le persone che hanno un *account twitter*, univocamente identificato da un nome utente, una *password* e alcune informazioni di carattere generale (nazionalità, età, ect). La peculiarità dei messaggi inviati su *twitter* è la possibilità di etichettare con alcune parole chiavi il contenuto del messaggio. Operativamente, ponendo un ashtag d'avanti alle parole chiavi è possibile la categorizzazione dei messaggi.

La raccolta dei dati Per la raccolta dei dati,inizialmente, si è fatto uso delle informazioni di localizzazione considerando solo gli utenti di nazionalità olandese, mediante un controllo manuale, ad intervalli regolari. Gli utenti di altre nazionalità sono stati rimossi. Per gli utenti identificati, sono stati raccolti e conservati in un data base i loro messaggi più recenti (fino ad un massimo di 200 *tweets* per ogni utente). Successivamente, i *tweets* raccolti sono stati analizzati con i metodi tradizionali di *text mining*.

I risultati Da un'analisi iniziale del contenuto dei *tweets*, è apparso con evidenza che questo strumento viene utilizzato per discutere su una grande varietà di argomenti, per cui ci si è concentrati solo sui messaggi contenenti *hashtag*, dove gli utenti indicano le parole chiave del loro messaggio. L'analisi è stata condotta con il *software* LingPipe con un'implementazione del DynamicLMClassifier. I risultati rivelano che molti messaggi raccolti sul *social network*, potranno essere di potenziale interesse per le statistiche, in particolare, la politica e gli eventi potrebbero essere utilizzati per ottenere informazioni sulle opinioni, atteggiamenti e sentimenti nei Paesi Bassi. Anche se i primi risultati sembrano positivi, è molto difficile considerare che questi siano condivisi dalla popolazione olandese nel suo complesso, sia perché non tutti i cittadini olandesi sono attivi su *Twitter*, sia perché l'analisi ha incluso solo i messaggi contenenti *hashtag* (funzione non utilizzata da tutti gli utenti).

3.2.3 Dispositivi di posizionamento

La posizione dei telefoni cellulari da la possibilità di identificare la posizione degli utenti attraverso il loro utilizzo attivo. I proprietari dei dati sono le società di telecomunicazioni. Uno studio su alcuni dei dati ha rivelato alcuni problemi di rappresentatività del set di dati (dati anonimi), ma ha identificato le attività di mobilità (turismo), come aree potenziali nelle quali poter utilizzare questi dati.

Il caso Estone

Introduzione Nel settore delle statistiche delle imprese turistiche l'estrazione di dati dal posizionamento dei telefoni cellulare, mostra i suoi vantaggi nei confronti di indagini tradizionali [26]. Con l'apertura delle frontiere interne dell'UE e il nuovo modo di viaggiare (senza il supporto di un'agenzia di viaggio) la fase di raccolta dei dati sul turismo è diventata più complessa. In questa zona il movimento dei telefoni cellulari è una delle più facili fonti utili a registrare l'attraversamento delle frontiere e dei flussi di traffico. La figura 3.1 mostra il confronto tra i dati raccolti dagli agenti di viaggio e i dati raccolti dal posizionamento mobile per i viaggi, all'interno dell'Unione europea e al di fuori dell'UE. L'estonia utilizza questi dati per l'analisi del traffico, del pendolarismo (chiamate nazionali) e per le statistiche sul turismo (chiamate in *roaming*).

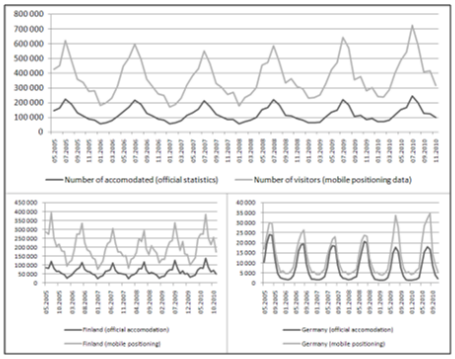
La raccolta dei dati In Estonia, LBS Positium e il dipartimento di geografia dell'università di Taru forniscono alle oganizzazioni governative diversi set di dati provenienti dalle reti di telefonia mobile. Il data base

contiene tutte le attività di chiamate effettuate dai telefono cellulari. Ogni telefono è descritto dal suo ID, contenente uno pseudonimo e il paese di origine (questi dati sono forniti in anonimato). Come precedentemente anticipato, questi dati sono raccolti per le statistiche del turismo, studi di traffico e studi pendolarismo, che rappresentano un'informazione preziosa, ad esempio, per la pianificazione dello sviluppo delle infrastrutture stradali.

I risultati Le statistiche sul turismo prodotte con l'utilizzo dei dati di posizionamento mobile differiscono notevolmente da quelle prodotte da statistiche ufficiali. I principali motivi devono ricercarsi nelle diverse metodologie di analisi utilizzate, e nella popolazione *target* identificata. Le statistiche ufficiali, rispetto ai dati di posizionamento mobile, includono nella raccolta dei dati anche i turisti giornalieri, i visitatori di transito e i visitatori che pernottano in albergi non ufficiali. Quindi questo divario, genera risultati simili dei due approcci per i visitatori che soggiornano in hotel per più notti, ma risultati molto diversi per i visitatori di paesi limitrofi, che transitano nella città solo per una giornata (vedi figura 3.1).Questo risultato, pone la seguente domanda: sono più precise e pertinenti, le statistiche ufficiali sugli alloggi o le statistiche prodotte con i dati di posizionamento mobile?

Il caso Olandese

Introduzione Attualmente il 92 per cento del popolo olandese possiede un cellulare. Le aree di copertura cellulare sono suddivise in celle nelle quali viene registrata, con un codice identificativo, ogni apparecchiatura cellulare che invia o riceve, una telefonata o un messaggio, nella zona in



Fonte: Mobile Telephones and Mobile Positioning Data come fonte per la produzione di statistiche: Esperienza Estone (Ahas et al, 2011)

Figura 3.1: I dati degli agenti di viaggio e di posizionamento mobile.

cui è localizzata la cella di copertura. I codici registrati identificano dove è localizzato il telefono cellulare, e quindi di conseguenza il suo utilizzatore. Questa fonte di dati, oltre a rappresentare per i gestori di telefonia mobile, la loro fonte informativa primaria, da un punto di vista statistico consente agli INS di stimare la mobilità, il turismo e l'attività economica.

La raccolta dei dati Lo studio dell'istituto di statistica olandese è stato condotto su dati ricevuti da un'unica società di telecomunicazioni, che gestisce un terzo delle comunicazioni di telefonia mobile dei paesi bassi. Il set di dati ottenuto era composto da: record di ogni chiamata-evento (chiamate e messaggi), informazioni sulla data e ora di inizio della chiamata-evento, il tipo di evento (chiamata ricevuta, inviata, ect) e una chiave di identificazione anonima univoca per ogni apparecchiatura cellulare. Inoltre è stato ottenuto un piano contenete informazioni relative alle geo-posizioni delle antenne utilizzate dalla società di telecomunicazioni. In proposito, questo piano cellulare è stato trasformato in zone non sovrapposte [24].

I risultati Lo studio condotto sui dati provenienti dal traffico cellulare è nato con il principale obiettivo di verificare se e come si modifica l'attività geo-spaziale di un utente di telefonia mobile durante il giorno. Per questo studio sono state fatte delle assunzioni, sintetizzate di seguito:

- la posizione e il movimento di un telefono è strettamente legata al comportamento dell'utilizzatore;
- l'apparecchiatura cellulare si collega sempre alla cella disponibile più vicina;
- le aeree servite dalle cellule non si sovrappongono;
- ad ogni ora registrata puó corrispondere un'unica chiamata-evento (invece nella realtà un telefono cellulare puó contemporaneamente ricevere una chiamata ed un messaggio).

Lo studio ha rivelato che le zone ad alta intensità di chiamata coincidono con le aree ad alta densità di popolazione, come ad esempio la città di Amsterdam e di Rotterdam. Inoltre si è evinto che l'attività di chiamata segue un modello giornaliero. Di norma l'attività di notte è bassa, aumenta rapidamente di mattina, con due picchi nell'ora di pranzo e all'uscita dall'ufficio. In riferimento ai giorni della settimana, questo modello di attività di chiamata caratterizza i giorni feriali: invece durante il *weekend* il maggior flusso di attività si registra durante la mattina. L'analisi dei giorni è risultata utile al fine di identificare aree commerciali/industriali, dove il traffico cellulare nei *weekend* si comporta in maniera opposta al modello generale.

3.2.4 Strumenti automatici

Nel contesto moderno delle statistiche ufficiali, al fine di ridurre l'onere di risposta delle imprese e contestualmente ridurre i tempi e i costi connessi alla raccolta di informazioni, in ambito europeo, sono stati sviluppati strumenti automatici di raccolta dati. Con questi strumenti i rispondenti forniscono i loro dati direttamente dai propri uffici, garantendo agli istituti di raccolta il loro tempestivo ricevimento.

Statistiche sugli alloggi: il caso Finlandese

Introduzione Un'esperienza significativa proviene dall'istituto statistico Finlandese che, mediante un sistema di raccolta automatica dei dati, ha prodotto in maniera più efficiente le statistiche sugli alloggi. Lo strumento di raccolta automatica dei dati nell'istituto finlandese viene utilizzato in

molte altre aree di indagine per la cattura istantanea delle informazioni statistiche che, attraverso una trasmissione crittografica elettronica, vengono inviate direttamente alla banca dati dell'istituto.

La raccolta dei dati L'istituto statistico Finlandese, dal 1971, produce statistiche sugli alloggi riguardanti la capacità di occupazione (arrivi, pernottamenti) degli esercizi ricettivi turistici. In passato, si utilizzavano questionari cartacei e fax per inviare i report del proprio sistema alberghiero all'istituto, e quest'ultimo registrava manualmente i dati. È dal 2005 che l'istituto nazionale Olandese utilizza, per la produzione delle statistiche sugli alloggi, un sistema di raccolta dati automatico [**?**]. È stato creato un questionario Internet basato sul linguaggio XML (Extensible Markup Language), ossia un linguaggio di markup che definisce un insieme di regole di codifica per i documenti in un formato leggibile. Pertanto, i rispondenti possono trasmettere il file XML direttamente dal sistema di gestione alberghiera alla struttura di raccolta dati semplicemente premendo un pulsante.

I risultati Nel 2011, l'istituto statistico Filandese ha ricevuto circa il 66 per cento dei dati direttamente per via elettronica e il 34 per cento da altri metodi di comunicazione (questionario cartaceo, e-mail, fax, ecc.). L'uso di questo strumento ha migliorato notevolmente la qualità, la tempestività e la comparabilità delle statistiche, ottenendo una significativa riduzione dell'onere statistico gravante sui rispondenti e dell'onere economico gravante sull'istituto di raccolta. Seppur le esperienze di raccolta automatica di dati sono state molto incoraggianti, hanno elevati costi di implementazione.

Quindi è necessario che gli INS continuino a promuovere ed incentivare le piccole e medie imprese ad investire nei nuovi sistemi informativi.

Le altre esperienze Europee L'istituto statistico Filandese è stato l'istituto pioniere della raccolta automatizzata dei dati. A partire dal 2005, infatti, gli Stati membri dell'Unione europea hanno sperimentato e realizzato sistemi più o meno simili. L'INE in Spagna ha introdotto un sistema automatizzato di raccolta dei dati nel 2008; da allora gli alberghi hanno avuto la possibilità di inviare i dati in modo automatico dai propri sistemi di gestione o attraverso la pagina *web* dell'istituto. Nel 2012, ormai la maggior parte degli stabilimenti utilizza la raccolta automatizzata dei dati ogni mese. I risultati incoraggianti hanno portato a un ulteriore sviluppo anche in altri istituti nazionali di statistica. Eurostat ha avviato un progetto definito 'ESSnet' che, dal 2010, coinvolge otto Stati membri dell'Unione europea (Spagna, Belgio, Bulgaria, Finlandia, Lettonia, Lituania, Polonia e Slovacchia). Gli obiettivi di questo progetto, conclusosi ad ottobre 2012, si sono concentrati nella riduzione dell'onere di risposta, nel miglioramento della tempestività, della comparabilità internazionale e della qualità delle statistiche sulle strutture ricettive. Si è cercato di generare informazioni statistiche automaticamente dai sistemi di gestione, attraverso lo sviluppo di un sistema comune a tutti i paesi membri.

3.3 Discussione

Sebbene le nuove tecnologie forniscano importanti innovazioni per gli INS nella raccolta dei dati *'up-to-date'*, ci sono molti problemi relativi alla qualità di tali dati e alla possibilità di trasformarli in informazioni statistiche

degne di questo attributo. Il primo problema, e forse il più rilevante, è la distorsione di selezione, principalmente derivante da una distorsione del 'campione selezionato' e di un 'pregiudizio osservatore'. Per quanto riguarda il primo tipo di polarizzazione, il divario tecnologico crea un forte divario di selezione tra persone che utilizzano, ad esempio, GPS o *Twitter*, e le persono che non li utilizzano. Inoltre, gli utenti più esperti possono rendere i dati più difficili da raccogliere, utilizzando funzioni che limitano la visibilità dei loro *account*. Considerando ad esempio *Twitter*, il sito consente di rendere invisibile la tua posizione. Inoltre, studi recenti dimostrano che *Twitter* è principalmente uno strumento utilizzato da persone adulte, e questo chiaramente genera un'ulteriore distorsione legata all'età. Così come si possono avere solo i *tweet* scritti da persone interessate al tema in discussione (*trending topic, hashtags*), e quindi è possibile avere un cosiddetto 'pregiudizio osservatore', relativo a tutti i fenomeni che non stimolano la pubblicazione di un *tweet*. Considerazioni analoghe possono essere fatte per tutti i dati raccolti in rete. L'enorme diffusione dei telefoni cellulari possono renderli più interessanti dal punto di vista di copertura, ma i dati che possiamo raccogliere non sono molto ricchi, confrontati con i dati provenienti dalle altre fonti in rete. La distorsione di selezione non è una nuova domanda per la teoria statistica, ma diventa sempre più importante in quanto si rendono disponibili sempre di più nuove modalità di raccolta dei dati. È importante tenere presente i due lati della medaglia: da un lato, nuove modalità di raccolta dei dati sono a disposizione per raccogliere informazioni difficili da raggiungere con gli strumenti tradizionali, ma sorgono problemi relativi alla qualità dei dati, come la generalizzazione e la riservatezza. Tutte queste questioni devono essere considerate come nuove sfide per la teoria statistica.

Capitolo 4

Il mining da basi di dati documentarie

Il progresso della tecnologia, oltre ad aver stravolto l'intero pianeta, ha prodotto una vera e propria rivoluzione nella produzione dei dati statistici. Lo sviluppo dei motori di ricerca, del linguaggio XML, degli algoritmi di intelligenza artificiale, ha prodotto la possibilità di analizzare nuove informazioni (testi, documenti, *chat*, domande a risposta aperta etc.) con tecniche di elaborazione automatica. Se pensiamo alla enorme quantità dei documenti e delle informazioni contenute nel Web, è facile spiegare il grande interesse sviluppato nell'ultimo ventennio verso il cosiddetto *'text mining'*, termine con il quale si identifica l'insieme delle tecniche utili ad analizzare, esplorare e interrogare raccolte di grandi basi di dati testuali. Con queste tecniche sono possibili indagini evolute, che consentono di estrarre informazioni utili a produrre valore (conoscenza). Il *text mining* trae le sue origini dai più antichi studi sul linguaggio, noti oggi con il termine di analisi dei dati testuali (ADT). In questo capitolo si presenta una rassegna delle tecniche di

analisi dei dati testuali e del più recente filone di ricerca del *text mining*, con una particolare attenzione alla procedura di trasformazione del testo in dato statistico, analizzabile con tecniche appropriate.

4.1 L'analisi testuale dei dati

Il linguaggio naturale è lo strumento attraverso il quale l'uomo esprime il suo pensiero e riesce ad interagire con il mondo esterno; il linguaggio è la caratteristica che differenzia l'uomo dalle altre specie viventi. Dagli inizi del XX secolo il linguaggio è stato oggetto di interesse di molteplici ambiti disciplinari. I primi studi quantitativi sono stati sviluppati da linguisti, con l'obiettivo di codificare le regolarità esistenti nella lingua. È a partire dagli anni Sessanta, che gli studi statistici su dati espressi in linguaggio naturale hanno subito notevoli cambiamenti, strettamente legati all'evoluzione dell'informatica, fino a produrre l'analisi automatica dei testi e la statistica testuale [28]. Più di recente, la crescente disponibilità di risorse linguistiche informatizzate [29] e la crescente diffusione di testi consultabili online, quindi direttamente analizzabili, ha ulteriormente rivoluzionato criteri e tecniche. Gli strumenti di analisi non sono più solo di natura statistica, ma scaturiscono da un'intensa multidisciplinarità che associa a questi, strumenti informatici e linguistici. Nel corso del tempo gli studi quantitativi intorno alla lingua hanno cambiato progressivamente il loro obiettivo, spostandosi da una logica di tipo linguistici, ciò concentrando l'attenzione su morfemi, lessemi e n-grammi [30], ad una di tipo lessicale, in cui l'analisi del linguaggio si fonda sullo studio dei lemmi [31], per approdare negli anni Ottanta ad analisi di tipo testuale, basata sull'analisi per forme grafiche (e non più per lemmi) [32] ed infine lessico-testuale. L'approccio lessico-testuale in-

tegra un'analisi statistico-linguistica con l'apporto di meta-informazioni di carattere linguistico (dizionari elettronici, lessici di frequenza, grammatiche locali) e con interventi sul testo (normalizzazione, lemmatizzazione e lessicalizzazione).

4.1.1 Le variabili di un testo

I testi, a differenza dei dati in forma numerica, non hanno una forma idonea per un loro pronto utilizzo a fini di analisi statistica. Per loro natura, i dati testuali sono dati non strutturati, pertanto necessitano di molteplici operazioni di rielaborazione, che consentono di trasformarli in dati statistici, e quindi in una matrice analizzabile con tecniche statistiche.
La prima considerazione da fare è, come per ogni ricerca, identificare gli obiettivi. In base a questi, viene selezionata la base di dati (in questo caso di testi) oggetto di interesse. Una base di dati testuali è composta da una collezione di testi *(corpus)* selezionati secondo un criterio di ricerca, precedentemente fissato. I *corpora* possono essere costituiti da libri, *chat*, opinioni, articoli di giornale, *e-mail*, risposte a domande aperte, etc. I documenti contenuti nella base dati selezionata, rappresentano le unità statistiche (gli individui di un'indagine classica) della popolazione investigata. Ma quali sono le variabili? In una analisi classica, le variabili sono rappresentate sia da informazioni riguardanti le caratteristiche socio-demografiche degli individui sia da informazioni specifiche, utili a conoscere il fenomeno indagato. L'obiettivo dell'analisi di un testo è identificare il contenuto del testo e le informazioni contenute in esso. Quindi le variabili da studiare sono proprio le parole che lo caratterizzano e che quindi ricorrono in esso. Una 'parola' è definibile come una sequenza di caratteri, a cui è associabile un significato. Se pensiamo, per esempio al caso della lingua italiana, già

possiamo immaginare che la variabilità di un testo può raggiungere dimensioni difficilmente gestibili. Inoltre, se pensiamo che alle differenti parole esistenti nel vocabolario della lingua italiana si aggiungono le forme flesse di ogni parola (bello-a-e-i), la dimensione della variabilità aumenta esponenzialmente. Quindi, quasi mai, nella costruzione della matrice di dati, vengono considerate tutte le parole presenti nel *corpus*. Le considerazioni fatte finora sono valide se consideriamo il significante di una parola (definito da Saussure come la 'faccia esterna' del segno, mentre la 'faccia interna' è il significato. Si veda l'enciclopedia Treccani.). La parola acquisisce un significato di contesto, e quindi il suo significato viene definito dalle parole che la circondano. Questo concetto rappresenta l'ambiguità di una parola, che chiaramente rende ancora più complessa la comprensione del testo. Le ragioni dell'ambiguità possono essere di natura lessicale e semantica, e quindi: può essere dovuta al fatto che talune forme sono identificate da una medesima stringa di caratteri dell'alfabeto (omografia) o dalla medesima pronuncia (omofonia), o una stessa forma, con un unico significante, può assumere significati differenti a seconda dei contesti in cui è usata (polisemia). Per condurre un'analisi testuale che produca risultati rilevanti, cioè riducendo l'ambiguità esistente nel testo, è necessario scegliere l'unità di analisi. In questo contesto, l'unità di analisi è il tipo di 'parola considerata', e cioè, il lemma (bello), la forma grafica (belle), i poliformi (carta di credito) (di cui parleremo di seguito nel dettaglio). La scelta di un'unità di analisi rispetto ad un'altra condiziona l'intera analisi, in quanto riduce la variabilità del testo, ma perdendo parte dell'informazione; per questo la scelta dell'unità di analisi deve essere molto ragionata. Operativamente, dopo la scelta dell'unità di analisi, bisogna procedere all'effettiva trasformazione del testo (secondo il criterio di unità prescelto) e alla pulizia dello

stesso. Queste operazioni possono essere effettuate automaticamente con il supporto di *software* specifici, che consentono di visualizzare la distribuzione statistica delle parole all'interno del *corpus*, trasformando il testo in un vocabolario (lista di parole contenute nel testo, per ognuna delle quali viene misurata il numero di volte in cui occorre). L'ampiezza del vocabolario V è definita dal numero di parole presenti nel testo:

$$V = V_1 + ... + V_k + ... + V_{fmax} \tag{4.1}$$

dove V_1 rappresenta il numero di parole che si presentano una volta sola all'interno del testo (detti hapax), V_k il numero di parole che si presentano K volte, e V_{fmax} la frequenza della parola con il maggior numero di occorrenze nel vocabolario. A questo punto, con le procedure di pretrattamento si trasformano le parole nel tipo di unità prescelto (lemma, poliforme, forma grafica). Tuttavia, il problema della dimensione della variabilità e dell'ambiguità delle parole resta ancora irrisolto. Per ridurre ulteriormente la variabilità del testo è necessario procedere con la *feature selection*, operazione attraverso la quale, fissato un criterio, si selezionano solo le parole idonee a soddisfarlo. Delle tecniche di *feature selection*, si parlerà nel dettaglio successivamente. Una volta selezionate le parole (le variabili di interesse), si può procedere alla costruzione della matrice dei dati, chiamata matrice lessicale, dove le righe sono rappresentate dai testi e le colonne dalle parole, a cui può essere associato un sistema di pesi (vedi paragrafo 4.1.4).

4.1.2 La scelta dell'unità di analisi

A seconda degli obiettivi un'unità di analisi può essere una forma grafica, un lemma, un poliforme o una forma mista (lessia), in grado di catturare

al meglio il contenuto presente nel testo[**?**].
Nella statistica testuale le analisi basate sulle forme grafiche (sequenza di parole delimitata da due separatori) hanno il vantaggio di essere indipendenti dalla lingua. Si tratta di un approccio puramente formale, che privilegia i segni per arrivare al senso come rappresentazione del contenuto. L'uso delle forme grafiche risolve l'ambiguità del linguaggio attraverso l'uso di strumenti di analisi multidimensionale. Tale analisi, misurando la somiglianza lessicale di profili, considera i profili vicini tra loro come profili simili [33]. Dall'altra parte, i linguisti computazionali considerano il lemma (parola che per convenzione è scelta per rappresentare tutte le forme di una flessione) come unità di analisi [34]. Dizionari elettronici, risorse statistico-linguistiche, sono gli strumenti adottati in accordo con questo approccio, dipendente dalla lingua. Nel campo dei dati testuali, Bolasco [63] considera un'unità mista dipendente dal linguaggio (forma grafica/-lemma/polirematica) denominato 'forma testuale'. La forma testuale può essere rappresentata da una o più parole che esprimono un unico significato. Questo approccio risolve l'ambiguità attraverso l'uso di metadati.

4.1.3 Procedure di Pretrattamento

Per effettuare un'analisi automatica del testo è necessario trasformare il *corpus* oggetto di analisi in un formato direttamente trattabile con strumenti statistici. L'insieme delle operazioni da effettuare, note come fase di pretrattamento, consistono nello svolgimento di diversi passi integrati fra loro:

- *Parsing*
- *Normalizzazione*

- *Costruzione del vocabolario di lavoro*
- *Estrazione dei segmenti*
- *Lessicalizzazione*
- *Tagging Grammaticale*
- *Tagging Semantico*
- *Lemmatizzazione.*

Il primo passo da compiere è l'individuazione delle successioni di caratteri dell'alfabeto comprese tra i separatori, attraverso una procedura, detta *parsing*. Questa procedura scansiona il testo, identificando le forme grafiche presenti in esso. La *normalizzazione*, invece, agisce sull'insieme dei caratteri non separatori, eliminando le possibili 'repliche' del dato, ne sono esempio le forme grafiche con lettera iniziale maiuscola o minuscola. Inoltre, attraverso questa procedura è possibile uniformare le forme che presentano forte variabilità, come ad esempio date, sigle e nomi propri. Dopo la fase di normalizzazione si procede con la costruzione del vocabolario di lavoro. Sul vocabolario è possibile identificare l'unità minimale di senso attraverso la fase di *lessicalizzazione*. Per effettuare questa procedura è, però, necessario identificare, fissando a priori una soglia di frequenza, i poliremmatiche (es. carta di credito) e le poliformi presenti nel testo (segmenti ripetuti). Dalla loro identificazione è possibile marcare gli stessi all'interno del corpus, creando un nuovo vocabolario di forme testuali. Per integrare le conoscenze sul corpus, e quindi carpire le informazioni ivi comprese, è possibile procedere con una serie di operazioni che consentono di annotare il vocabolario con meta informazioni. Il *tagging grammaticale* rappresenta

una delle principali operazioni, che attraverso l'identificazione della categoria grammaticale delle parole, consente di suddividere le POS (*Part Of Speech*) funzionali (articoli, preposizioni, congiunzioni) dalle POS formali (nome, aggettivo, verbo). Un altro tipo di *tagging* è quello semantico, che attraverso l'uso di risorse statistico-linguistiche consente di annotare il vocabolario con meta-informazioni di tipo semantiche. In ultimo, considerando l'unità di analisi scelta, è possibile procedere alla *lemmatizzazione* del testo. Questa procedura trasforma le forme grafiche in lemma, producendo una grande standardizzazione del testo e una grande riduzione della dimensione. Ma è una scelta da ponderare, in quanto cancella gran parte dell'informazione.

4.1.4 La codifica delle basi di dati testuali

Lo schema maggiormente utilizzato per la codifica di corpora è il cosiddetto *Bag-of-Words* [58]. Con questa codifica ogni documento (o frammento di testo) contenuto nel corpus viene trasformato in un vettore i cui elementi sono i pesi attribuiti alle parole contenute nel vocabolario. Formalmente, ogni documento D_j è visto come un vettore nello spazio del vocabolario:

$$D_j = (w_{1j}, w_{2j}, w_{3j}, ..., w_{pj}) \tag{4.2}$$

dove ogni termini w_{ij} è il peso della *i-ma* forma nel *j-mo* documento. In base all'analisi da effettuare è possibile scegliere differenti schemi di ponderazione. Seguendo lo schema *Bag-of-words*, i documenti sono organizzati in una matrice T, nota come tabella lessicale, con p righe e q colonne. Sulle righe si trovano le p parole contenute nel vocabolario, mentre sulle colonne ci sono i q documenti considerati. Le celle di questa matrice *parole x documenti* (Figura 4.1) possono contenere le frequenze delle parole (numero di

volte che occorrono nel documento), oppure la semplice presenza o assenza della parola in un documento, oppure il valore di un indice statistico.

	doc 1	doc 2	doc 3	$\cdots$	doc q
forma 1	1	0	0	$\cdots$	1
forma 2	0	2	1	$\cdots$	0
forma 3	1	0	1	$\cdots$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
forma p	1	0	0	$\cdots$	0

Figura 4.1: La matrice lessicale.

4.1.5 Feature selection

La matrice di dati così costruita, è una matrice sparsa, ed è inoltre di dimensioni elevate, si pensi a quante parole vengono utilizzate per scrivere una singola frase. Ai fini di un'analisi statistica è importante eliminare il rumore contenuto nei dati, per effettuare l'analisi solo sulle variabili considerate rilevanti. Quando la base di dati è di tipo testuale, le variabili sono rappresentate dalle parole, e quindi è opportuno ridurre la dimensione dei dati, selezionando le parole rilevanti. La procedura utilizzata per la selezione è detta *feature selection* in analogia alle procedure cui si ricorre per l'analisi di grandi basi di dati numerici. La funzione di selezione più comune è l'uso della lista di *stop-words*, attraverso la quale si eliminano le parole considerate poco rilevanti, ne sono esempio il verbo essere e avere,

gli articoli, le preposizioni, le congiunzioni e le parole comuni a tutti i documenti (quindi poco discriminanti). Inoltre, viene utilizzato lo *stemming*, operazione attraverso la quale si riportano le forme flesse di una parola alla sua radice. Ma questi metodi non risolvono problemi di selezione specifici, e vengono solitamente utilizzati in una varietà di applicazioni non supervisionate. Seguendo un approccio di classificazione supervisionato, è possibile effettuare la selezione con l'uso di etichette di classe. Questo procedimento garantisce che le carattteristiche polarizzate vengono raccolte per il processo di apprendimento. Una vasta gamma di metodi utilizzati per la *feature selection* (basati su indice di Gini, *information gain*, *Fmutual information*, Chi-quadrato, *linear discriminant analysis*, *generalized singular value decomposition*) sono discussi in Aggarwal e ChengXiang [46].

Capitolo 5

Alcune proposte metodologiche per la produzione di informazione statistica ufficiale da fonti documentarie

Le opportunità offerte dal ricorso a basi di dati non strutturate per la costruzione di informazione statistica, illustrata nei capitoli precedenti, si realizzano qui, attraverso la proposta di alcune procedure utili alla produzione di statistiche ufficiali partendo da basi documentarie. Da prima si affronta il problema della costruzione di risorse statistico linguistiche che consentano, in una successiva fase di strutturazione della base di dati (e, quindi, nella fase di analisi statistica), di tener conto di informazioni di contesto ignorate dagli strumenti classici di analisi dei dati testuali. Si propone una strategia basata sull'utilizzo congiunto di strumenti propri dell'analisi

delle Corrispondenze Lessicali e della *network text analysis* [60], così da annotare il testo con metainformazioni, utili per selezionare i termini rilevanti, e, in definitiva, per l'identificazione del contenuto del testo oggetto di analisi. Il problema della *high dimensionality*, caratteristico delle basi di dati documentarie, viene affrontato, in un ambito più legato all'*information retrieval*, attraverso la proposta di una strategia di *text classification* finalizzata alla costruzione di strumenti di interrogazione di testo più efficienti, perché riferite a porzioni di corpus ritenute rilevanti sulla base delle relazioni fra termini identificate all'interno di un *training set* di documenti e, successivamente validate [61]. In ultimo, si affronta un problema di grande rilievo al fine di produzione statistica da fonti secondarie, quando si dispone di informazioni sia numeriche che testuali. In questo ambito, si propone un metodo di analisi fattoriale che analizza congiuntamente variabili numeriche (siano esse continue o categoriche) e testuali, al fine di costruire un'informazione statistica sulla base di informazioni numeriche, ma con l'ausilio di informazioni testuali [62]. Il contesto metodologico di riferimento è quello dell'analisi delle corrispondenze canoniche.

5.1 L'uso dell'analisi delle Corrispondenze Lessicali unitamente alla *network analysis* per la costruzione di risorse statistico-linguistiche

Le risorse statistico-linguistiche sono strumenti utili ad arricchire le conoscenze del ricercatore sul testo analizzato, in quanto consentono di annotare il testo con meta informazioni, utili a comprenderne i contenuti. Inoltre, le risorse possono essere utili a selezionare la presenza nel testo di alcune parole rilevanti, rispetto ad un qualche argomento. Esistono molteplici risorse

statistico-linguistiche, già implementate nei principali *software* di analisi dei dati testuali, ma ognuna di loro è stata creata per uno specifico argomento. Quì si propone la costruzione di una risorsa statistico-linguistica, con l'uso di tecniche che studiano le relazioni esistenti tra parole, a partire dalla base di dati oggetto di interesse. Questo consente di identificare insiemi di parole che esprimono i 'concetti' *ad-hoc* contenuti nel testo analizzato. Dall'identificazione di 'concetti' proponiamo uno strumento per affrontare la disambiguazione presente nei documenti espressi in linguaggio naturale.

5.1.1 L'analisi delle corrispondenze lessicali

L'Analisi delle Corrispondenze (AC) è un metodo fattoriale, solitamente applicato per analizzare tabelle di contingenza. Da un punto di vista computazionale e matematico, l'AC è una tecnica semplice e il suo successo è dovuto soprattutto alla capacità di attrazione delle sue rappresentazioni grafiche. Greenacre [36] dà la seguente definizione: l'AC è una tecnica per la visualizzazione delle righe e colonne di una matrice di dati (in primo luogo, una tabella di contingenza a due dimensioni) come punti di un spazio vettoriale di dimensioni ridotte. Secondo Lebart [35], l'AC è principalmente caratterizzata da: una elaborazione simmetrica di righe e colonne, l'uso di una speciale distanza euclidea ponderata, nota come metrica del chi-quadrato; formule di transizione semplici, che consentono una rappresentazione simultanea di righe e colonne (grafico congiunto). Generalmente l'AC viene eseguita per analizzare una tabella lessicale T, di dimensioni (I, J), dove I rappresentano le parti di un corpus, e J le parole, utile ad identificare le strutture semantiche latenti esistenti nel corpus e per rappresentare graficamente le relazioni lessicali latenti.

5.1.2 *Network Analysis*

La *Network Analysis* (NA) è l'analisi di un insieme di relazioni esistenti tra oggetti. Le relazioni non sono le proprietà degli oggetti, ma quelle di 'sistemi di oggetti', appartenenti ad un sistema relazionale più grande [37]. I grafici delle reti sono lo strumento per rappresentare il sistema relazionale esistente nei dati. In una rappresentazione grafica di una rete, gli oggetti (vertici) sono rappresentati da punti, e le relazioni sono disegnate come linee che connettono coppie di vertici. La NA ha trovato applicazione in molti campi scientifici come la biologia, l'economia, la linguistica e, in particolare nelle scienze sociali, dove, negli ultimi anni, è stata riposta una notevole attenzione, sia da un punto di vista metodologico che applicativo verso questo strumento (che prende il nome di *social network analysis*). Popping [38] ha definito la *network text analysis* come un metodo utile, per la codifica dei rapporti esistenti tra parole in un testo, e per la loro rappresentazione grafica attraverso la rete. Il presupposto è che il linguaggio e la conoscenza possono essere modellati come reti di parole e di relazioni esistenti tra di loro. Una panoramica dei diversi modi per costruire reti di dati testuali è presentata in Batagelj *et al.* [39].

5.1.3 L'interazione tra l'analisi delle corrispondenze e le tecniche di *network analysis*: un dibattito aperto

In letteratura, è stato affrontato un ampio dibattito sull'uso dell'AC per analizzare i dati relazionali. Anche se l'AC è inclusa nei principali pacchetti di NA (ad esempio Ucinet, Pajek,...), sono sorte molte obiezioni sulla sua idoneità ad analizzare questo tipo di dati. Contro l'uso dell'AC per la rappresentazione dei dati relazionali, Borgatti e Everett [40] sostengono che le mappe bidimensionali saranno quasi sempre gravemente imprecise e

fuorvianti nella visualizzazione delle reti; inoltre le distanze non sono euclidee, ma gli utenti della tecnica trovano molto difficile la comprensione delle mappe in un qualsiasi altro modo. A nostro parere, il primo argomento di Borgatti e Everett è legato alla riduzione dimensionale eseguita dall'AC e alla pratica comune di riferirsi solo al primo piano fattoriale, quando le informazioni utili possono essere ottenute con mappe fattoriali costruite su i successivi assi. Inoltre, i fattori dell'AC non possono essere utilizzati solo per la visualizzazione, in quanto possono essere utili per interpretare e sintetizzare la struttura relazionale tra gli oggetti. Dobbiamo considerare un numero adeguato di fattori, che ricostruiscono la struttura relazionale in uno spazio dimensionale ridotto. Inoltre, l'AC dà una misura del raccordo dei dati alla struttura latente. Il secondo punto di Borgatti e Everett mostra un malinteso. La distanza del Chi-quadrato dei profili può essere osservata in uno spazio fisico, trasformando i profili prima di tracciare gli assi, in modo che un'unità su ciascun asse ha una lunghezza fisica inversamente proporzionale alla radice quadrata del corrispondente elemento del profilo medio [43]. Uno dei principali sostenitori dell'AC per l'analisi dei rapporti sociali è Bourdieu [42]. Egli afferma che la NA riduce la struttura di interazione, mentre l'AC consente l'estrazione delle relazioni latenti, non direttamente visibili. Pertanto L'AC soddisfa i requisiti del pensiero relazionale e l'altra classe di modalità relazionali deve essere respinta. Tuttavia, secondo de Nooy [41], l'AC non può andare in profondità nei rapporti manifesti, a causa della riduzione della dimensionalità e quindi gli strumenti di NA sono indispensabili, fornendo un modo per rappresentare le relazioni reali. Noi vogliamo entrare nel dibattito, mostrando che l'efficacia dell'uso congiunto dei due approcci può essere utile per l'analisi dei dati testuali. La nostra proposta è quella di indagare l'insieme dei termini, identificati

come 'rilevanti' dall'AC, grazie ad uno strumento particolare di NA, le *ego network*.

5.1.4 La nostra proposta

Analisi delle corrispondenze Il primo passo della nostra strategia consiste nell'effettuare un'analisi delle corrispondenze lessicali per identificare la struttura semantica latente nella tabella lessicale. La matrice di dati che andiamo ad analizzare è la tabella lessicale T, che è una tabella di contingenza in cui sono catalogate I parti e le J parole di un corpus. Dove i suoi elementi $t(i, j)$ esprimono rispettivamente il numero di volte che viene trovato il termine j nella parte i del corpus. L'interpretazione dei risultati dell'AC include la comprensione dei due risultati numerici e grafici. L'AC decompone una misura di associazione, cioè la Φ^2, che rappresenta la dispersione (varianza) delle nuvole di punti (sia nello spazio occupato dai termini che nello spazio delle parti) attorno al baricentro comune definito dall'ipotesi d'indipendenza. La parte di varianza spiegata l'alpha-esimo asse è data dalla alpha-esimo autovalore, λ_α, ottenuta con una decomposizione generalizzata in valori singolari della matrice T, con i vincoli di orto-normalizzazione. La decomposizione in valori singolari è dato da:

$$T = U^T \tag{5.1}$$

and

$$U^T D_I^{-1} U = V^T D_J^{-1} V = I \tag{5.2}$$

dove D_I^{-1} rappresenta la distribuzione marginale delle parti del corpus e D_J^{-1} rappresenta la distribuzione marginale dei termini. La percentuale di

varianza τ_α valutate con il fattore $\alpha - esimo(\tau_\alpha = \lambda_\alpha/\Sigma_\alpha)$ è un indice descrittivo che ha la sua importanza. Si noti che la misura di associazione Φ^2 è uguale a $\Sigma_\alpha\lambda_\alpha$. Il numero di fattori da conservare è relativo alla somma dell' inerzia spiegata dai fattori maggiori e in ogni caso i cui corrispondenti autovalori sono superiori alla media. Altre due misure sono importanti per l'interpretazione dei risultati di un'AC: il contributo assoluto e il contributo relativo. Il contributo assoluto (CA) di un elemento ad un asse indica la percentuale di varianza dell'asse principale, spiegata dall'elemento. È dato dal quadrato della coordinata moltiplicato per la frequenza dell'elemento. Il contributo relativo (CR), invece, è una misura della qualità della rappresentazione di un punto su un asse, ed è dato dal quadrato del coseno dell'angolo formato dalla proiezione del punto sull'asse principale e il vettore che congiunge il punto al centro di gravità. Usiamo questi due tipi di contributi per individuare i termini rilevanti che caratterizzano la struttura latente semantica.

Network Analysis Ai fini di un'analisi di rete abbiamo bisogno di dicotomizzare la tabella lessicale T, per ottenere una matrice binaria A di dimensioni (IXJ) in cui il generico elemento $a(i,j)$ con $i = (1,\ldots,I)$ e con $j = (1,\ldots,J)$ è uguale a 1 se il termine j ricorre almeno una volta nella parte i e 0 nel caso contrario. Dalla matrice A deriva la matrice di co-occorrenza W di dimensioni (JXJ), calcolata dal prodotto:

$$W = A^T A \tag{5.3}$$

L'elemento $w(k,j)$ rappresenta il valore di co-occorrenze dei termini k e j, in cui $w(j,j)$ è il numero di occorrenze dei j termini. Secondo la NA, la matrice W è una matrice di adiacenza pesata, non orientata che, può essere

utilizzata per analizzare le relazioni esistenti tra i parametri valutati. Al fine di normalizzare le co-occorrenze introduciamo un noto indice di similarità, cioè l'indice di Jaccard. Secondo questa misura, la somiglianza tra due termini k e j viene definita come:

$$S_{ij} = \left(\frac{w_{kj}}{w_{kk} + w_{jj} - w_{kj}} \right) \tag{5.4}$$

Si ottiene una matrice simmetrica $S(JXJ)$ se il valore della cella $s(k, j)$ indica la misura normalizzata delle co-occorrenze per i termini k e j. Dopo aver ottenuto una matrice di adiacenza binaria X, andiamo a dicotomizzare la matrice S come segue: per ogni $s(k, j)$ con il valore superiore ad una soglia predefinita si imposta il valore dell'elemento $x(k, j)$ uguale a 1 e 0 nel caso contrario. La scelta del valore di soglia è basato sulla distribuzione effettiva dell'indice di *Jaccard* nei dati. Come abbiamo sottolineato in precedenza, andiamo a studiare le relazioni tra i termini che ricorrono nel *corpus*. Ai fini dell'analisi abbiamo scelto di studiare le *ego network* delle parole rilevanti, che illustrano le aree locali di un'intera rete. Una *ego network* nasce 'estraendo' dai dati di una rete regolare, una rete che consiste di un nodo focale ('*ego*'),e di tutti i soggetti collegati a quel nodo (*'alters'*), e tutti i collegamenti tra questi e gli altri attori [44]. Studiare le *ego network* dà la possibilitàdi comprendere il ruolo che gioca un nodo in una struttura relazionale e fornisce alcune informazioni sulla rete nel suo complesso.

L'unione di AC e NA per l'analisi testuale dei dati Per unire l'AC con la NA abbiamo scelto il punto di vista delle rappresentazioni grafiche, in modo da rappresentare sia le relazioni lantenti che manifeste esistenti

tra le parole. Inoltre, proponiamo strumenti di text mining per ulteriori analisi. Nella fase dell'AC abbiamo individuato le parole più rilevanti come l'intersezione di due insiemi di termini CA e CR. I CA sono composti dai termini con i più alti contributi assoluti sul primo piano fattoriale, che sono importanti per capire la struttura latente. I CR si compongono dei termini con i più alti contributi relativi sul primo piano fattoriale, che sono i migliori rappresentati. Nella fase successiva di NA, consideriamo l'intersezione di CA e CR. Le condizioni selezionate sono scelte come nodi focali per le ulteriori analisi. Per visualizzare le relazioni esistenti tra i nodi focali, costruiamo una rete. Con l'utilizzo di *software standard* per l'analisi di rete (ad esempio Ucinet), si ottiene una rappresentazione grafica da un algoritmo *spring embedding*. Con l'obiettivo di ottenere una rappresentazione metrica, in linea con la struttura latente, proiettiamo la rete sulla prima mappa fattoriale. Inoltre, per ogni nodo focale viene costruita la sua *ego network*, utile a definire concetti e termini ad esso collegato, esplorando i rapporti di ogni nodo focale con i suoi *alters* (non sono state analizzate le connessioni tra gli *alters*). Questa procedura può portare a situazioni differenti: se la base testuale è stata correttamente pre-elaborata e le relazioni latenti e manifeste risultano coerenti, si ottiene una rappresentazione metrica della rete, e possono essere costruite le risorse statistico-linguistiche per ulteriori analisi. Le incoerenze possono, invece, produrre configurazioni difficili da interpretare. Ad esempio, se la trama sembra difficile da leggere, problemi come la disambiguazione devono essere esaminati.

5.2 Categorizzazione di testi: una metodologia per la costruzione semi-automatica di *query*

L'uso di fonti secondarie, tipicamente in formato di testi, è un'occasione non sufficientemente esplorata dagli istituti nazionali di statistica. L'uso di documenti, espressi in linguaggio naturale, può creare alcuni problemi, legati alla complessità e alle costose procedure di pretrattamento di cui necessitano i testi. Inoltre, non è assolutamente chiaro come strutturare il *set*set di dati da analizzare. In questo lavoro si propone una strategia basata sulle tecniche di classificazione del testo per la produzione di statistiche ufficiali da nuove fonti testuali. Questa strategia si basa sulla classificazione del testo, sfruttando l'organizzazione strutturale dei documenti in frasi e la conoscenza esterna [45]. La strategia consiste di due fasi: nella prima fase, fissato un bisogno informativo, la categorizzazione del testo è impegnata a discriminare un insieme di parole interessanti da quelle prive di interesse (avente una conoscenza approfondita di ingresso). Nella seconda fase, si applicano le regole logiche costruite nel passo precedente, per identificare in quali parti del documento è presente l'informazione espressa dalla regola. Questo ci consente di costruire delle *query* testuali.

5.2.1 Categorizzazione di testi

Il problema della classificazione è stato ampiamente studiato nelle comunità di *data mining* e *information retrieval*. Abbiamo un insieme D di N *training record* $= d_1, \ldots, d_n$, in modo che ogni *record* è etichettato con un valore di classe provenienti da un insieme di K valori discreti diversi $1, \ldots, k$. Il *training set* di dati viene utilizzato per costruire un modello di classificazione che lega le caratteristiche (parole) delle parti a differenti etichette di

classi, definite a priori. I problemi generali della classificazione statistica nel dominio testuale possono essere riassunti come segue: abbiamo un *training set* di vettori di documenti, ognuno con una classe, denominata valore obiettivo. Ogni oggetto del *training set* è rappresentato nella forma (d_i, c_k) dove d_i è il vettore di documenti e c_k il valore dell'etichetta della classe. L'obiettivo è quello di imparare una mappa o una funzione che sia in grado di prevedere un valore della classe c_k di un nuovo documento. Il grado di 'coincidenza' tra la funzione obiettivo e il classificatore determina l'efficacia dell'algoritmo di classificazione. Questa operazione simula il processo umano di valutare la rilevanza di un documento rispetto al tema di interesse. In termini generali, abbiamo tre principali famiglie di tecniche statistiche per la categorizzazione del testo: alberi di decisione, metodi di regressione e le reti neurali; inoltre, nell'ambito del *machine learning* sono stati proposte diverse metodologie, tra le quali citiamo *support vector machines*, analisi discriminante lineare, metodi di inferenza bayesiana, modelli massima entropia, algoritmi genetici e le regole di associazione. Per una panoramica completa delle metodologie sviluppate nell'ambito della categorizzazione dei testi si può consultare Charu Aggarwal e Chengxiang [46]. Quì si descrive la metodologia utilizzata nella strategia proposta: gli alberi di decisione. La scelta di questa metodologia risiede nella sua semplice applicabilità e interpretazione dei risultati. Infatti gli alberi di decisione producono regole logiche semplici da interpretare e quindi da generalizzare a contesti d'indagine simili. Ci sono un numero enorme di algoritmi sviluppati per costruire un albero di decisione. Un albero di decisione è essenzialmente una scomposizione gerarchica dello spazio dei dati in cui viene utilizzato un predicato o una condizione sul valore dell'attributo per dividere gerarchicamente lo spazio dei dati. L'albero è costruito su un *training set* con

un algoritmo iterativo. Il *training set* consiste in un insieme di documenti etichettati, cioè documenti per i quali la categoria è nota a priori. Ad ogni passo, è scelta la variabile che maggiormente discrimina (divide) i dati in gruppi, che presentano il più grande miglioramento in termini di punteggio, della funzione scelta. L'iterazione si ferma se ciascun gruppo ('foglia') contiene un unico individuo (nel caso dei testi un documento) oppure se è soddisfatta la regola di arresto fissata. L'obiettivo è quello di trovare un modello complesso sufficiente a catturare le strutture esistenti nei dati, ma non così complesso da generare *overfitting*. Gli algoritmi più utilizzati sono CART [47] e C4.5 [48].

5.2.2 Una strategia per la categorizzazione dei testi

In questo lavoro si propone una strategia di classificazione del testo che mira a discriminare un documento interessante da quelli non interessanti. L'interesse dipende dalle esigenze informative specifiche. La strategia si compone di due fasi. Dato un bisogno informativo, nella prima fase, vengono identificati i documenti interessanti da quelli non interessanti; operativamente i primi saranno etichettati con il valore 1 e i secondi con il valore 0. Questa procedura viene effettuata solo su una parte dei documenti, attraverso l'input esterno di conoscenza esperta. Ciò consente di costruire il *training set*, che sarà l'input necessario per la seconda fase della strategia. Nella seconda fase, il *training set* viene analizzato con l'algoritmo di segmentazione (C4.5). Durante questa procedura, l'algoritmo di apprendimento tramuta la classificazione del *training set* in regole logiche. Nel contesto dei dati testuali, l'algoritmo identifica l'insieme delle parole rilevanti che devono essere presenti in un testo per considerarlo interessante o meno. Le regole prodotte vengono applicate sul *test set*, attribuendo ai documenti non clas-

sificati un'etichetta di classe. Come precedentemente detto, la funzione di selezione consiste nell'identificare le parole rilevanti per ridurre la dimensionalità del corpus analizzato. Un nuovo corpus di dimensione inferiore è ottenuto, considerando solo i testi relativi al tema di interesse.

5.3 Introduzione all'analisi delle Corrispondenze Canoniche

Gli INS diffondono le statistiche economiche da loro prodotte sottoforma di microdati e di grafici. Non sempre però questi strumenti di comunicazione sembrano essere efficaci ad accrescere la conoscenza dei loro utilizzatori. Da un'indagine condotta nell'ambito del progetto BLUE-ETS sembrano essere pochissime le imprese che riescono ad utilizzare le statistiche ufficiali per supportare le decisioni in campo aziendale. Questo fenomeno è spesso determinato dalla struttura dei metadati che accompagnano le statistiche. Ciò genera una bassa interazione tra gli INS e le imprese, le quali si trovano solo a fornire dati, senza riuscire a fruire in modo soddisfacente del servizio statistico. In questo lavoro si propone l'uso di un metodo fattoriale, che analizza congiuntamente variabili quantitative e testuali. La scelta dell'uso di questo metodo (noto in ambito ecologico) è stata determinata dalla sua rappresentazione grafica, che cosente di proiettare sullo stesso piano (piano vincolato allo spazio delle variabili quantitative) le variabili quantitative e quelle testuali. Il metodo è noto come analisi delle corrispondenze canoniche. Dalla lettura del piano fattoriale è possibile chiarire ed arricchire le informazioni numeriche con variabili testuali (le parole).

5.3.1 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 [49] è un esempio di queste tecniche di analisi multidimensionale dei dati e può essere considerata come un metodo di ordinamento non vincolato [50]. 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 [51]. 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. 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$el'ACCnonèaltrocheunaAC$.

5.3.2 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.* [52] hanno proposto una tecnica chiamata Ordinamento Gaussiano. Pertanto il modello di risposta per le specie è rappresentato dalla funzione campanulare:

$$E(y_{ik}) = c_k exp[1/2(x_i - u_k)^2/t_k^2]. \tag{5.5}$$

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: ck, massimo della curva di risposta delle specie; uk, la moda, ossia il valore di x per cui si ottiene il massimo e tk, 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:

$$x_i = b_0 + \sum_{j=1}^{q} b_j z_{ij} \tag{5.6}$$

Dove b_0 è l'intercetta, b_j è il coefficiente di regressione della $j - esima$ variabile ambientale e xi 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 [53] 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:

$$\lambda U_k = \sum_{j=1}^{n} y_{ik} x_i / y_k \tag{5.7}$$

$$x_i^* = \sum_{k=1}^{m} y_{ik} u_k / y_i \tag{5.8}$$

$$b = (Z'RZ)^{-}1 Z'Rx^* \tag{5.9}$$

$$x = Zb \tag{5.10}$$

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 nxn; $Z = z_{ij}$ è una matrice di dimensioni $n(q+1)$ contenente i valori delle variabili ambientali e una colonna di 1; b, x , x^* sono tre vettori colonna: $b = (b_0, b_1, \ldots, b_q)'$, $x = (x_1, \ldots, x_n)'$ e $x^* = (x_1^*, \ldots, x_n^*)'$. 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 con $\lambda = 1$);
- *Step3*: Calcolare i nuovi scores dei siti 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: $\sum y_i x_i = 0$ e $\sum y_i x_i^2 = 1$

- *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.

5.3.3 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.

Capitolo 6

Produzione di informazione statistica in campo aziendale

I metodi proposti nel precedente capitolo, sono stati applicati alla relazione sulla gestione delle società quotate sul mercato regolamentato italiano. La relazione sulla gestione è un documento che viene allegato al bilancio di esercizio che contiene informazioni, quantitative e testuali, relative all'andamento della gestione dell'anno considerato. Questo documento rappresenta un esempio calzante, di una risorsa secondaria (dato amministrativo), ma sottoforma di testo (nuova fonte). In questo capitolo vengono presentati i risultati ottenuti.

6.1 L'unione di AC e NA per la costruzione di risorse statistico-linguistiche

6.1.1 Le operazioni di pre-trattamento

Per la redazione della relazione sulla gestione, la legge italiana non prescrive una struttura definita. Diversamente, nel mercato statunitense, la legge spiega accuratamente tutte le informazioni che la relazione sulla gestione deve contenere. Per questo motivo e per il nostro obiettivo, abbiamo scelto una società quotata in entrambi i mercati: *Luxottica Group*, *leader* mondiale di *eyewears*. È stato analizzato il documento relativo all'esercizio 2009. Il commento di gestione di Luxottica è composto da 44 sezioni e 22.621 *tokens*. È ben noto che la procedura di pre-trattamento del testo, ha un ruolo fondamentale, ed influisce sui risultati finali di qualsiasi analisi testuale dei dati o di procedure di *text mining*. Abbiamo eseguito la seguente strategia di pretrattamento, con l'utilizzo del *software* TALTAC 2.10:
STEP 1. è stata applicata la procedura di normalizzazione, e di pulitura del testo con l'uso di una lista di *stop-words*, riducendo il *corpus* a 4.315 *types*.
STEP 2. è stato effettuato il tagging grammaticale del *corpus*, e sono stati selezionati solo i termini classificati come aggettivi, sostantivi e verbi.
STEP 3. Tra i termini non inclusi nella precedente selezione (*STEP* 2) sono stati selezionati quelli con un alto valore dell'indice Tf-idf [58]. Si noti che lo *STEP* 3 è motivata dall'alta percentuale (circa il 40 per cento) delle forme ambigue (termini ai quali non poteva essere associata una sola categoria grammaticale, e quindi il *software* identifica

la parola come forma ambigua). Dopo queste procedure, il *corpus* pre-trattato è costituito da 5627 occorrenze e 403 forme grafiche. La tabella lessicale utilizzata per ulteriori analisi è una matrice rettangolare, documenti (per documenti quì si intendono le sezioni contenute nella relazione sulla gestione) per termini, di dimensione (44 x 403).

6.1.2 I risultati dell'analisi delle corrispondenze e della *network analysis*

Eseguiamo una CA sulla nostra tabella lessicale, utilizzando il *software* SPAD. Da un punto di vista statistico lo scree plot mostra che i primi 10 assi rappresentano la struttura latente esistente nei dati. I primi 10 assi spiegano il 75 per cento dell'associazione nella tabella e ognuno con una percentuale più elevata rispetto alla media. Tuttavia, dato che il nostro obiettivo non è analizzare la relazione sulla gestione di Luxottica da un punto di vista economico, ma fondamentalmente trovare una mappa dove poter rappresentare le relazioni esistenti tra le parole, abbiamo scelto un punto di taglio più stretto, concentrando la nostra attenzione sul primo piano fattoriale (che spiega il 26 per cento della variabilità). Nella figura 6.1, il primo fattore (14,5 per cento della variabilità) oppone alla sezione relativa alle prestazioni economiche ('andamento economico') rappresentata sul lato sinistro, la sezione dedicata ai principali eventi accaduti durante l'anno ('principali eventi del 2009'). I termini con maggiori contributi assoluti mostrano una contrapposizione tra il linguaggio usato per descrivere i componenti del bilancio d'esercizio (vendita, periodo, milione, euro, utile, attività) e i termini che descrivono eventi economici (finanziaro, mercato, marchio, risultato) posizionati sul lato destro della mappa.

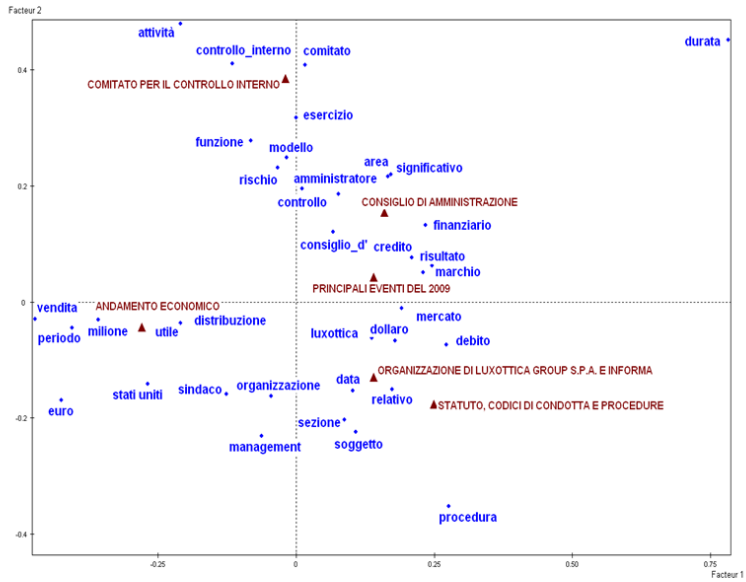


Figura 6.1: L'analisi delle corrispondenze: il primo piano fattoriale.

Il secondo fattore (11,4 per cento di varianza spiegata) oppone alla sezione relativa al controllo interno ('comitato per il controllo interno') la sezione che descrive gli aspetti organizzativi ('organizzazione') di luxottica group e statuto, codici di condotta, procedura rappresentato nella parte bassa del grafico. I termini di individuazione degli aspetti di controllo sono rispettivamente 'controllo interno', 'comitato', 'transazione sul mercato' e 'funzione'. In questo modo siamo in grado di definire le strutture latenti semantiche dei termini utilizza-

ti nella relazione sulla gestione. Il passo successivo della procedura è quello di individuare i termini che saranno trattati come nodi di analisi della rete. A tale scopo abbiamo scelto i termini che hanno sia il più altro contributo assoluto che relativo. Con l'intersezione del 90-esimo percentile dei contributi assoluti e il 90-esimo percentile dei contributi relativi si ottengono 14 termini: 'area', 'attività', 'comitato', 'controllo-interno', 'transazione sul mercato', 'euro', 'funzione', 'milione', 'periodo', '*section*', 'significativo', 'stati-uniti', 'utile' e vendita.

La Figura 6.2 illustra la migliore rappresentazione bidimensionale della struttura semantica latente del *corpus* e la posizione relativa dei termini selezionati, ma ignora i collegamenti reali. Per andare in profondità e scoprire le relazioni reali, disegnamo una rete di relazioni tra i termini ('nodi focali') individuati dall' analisi delle corrispondenze. La rete in Figura 6.3 mostra tre componenti: un nodo isolato (vendita), un componente con 8 nodi e 14 collegamenti e un componente con 5 nodi e 10 collegamenti (tutti i nodi sono adiacenti).

La Figura 6.4 illustra l'efficacia della nostra strategia di collaborazione tra AC e NA. La componente con i 5 nodi (comitato, controllo-interno, transazione sul mercato, funzione, section) ha una buona (leggibile) rappresentazione sulla mappa e può essere utilizzata per la costruzione di un risorsa lessicale utile ai fini di uteriori analisi sulle attività di controllo interno. La componente con 8 nodi (area, attività, euro, milione, periodo, significativo, stati uniti, utile) mostra chiaramente problemi linguistici: i termini attività, area e significativo sono posizionati lontano dal centro della rete e quindi necessitano

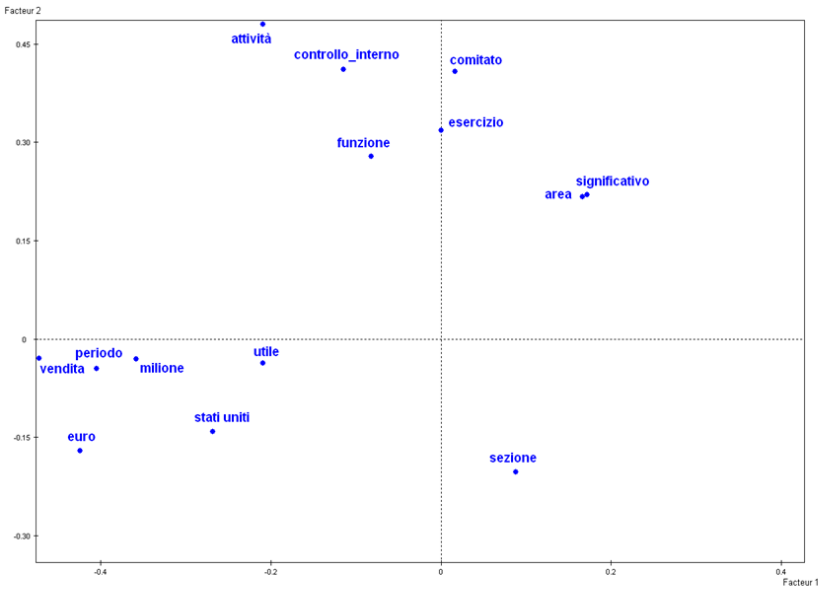


Figura 6.2: Il primo piano fattoriale dell'analisi delle corrispondenze: i termini selezionati.

di ulteriori approfondimenti. Consideriamo la *ego network* costruita intorno alla zona del nodo focale. In figura 6.5 sono rappresentati i suoi *'alters'* nel primo piano fattoriale dell'AC. Emergono due diversi usi del termine area. Se guardiamo sul lato sinistro uno dei suoi *alters*, 'euro', definisce un significato monetario del termine area ('area dell'euro'), mentre dagli '*alters*' posizionati sul lato destro si definisce il concetto di area come mercato di sbocco ('Sud Africa', 'Europa',

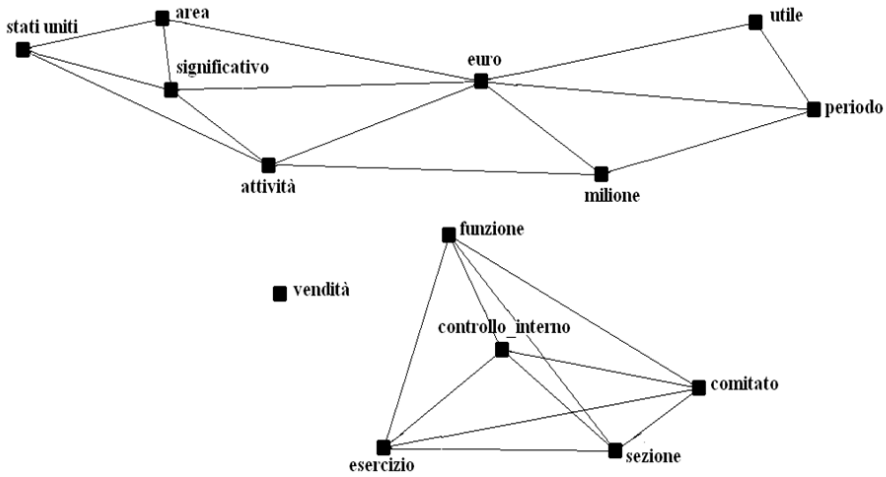


Figura 6.3: La rete dei termini selezionati.

'Australia'). La posizione interessante di stati Uniti e distribuzione che si oppongono al commercio all'ingrosso, fornisce una nuova visione del primo asse fattoriale, che viene interpretato come i canali di distribuzione utilizzati da *luxottica*.

Del punto isolato, 'vendita', viene illustrata la sua *ego network* nella figura 6.6. Dal grafico possiamo vedere che tutti gli *alters* di 'vendita' descrivono le attività aziendali ('*luxottica*', 'prodotto', 'marchio', 'mercato', 'finanziario', 'operativo', 'risultato', 'affari', 'negozio', 'suc-

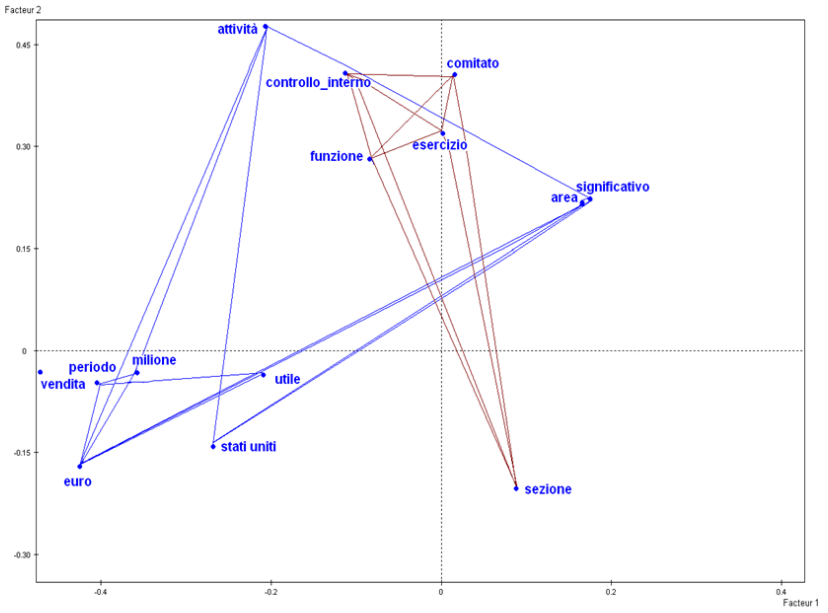


Figura 6.4: Output della strategia proposta.

cesso', 'vista', 'produzione', 'licenza' e 'distribuzione').

Dall'esplorazione della *ego network* del nodo 'vendita' proiettato sul primo piano fattoriale dell'AC (figura 6.7), è possibile notare che alcune parole rilevanti ('*luxottica*', 'mercato', 'marchio', 'finanziario', 'risultato', 'distribuzione'), presentano un elevato contributo assoluto, ma un basso contributo relativo. Anche se questoa circostanza meriterebbe ulteriori approfondimenti, ci induce a sospettare che la parola vendita assume diversi molteplici significati, in corrispondenza

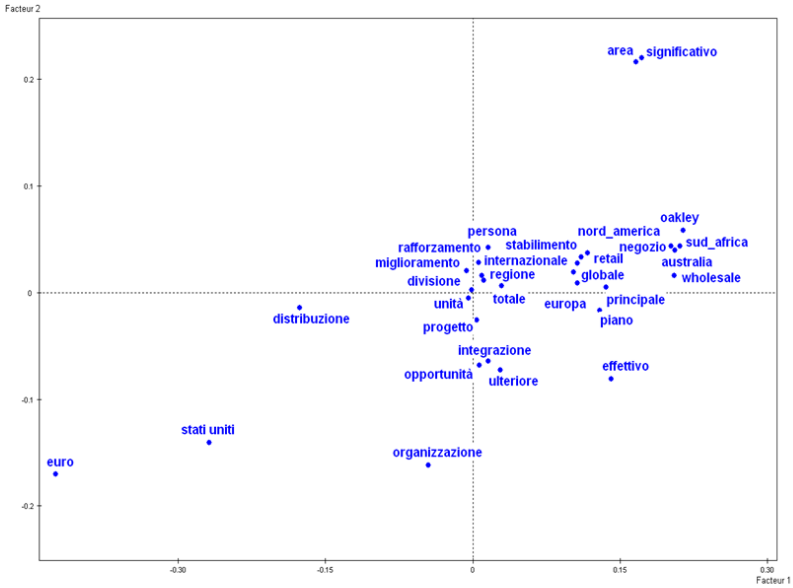


Figura 6.5: Gli *alters* del nodo focale 'area' sul primo piano fattoriale.

della sua collecazione, come ad esempio 'punti vendita'. Questa ipotesi è sottolineata dalle relazioni manifeste alla base della *ego network* (figura 6.6). Dall'analisi dell'AC si individuano i collegamenti della parola vendita agli aspetti economici dell'attività dell'impresa.

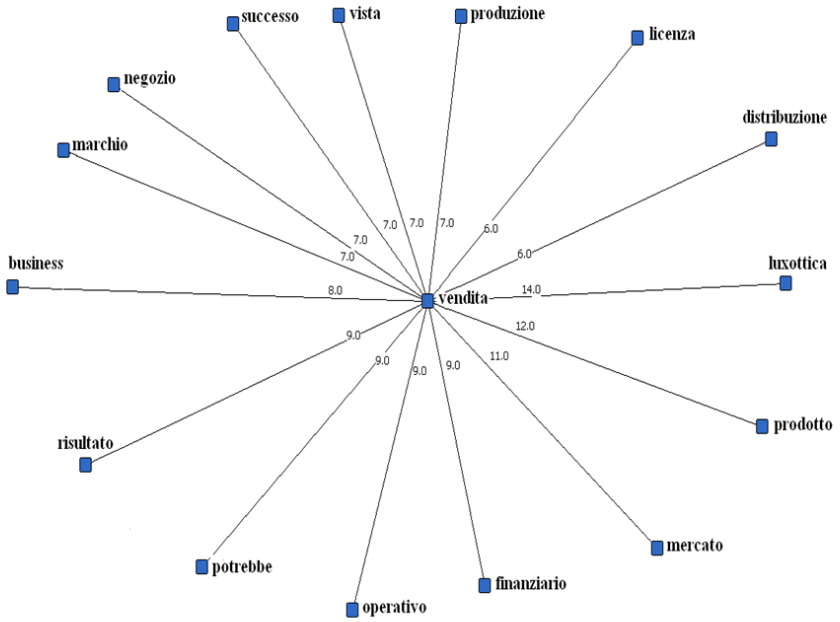


Figura 6.6: Ego network di 'vendita'.

6.1.3 Conclusioni

In questo lavoro si è dimostrata l'efficacia dell'uso congiunto di analisi delle corrispondenze e *network analysis* per i dati testuali. La strategia che proponiamo può essere uno strumento utile per la produzione di fonti linguistiche e per individuare i problemi relativi alle procedure di pretrattamento di cui necessita un documento espresso in linguaggio naturale, e quindi a ridurre la l'ambiguità delle parole

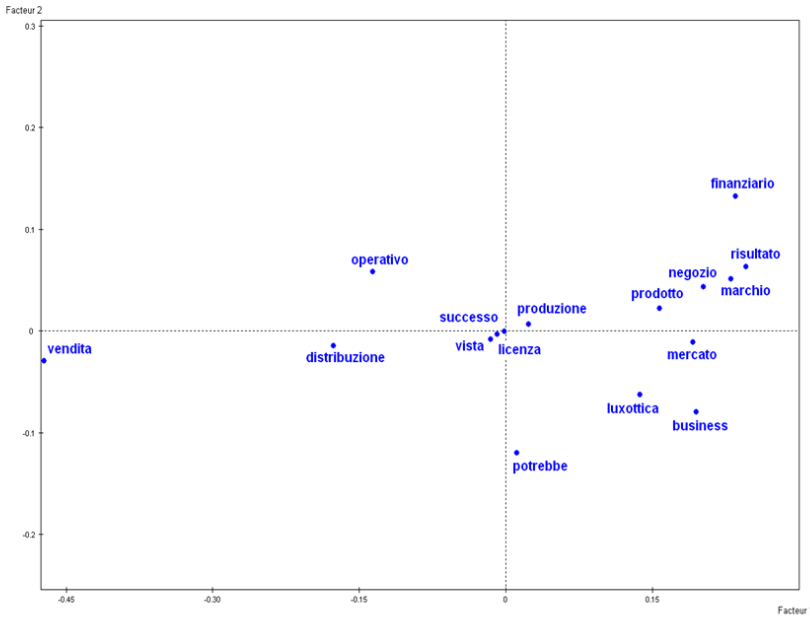


Figura 6.7: Gli *alters* del nodo focale 'vendita' sul primo piano fattoriale.

presenti in esso.

6.2 Una strategia di classificazione del testo per l'estrazione di informazioni sulle imprese

Come precedentemente affermato, la relazione sulla gestione, è un documento espresso in linguaggio naturale. Da un punto di vista statistico, per utilizzare questa fonte per la produzione di statistiche,

è necessario trasformare i dati non strutturati in dati strutturati. A tal fine, si prropone una strategia basata sulla classificazione del testo. La strategia, a titolo di esempio è stata applicata ad ad un'unica sezione della relazione sulla gestione delle società quotate sul mercato italiano regolamentato.

6.2.1 Il primo *step*: la costruzione di regole logiche per identificare le parole rilevanti di un testo

La base di dati analizzata è stata selezionata con l'uso di un campionamento per quote che rispecchia il settore di appartenenza delle società, secondo la classificazione NACE. Le relazioni sulla gestione considerate sono relative all'anno 2011. È stata considerata solo una sezione del documento (omogenea a tutte le società considerate) relativa ai principali eventi accaduti nell'esercizio considerato. Al fine di ridurre il range di variabilità, la sezione di interesse è stata suddivisa in frasi (una frase è un insieme di parole che si combinano tra loro secondo precise regole grammaticali ed è la massima sequenza di un testo in cui vigono relazioni sintattiche e in cui le parole seguono un certo ordine), per cui le frasi rappresentano i nostri documenti. Abbiamo raccolto 2.110 frasi. Abbiamo scelto casualmente un loro sottoinsieme per costruire il *training set*, che consiste in 30 società quotate, composto da 597 frasi. Dopo una codifica manuale dei commenti di gestione, abbiamo etichettato 452 frasi con il valore 0 (considerate non interessanti) e le restanti 145 con il valore 1 (frasi interessanti). Una volta che il *training set* è stato costruito, si è avviata la procedura di segmentazione automatica. L'analisi è stata effettuata con il *software R* (l'algoritmo utilizzato è il C5.0, che è una

estensione del C4.5). La procedura ha prodotto le regole per l'identificazione delle *'performance* economiche' di un'impresa (vedi figura 6.8). prestazioni economiche possono essere identificati con la regola:

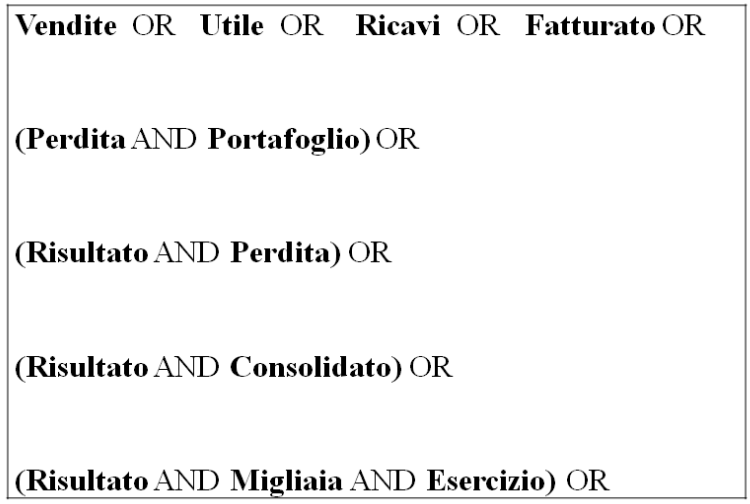


Figura 6.8: Regole logiche.

6.2.2 Il secondo *step*: le *performance* delle imprese

Con la costruzione delle regole logiche, sono state selezionate dall' intero data base solo le frasi che parlano delle *performance* economiche delle imprese. Così è stata ottenuta una notevole riduzione

della dimensione del *corpus*. Dall originale *set* di dati (753 MB) è stato estratto un sottoinsieme contenente solo 597 frasi (interessanti). Abbiamo così ottenuto una riduzione dimensionale pari al 95 per cento.

6.2.3 Conclusioni

Qui abbiamo proposto una metodologia, basata sulla *text classification*, con lo scopo di estrarre modelli di parole interessanti (in un dominio specifico). In questo modo, alcuni confini sequenziali sono stati costruiti, trasmettendo le informazioni specifiche di interesse. Questa divisione strutturale può essere più o meno evidente. Lo scopo di questa strategia è quello di sfruttare la struttura incorporata nel documento di testo e di migliorare le prestazioni delle attività di *text mining*, e dare ai ricercatori un modo semplice di analisi.

6.3 La relazione sulla gestione delle società quotate sul mercato regolamentato

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.

6.3.1 I 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.

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 poliremathiche 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 [55], 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.

Indicatori di *'Perfomance'* 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 'percentuale 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 e Young, Deloitte e 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*,proposti nell'analisi del contenuto in ambito di *financial analysts* [57].

L'indice Gulpease si calcola applicando la seguente formula:

$$89 - (\frac{Lp}{10}) + (3xFr). \tag{6.1}$$

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.

I risultati Per le similitudini già esplicitate nel precedente paragrafo, 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 Figura 6.9 e che complessivamente spiegano il 58,63 per cento della variabilità totale.

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 semias-

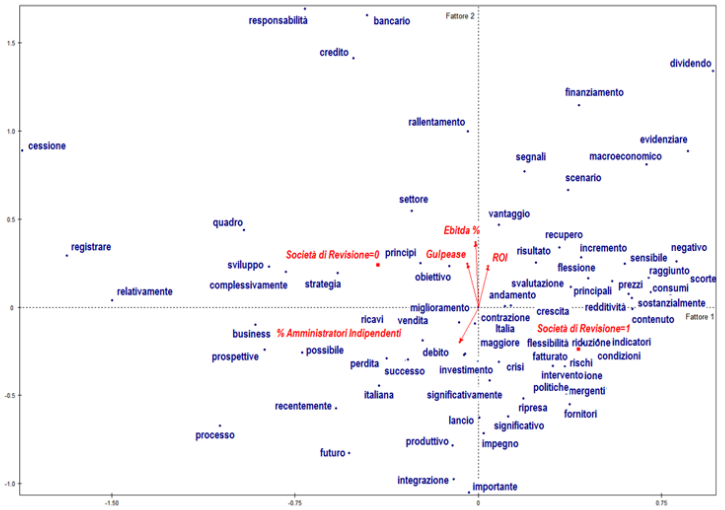


Figura 6.9: Rappresentazione congiunta delle parole e delle variabili di *'performance'*.

se 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'*.

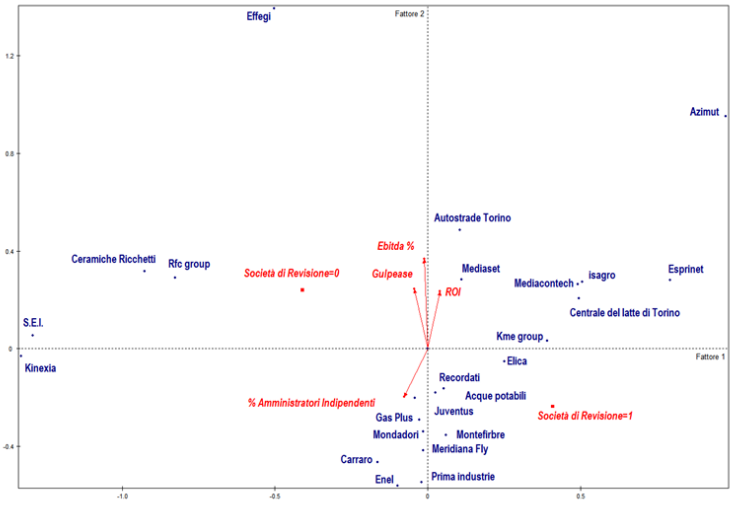


Figura 6.10: Rappresentazione congiunta delle società e delle variabili di '*performance*'.

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 ci 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*.

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 [54], 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 [56], che propon-

gono una tecnica denominata analisi delle corrispondenze canoniche non simmetriche (ACNC) applicata a dati ecologici.

Conclusioni

La crisi economica degli ultimi anni ha generato un nuovo modo di pensare *'up-to-date'*, cioè essere in grado di reagire tempestivamente ed in modo adeguato ai cambiamenti e agli eventi inattesi. Gli utenti: le imprese, le persone fisiche e soprattutto le istituzioni governative richiedono dati sempre più dettagliati, tempestivi e di buona qualità. Contestualmente i fornitori di dati, in particolare le imprese, richiedono di diminuire il carico amministrativo e statistico gravante su di loro.

Il contesto descritto ha sottolineato le carenze del sistema statistico esistente, ponendo quest'ultimo di fronte a nuove sfide. L'incessante sviluppo tecnologico e la crescente diffusione di dispositivi collegabili alla rete Internet sta creando una nuova miniera informativa (fonti secondarie) utile per la produzione di informazioni. Lo sviluppo della rete internet, dei motori di ricerca, del linguaggio XML, degli algoritmi di intelligenza artificiale, ha prodotto la possibilità di analizzare nuove informazioni (testi, documenti, *chat*, domande a risposta aperta etc.) con tecniche di elaborazione automatica. Se pensiamo alla enorme quantità dei documenti e delle informazioni contenute

nel Web, è facile spiegare il grande interesse sviluppato nell'ultimo ventennio verso il cosiddetto *'text mining'*, termine con il quale si identifica l'insieme delle tecniche utili ad analizzare, esplorare e interrogare raccolte di grandi basi di dati testuali. Con queste tecniche sono possibili indagini evolute, che consentono di estrarre informazioni utili a produrre valore (conoscenza).

Le opportunità offerte dal ricorso a basi di dati non strutturate per la costruzione di informazione statistica, si sono realizzate qui, attraverso la proposta di alcune procedure utili alla produzione di statistiche ufficiali partendo da basi documentarie. Da prima si è affrontato il problema della costruzione di risorse statistico linguistiche che consentano, in una successiva fase di strutturazione della base di dati (e, quindi, nella fase di analisi statistica), di tener conto di informazioni di contesto ignorate dagli strumenti classici di analisi dei dati testuali. Qui si è proposta una strategia basata sull'utilizzo congiunto di strumenti propri dell'analisi delle corrispondenze lessicali e della *network text analysis*, così da annotare il testo con metainformazioni, utili per selezionare i termini rilevanti, e, in definitiva, per l'identificazione del contenuto del testo oggetto di analisi.

Il problema della *high dimensionality*, caratteristico delle basi di dati documentarie, è stato affrontato, in un ambito più legato all'*information retrieval*, attraverso la proposta di una strategia di *text classification* finalizzata alla costruzione di strumenti di interrogazione di testo più efficienti, perché riferite a porzioni di corpus ritenute rilevanti sulla base delle relazioni fra termini identificate all'interno di un *training set* di documenti e, successivamente validate. In ultimo, si è affrontato un problema di grande rilievo al fine di produzione statistica da

fonti secondarie, quando si dispone di informazioni sia numeriche che testuali. In questo ambito, è stato proposto un metodo di analisi fattoriale (analisi delle corrispondenze canoniche) che analizza congiuntamente variabili numeriche (siano esse continue o categoriche) e testuali, al fine di costruire un'informazione statistica sulla base di informazioni numeriche, ma con l'ausilio di informazioni testuali.
Sebbene le nuove tecnologie forniscano importanti innovazioni per gli INS nella raccolta dei dati, ci sono molti problemi relativi alla qualità di tali dati e alla possibilità di trasformarli in informazioni statistiche degne di questo attributo. Il primo problema, e forse il più rilevante, è la distorsione di selezione, principalmente derivante da una distorsione del 'campione selezionato' e di un 'pregiudizio osservatore'. Per quanto riguarda il primo tipo di polarizzazione, il divario tecnologico crea un forte divario di selezione tra persone che utilizzano le nuove strumentazioni e le persono che non li utilizzano. Inoltre, gli utenti più esperti possono rendere i dati più difficili da raccogliere, utilizzando funzioni che limitano la visibilità e l'accessibilità ai propri dati.
È importante tenere presente i due lati della medaglia: da un lato, nuove modalità di raccolta dei dati sono a disposizione per raccogliere informazioni difficili da raggiungere con gli strumenti tradizionali, ma sorgono problemi relativi alla qualità dei dati, come la generalizzazione e la riservatezza. Tutte queste questioni devono essere considerate come nuove sfide per la teoria statistica.

Bibliografia

[1] Giovannini E., (2010) *Statistica 2.0: The Next Level*, Decima Conferenza Nazionale della Statistica, Roma.

[2] Bavdaz M., Bergstrom Y., Biffignandi S., Bolko I., Deirdre Giesen D.,(2011) *Business use of NSI statistics based on external sources*, BLUE-Enterprice and Trade Statistics, Bruxelles.

[3] Statistic Canada (2003) *Survey Method and Practices*. Ottawa.

[4] Moser C.A., Kalton G., (1971) *Survey Methods in Social Investigation*. Heinemann Educational Books Limited, London.

[5] Raj D., (1972) *The design of Sample Surveys*. McGraw-Hill Series in Probability and Statistics,New York.

[6] Groves R.M., (1989) *Survey Errors and Survey Costs*. John Wiley and Sons, New York.

[7] Cochran W.G., (1977) *Sampling Techniques*. John Wiley and Sons, New York.

[8] Biemer P.P., Groves R.M., Lyberg L.E., Mathiowetz N.A., (1991) *Measurement Errors in Surveys*. John Wiley and Sons, New York.

[9] Lesseler J.T., Kalsbeek W.D., (1992) *Nonsampling Errors in Surveys*. John Wiley and Sons, New York.

[10] Cox B.G., Binder D.A., Chinnappa B.N., Christianson A., Colledge M.J., Kott P.S., (1995) *Business Survey Methods*. John Wiley and Sons, New York.

[11] Dolson D., (1999) *Imputation Methods*. Statistic Canada, Ottawa.

[12] Brackstone G., (1987) *Statistical Issues of Administrative Data: Issues and Challanges*. Statistic Canada, Ottawa.

[13] Calzaroni M., (2011) *Le fonti amministrative nei processi e nei prodotti della statistica ufficiale*. Istat, Roma.

[14] Calzaroni M., (2004) *La cooperazione Inter-istituzionale: il valore aggiunto dell'integrazione di informazioni*. VII Conferenza nazionale di statistica, Roma.

[15] Vale S., (2011) *Using Administrative and Secondary Sources for Official Statistics* in *Handbook 'Principles and Practices'*, Unece.

[16] Costanzo L., Di Bella G., Hargreaves E., Pereira H. J., Rodriguez S., (2011) *An Overview of the Use of Administrative Data for Business Statistics in Europe. ESSNET Admin Data: Workpackage 1*, Bruxelles. http://essnet.admindata.eu/WorkPackage?objectId=4251

[17] Daas P.J.H., Vis-Visschers R.J.W.M., Arends-Toth J., (2009) *Checklist for the Quality Evaluation of Administrative Data Sources*. Statistics Netherland, the Hague.

[18] Friedman J. H., (2011) *The role of statistics in the data revolution*. In *International Statistics Review*, 69, pp.5-10.

[19] Blue-ets Project: Work Package 5, (2012)*New Ways for Collecting and Analysing Information: Deliverible 5.1: Report on principles of fuzzy methodology and tools developed for use in data collection.*

[20] Balbi S., Triunfo N., (2011) *Statistical Tools for jointly analysing open and close ended questions in surveys.* In *Statistical Surveys: thinking about methodology and applications*, pp.61-72, Springer, Heidelberg.

[21] Tinto A., della Ratta Rinaldi F., (2011) *Le opinioni dei cittadini sulle misure del benessere*, DISA, Istat. http://www.misuredelbenessere.it/fileadmin/relazione-questionarioBES.pdf

[22] della Ratta Rinaldi F., Loré A., (2010) *Lavoro e i suoi contenuti. Un'applicazione di text mining per categorizzare le attività dettagliate di lavoro nell'indagine campionaria sulle professioni ISTAT.* Proceedings of JADT 2010 : 10th International Conference on statistical analysis of textual data. Vol.2.

[23] Canzonetti A., Misuraca M., Romano M.C., (2006) *Analyzing the language of everyday life: how textual statistics can support time use surveys. The Italian experience..* Atti del XLIII Riunione Scientifica della Societá Italiana di Statistica, pp.69-80.

[24] Daas P., Roos M., Blois C., Hoekstra R., Bosch O., Ma Y., (2011) *New Data Sources for Statistics: Experiences at Statistics Netherlands.* Paper for the 2011 European NTTS conference, Brussels, Belgium.

[25] Miller C. C., (2009) *Why Adults Have Fed Twitter's Growth.* The New York Times 26 August 2009.

[26] Ahas R., Tiru M., Saluveer E., Demunter C., (2011) *Mobile Telephones and Mobile Positioning Data as Source for Statistics: Estonian Experiences.* Paper for the 2011 European NTTS conference, Brussels, Belgium.

[27] Konttinen J.P., Gimeno M. V., (2011) *Electronic Data Collection in Accommodation Statistics.* Paper for the 2011 European NTTS conference, Brussels, Belgium.

[28] Lebart L., Salem A. (1994) *Statistique textuelle*, Dunod, Paris.

[29] Zampolli A., Calzolari N., (1995) *Problemi, metodi e prospettive nel trattamento del linguaggio naturale: l'evoluzione del concetto di risorse linguistiche.* In Cipriani R., Bolasco S. (eds.), pp. 51-68.

[30] Guiraud P., (1954) *Les caractères statistiques du vocabulaire*, Puf, Paris.

[31] Brunet E., (1981) *Le vocabulaire francais de 1789 à nos jours*, Slatkine e Champion, Genève Paris.

[32] Lebart L., Salem A. (1988) *Analyse statistique des données textuelles*, Dunod, Paris.

[33] Benzécri J. P., (1981) *Pratique de l'Âanalyse des données Linguistique et lexicologie*, Dunod, Paris.

[34] De Mauro T. (1989) *I vocabolari ieri e oggi.* In *Il vocabolario del 2000* a cura di IBM Italia, Roma.

[35] Lebart L., Salem A., (1998) *Exploring textual data.* Kluver Academic Publishers, Dordrecht.

[36] Greenacre M.J., (1984) *Theory and applications of Correspondence Analysis*. Academic Press, London.

[37] Scott J., (2000) *Social Network Analysis: A Handbook*. Sage, London.

[38] Popping R., (2000) *Computer-assisted Text Analysis*. Sage, London.

[39] Batagelj V., (2002) *Network Analysis of texts*. Paper online: http://nl.ijs.si/isjt02/zbornik/sdjt02-24bbatagelj.pdf

[40] Borgatti S., Everett M., (1997) *Network analysis of 2-mode data*. Social Network, 19: 243-269.

[41] de Nooy W., (2003) *Fields and networks: corresponding analysis and social network analysis in the framework of field theory*. Poetics, 31: 305-327.

[42] Bourdieu P., (1991) *The Craft of Sociology*. Introduction in Bordieu P., Chamboredon J-C. and Passeron J-C. Walter de Guyter, Berlin.

[43] Greenacre M. J., (1993) *Correspondence analysis in practice*. Academic press, London.

[44] Hanneman R., Riddle M., (2005) *Introduction to social network methods*. Riverside, University of California. http://faculty.ucr.edu/ hanneman/

[45] Balbi S., Di Meglio E., (2004) *A text mining strategy based on the local contexts of words*. Proceedings of JADT 2004: 10th International Conference on Statistical Analysis of Textual data, Louvain-La-Neuve, Belgium, Vol 1, pp 79-87.

[46] Charu Aggarwal C., ChengXiang Z., (2012) *Mining Text Data*. Springer, New York. ISBN: 978-1-4614-3223-4.

[47] Breiman L., Friedman J. H., Olshen R. A., Stone C. J., (1984) *Classification and Regression Trees*. Champaman e Hall, New York.

[48] Quinlan J. R., (1986) *Introduction of decision tree*. In Machine learning 1, 81-106.

[49] Benzécri J. P., (1973) *L'analyse des Données: Tome I: La Taxonomie. Tome 2: L'analyse des Correspondance*. Dunod, Paris.

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

[51] Wittaker R.H., (1967) *Gradient analysis of vegetation*. Biological Reviews, vol.(49): 207-264.

[52] Gauch H. G., Chase G. B., Whittaker R. H., (1974) *Ordination of vegetation samples by Gaussian species distributions*. Ecology, vol. (55): 1382-1390.

[53] 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.

[54] Balbi S., (1995) *Non symmetrical correspondence analysis of textual data and confidence regions for graphical forms*. In Bolasco S. *et al.* Actes des 3es Journèes internationales d'Analyse statistique des donnèes Textuelles. Cisu, Roma, vol.(II): 5-12.

[55] Salton G., (1989) *Automatic text processing: the trasformation, analysis and retrieval of information by computer.* Addison-Wesley, Boston.

[56] Willem P., Galindo Villardon M. P., (2008) *Canonical non-symmetrical correspondence analysis: an alternative in constrained ordination.* Statistics and operations Research Trasactions, vol.(32): 93-111.

[57] Lehavy R., Li F., Merkley K., (2011) *The Effect of Annual Report Readability on Analyst Following and the Properties of Their Earnings Forecasts.* American Accounting Association, vol. (86): 1087-1115.

[58] Salton G., Buckley C., (1988) *Term-weighting approaches in automatic text retrieval.* Information Processing e Management, 24 (5): 513-523.

[59] Bolasco S., D'avino E., Pavone P., (2007). Analisi dei diari giornalieri con strumenti di statistica testuale e text mining. In *I tempi della vita quotidiana. Un approccio multidisciplinare all'analisi dell'uso del tempo.* Roma, Istat.

[60] Balbi S., Stawinoga A., Triunfo N., (2012). Text mining tools for extracting knowledge from firms annual reports. In *Actes des 11es Journées internationales d'Analyse statistique des donnèes Textuelles..* Liegi, Universitè de Liège.

[61] Triunfo N., (2013) *Obtaining the businesses information by applying a Text Classification strategy.*NTTS 2013, Bruxelles.

[62] Spano M., Triunfo N., (2012). La relazione sulla gestione delle societÃ Italiane quotate sul mercato regolamentato. In *Actes des*

11es Journèes internationales d'Analyse statistique des donnèes Textuelles.. Liegi, Universitè de Liège.

[63] Bolasco S., (1990). Sur diffèrentes stratègies dans une analyse des formes textuelles: une expèrimentation à partir de donnèes denquete. In *Jornades Internationals DÂAnalisi de Dades Textuals.* Barcellona, UPC.

[64] Bolasco S., (2005). *Statistica Testuale e Text Mining: alcuni paradigmi applicativi.* Quaderni di statistica, Vol. 7, Roma.