

XXXIX CONFERENZA ITALIANA DI SCIENZE REGIONALI

LAVORATORI AUTONOMI E SVILUPPO LOCALE: UN'ANALISI DI NETWORK

Paola Bosso¹, Stefano De Santis², Dario Ercolani³, Vincenzo Spinelli⁴

¹ Istat, V. Balbo, Roma, paola.bosso@istat.it

² Istat, V. Balbo, Roma, sdesantis@istat.it

³ Istat, V. Balbo, Roma, ercolani@istat.it

⁴ Istat, V. Balbo, Roma, vispinel@istat.it

ABSTRACT

Lo scopo del presente lavoro è fornire una rappresentazione esaustiva dei business network in cui sono inseriti i lavoratori autonomi, un aggregato socio-economico di grandissima rilevanza nel contesto nazionale che rappresenta oltre il 45% del totale delle imprese attive censite nell'Archivio Statistico delle Imprese Attive – ASIA e il 10% del totale del valore aggiunto.

L'enfasi del presente lavoro si concentra in maniera particolare sulle caratteristiche dei business network locali, ossia della rete commerciale (transazioni) in cui sono inseriti i lavoratori autonomi, tracciando un quadro ideale che spazia dai lonely players (“battitori liberi”) sino agli attori economici che costituiscono nodi rilevanti di tessuti economici importanti. Sfruttando gli strumenti analitici della network analysis, vengono definiti per aree territorialmente compatte (Sistemi Locali del Lavoro) dei cluster di relazioni ed indagata l'appartenenza alla rete in termini di maggiore redditività/efficienza della singola impresa;

Cardinalità della rete; profondità, densità ed estensione delle relazioni; proprietà dei nodi (fatturato, età dell'impresa, localizzazione, livelli di istruzione) consentono una innovativa rappresentazione del problema, fornendo al contempo importanti spunti per le politiche regionali

1 Introduzione

Lo scopo del presente lavoro è fornire una rappresentazione esaustiva dei business network in cui sono inseriti i lavoratori autonomi, un aggregato socio-economico di grandissima rilevanza nel contesto nazionale che rappresenta oltre il 45% del totale delle imprese attive censite nell'Archivio Statistico delle Imprese Attive – ASIA e il 10% del totale del valore aggiunto. Il lavoro autonomo in Italia ha caratteristiche sempre più variegate. Esso è costituito da circa 2.800.000 imprese che rappresentano quasi 4.500.000 di occupati. L'analisi delle fonti amministrative e fiscali disponibili consente sia di delineare i diversi profili professionali sia di osservare le reti commerciali tra lavoratori. Il nostro target di riferimento è il lavoro autonomo "individuale", inteso come le imprese in cui la titolarità dei diritti e delle obbligazioni giuridiche assunte nell'ambito dell'attività d'impresa è in testa ad una singola persona fisica.

Nel presente lavoro saranno descritti aspetti strutturali economici, il quadro di riferimento normativo, le fonti dei dati e le procedure di integrazione che hanno permesso di costruire e realizzato il dataset per l'analisi dell'impresa individuale e del network commerciale.

Inoltre è stata svolta una descrizione delle imprese individuali, con riguardo alla posizione settoriale e territoriale: a livello uni e bivariato, sia a livello di analisi multivariata. Infine, attraverso una modellazione della produttività, è stata sintetizzata l'analisi svolta con i precedenti passi al fine di mettere in luce come la le variabili proposte a descrivere un profilo dell'impresa individuale, segnatamente l'ubicazione territoriale e il sistema di relazioni commerciali in cui è inserita, influiscano sulla profittabilità e sulla capacità imprenditoriale.

2 Integrazione di fonti per una mappatura delle dotazione di capitale umano in Italia

La prima parte riguarda la costruzione della base dati, per cui concentra la sua enfasi sull'aspetto metodologico di integrazione delle fonti in un'unica base dati, con una serie di indici tesi a valutare la qualità delle fonti e dell'abbinamento.

2.1 Fonti di dati

Registro statistico Istat Asia-Imprese. Costituito dalle unità economiche che esercitano arti e professioni nelle attività industriali, commerciali e dei servizi alle imprese e alle famiglie, fornisce informazioni identificative (denominazione e indirizzo) e di struttura (attività economica, addetti dipendenti e indipendenti, forma giuridica, data di inizio e fine attività, fatturato) di tali unità. Oltre a contenere le informazioni per le analisi sull'evoluzione della struttura delle imprese italiane e sulla loro demografia, il registro rappresenta la base di tutte le indagini Istat sulle imprese, viene utilizzato per le stime di Contabilità Nazionale e individua la popolazione di riferimento per i piani di campionamento e per il loro riporto all'universo. Dal 2011, con l'introduzione di importanti innovazioni nel processo e nella stima dei caratteri delle imprese, dal punto di vista definitorio e metodologico, il Registro è stato utilizzato come base informativa per riprodurre i dati oggetto del Censimento Industria e Servizi. Il Registro fornisce informazioni sulle imprese integrando quelle desumibili dalle fonti amministrative, gestite da enti pubblici o da società private e quelle da fonti statistiche: (1) gli archivi gestiti dall'Agenzia delle entrate per il Ministero dell'economia e delle finanze, quali l'Anagrafe tributaria, le dichiarazioni annuali delle imposte indirette, le dichiarazioni dell'imposta regionale sulle attività produttive (Irap), gli Studi di settore, i dati del modello Unico, quadro Rh; (2) i registri delle imprese delle Camere di commercio, industria, artigianato e agricoltura e gli archivi collegati dei soci delle Società di capitale e delle "Persone" con cariche sociali; (3) gli archivi dell'Istituto nazionale di previdenza sociale: le denunce retributive mensili e Mens per gli occupati dipendenti; le dichiarazioni trimestrali della manodopera agricola (modello Dmag); la Cassa integrazione a pagamento diretto; le posizioni contributive degli imprenditori artigiani e commercianti; la gestione separata parasubordinati; l'archivio delle denunce contributive lavoratori dello sport e dello spettacolo (ex Enpals); le posizioni degli assicurati iscritti alla gestione ex-Inpdap; (4) l'archivio dell'Inail, delle assicurazioni per i lavoratori con contratto di somministrazione; (5) l'archivio delle utenze telefoniche; (6) l'archivio dei Bilanci consolidati e di esercizio; (7) l'archivio degli Istituti di credito gestito dalla Banca d'Italia; (8) l'archivio delle società di assicurazioni gestito dall'Isvap.

L'Anagrafe tributaria e il Registro delle imprese sono le fonti utilizzate per l'identificazione delle unità statistiche del registro Asia. Tutte le altre sono utilizzate, in maniera esclusiva o in concomitanza con le

precedenti, per la stima dei caratteri o per il controllo di particolari sottoinsiemi. Nell'aggiornamento del Registro svolge un ruolo di rilievo il Portale delle imprese per la raccolta e la restituzione di informazioni nell'ambito delle rilevazioni condotte dall'Istat: la gestione delle segnalazioni effettuate direttamente dalle imprese in tale sistema consente un tempestivo aggiornamento dei caratteri anagrafici, dello stato di attività e dell'attività economica principale. Le variabili comprese nel registro sono classificate secondo tre tipologie: variabili identificative (ragione sociale, indirizzo e altri caratteri per l'esatta individuazione dell'unità sul territorio); variabili di stratificazione (attività economica dell'impresa classificata secondo la classificazione Ateco, forma giuridica, dimensione dell'impresa, in termini di addetti indipendenti e dipendenti medi annui e di fatturato); variabili demografiche (data di nascita e cessazione dell'impresa, data di eventi quali scopori, fusioni o procedure concorsuali, fallimenti, liquidazioni, eccetera). La base dati che contiene le informazioni strutturali sull'occupazione delle imprese è il registro Asia-Occupazione. Oltre alle variabili occupazionali, diverse per tipologia di lavoratore (dipendente, indipendente), nel 2014 è stata introdotta un'importante innovazione sul versante informativo demo-sociale, utilizzata nel presente lavoro: l'assegnazione del titolo di studio agli individui-lavoratori, risultato di una procedura che integra il titolo di studio rilevato al Censimento della popolazione 2011 con l'aggiornamento proveniente dalla Base informativa su istruzione e titoli di studio (Bit).

Frame Istat. Realizzato dal Dipartimento per i conti nazionali e le statistiche economiche, è un sistema informativo complesso per la stima delle Statistiche economiche strutturali basato sull'uso massivo di dati amministrativi provenienti da fonti diverse - Bilanci civilistici, Studi di settore, Modello Unico, modello IRAP e dati Inps - integrati con i dati dell'indagine campionaria dell'Istat sulle piccole e medie imprese e con la base informativa costituita da Asia. Frame contiene oggi dati individuali per le principali variabili del conto economico (Ricavi vendite e prestazioni, Spese per beni e servizi, Costo del lavoro, Valore della produzione, Costi intermedi, Valore aggiunto, Margine Operativo Lordo) su tutte le imprese con meno di 100 addetti, circa 4,4 milioni di unità nel 2016, e stime di dominio per le altre voci del conto economico. Da Frame è possibile ottenere stime settoriali-dimensionali-territoriali di elevata accuratezza ed elevatissimo livello di dettaglio. In risposta agli stimoli internazionali che emerge anche dai nuovi regolamenti comunitari di settore (FRIBS), Frame consente stime più accurate e coerenti temporalmente, a fronte della riduzione dei costi e dell'onere statistico complessivi, garantendo un sensibile miglioramento del grado di armonizzazione e coerenza del sistema complessivo delle statistiche economiche sulle imprese, nonché maggiori livelli di coerenza tra le statistiche strutturali annuali e la Contabilità Nazionale.

Frame-SBS territoriale. Nuova base informativa rilasciata dall'ISTAT che consente la stima per Unità di Attività economica Locale (UAEL), effettuata integrando il sistema informativo "Frame SBS" con il Registro delle unità locali di Impresa, che costituisce il censimento virtuale della unità locali. Contiene dati su occupazione, attività economica esercitata e consente di ottenere una stima per ogni UAEL del valore aggiunto e del costo del lavoro in modo coerente con l'input di lavoro, privilegiando un approccio di tipo bottom-up che attribuisce all'unità locale un peso calcolato in termini di monte retributivo. I nuovi dati e indicatori territoriali sono elaborati con cadenza annuale a partire dalla stima delle principali variabili di conto economico per ciascuna delle unità locali delle imprese industriali e dei servizi non finanziari residenti sul territorio nazionale (oltre 4,7 milioni di unità) consentendo così una notevole flessibilità in termini di schemi di classificazione e dettaglio di analisi

Certificazione Unica. La Certificazione Unica (CU), precedentemente definita Certificato Unico Dipendente (CUD), è una certificazione dei redditi introdotta a partire dall'anno 2015 (anno di competenza 2014). Assolve la funzione di certificazione di molteplici tipologie di redditi da parte del sostituto di imposta (lavoro dipendente, pensione, lavoro autonomo o redditi diversi). Poiché costituisce il mezzo attraverso cui certificare ai fini legali redditi di natura assai diversa, ha assunto la denominazione di "Certificazione Unica". Riepiloga i redditi corrisposti dal datore di lavoro o dall'ente pensionistico nell'arco del cosiddetto anno fiscale "allargato"⁵. Come visto abbraccia tutti le tipologie di reddito da lavoro dipendente e assimilato (redditi da pensione e redditi da lavoro autonomo coordinato e continuativo, remunerazioni dei sacerdoti, assegni periodici corrisposti al coniuge ecc.), nonché emolumenti connessi a redditi da lavoro autonomo svolto sia professionalmente che occasionalmente (emolumenti connessi a fatture emesse, prestazioni

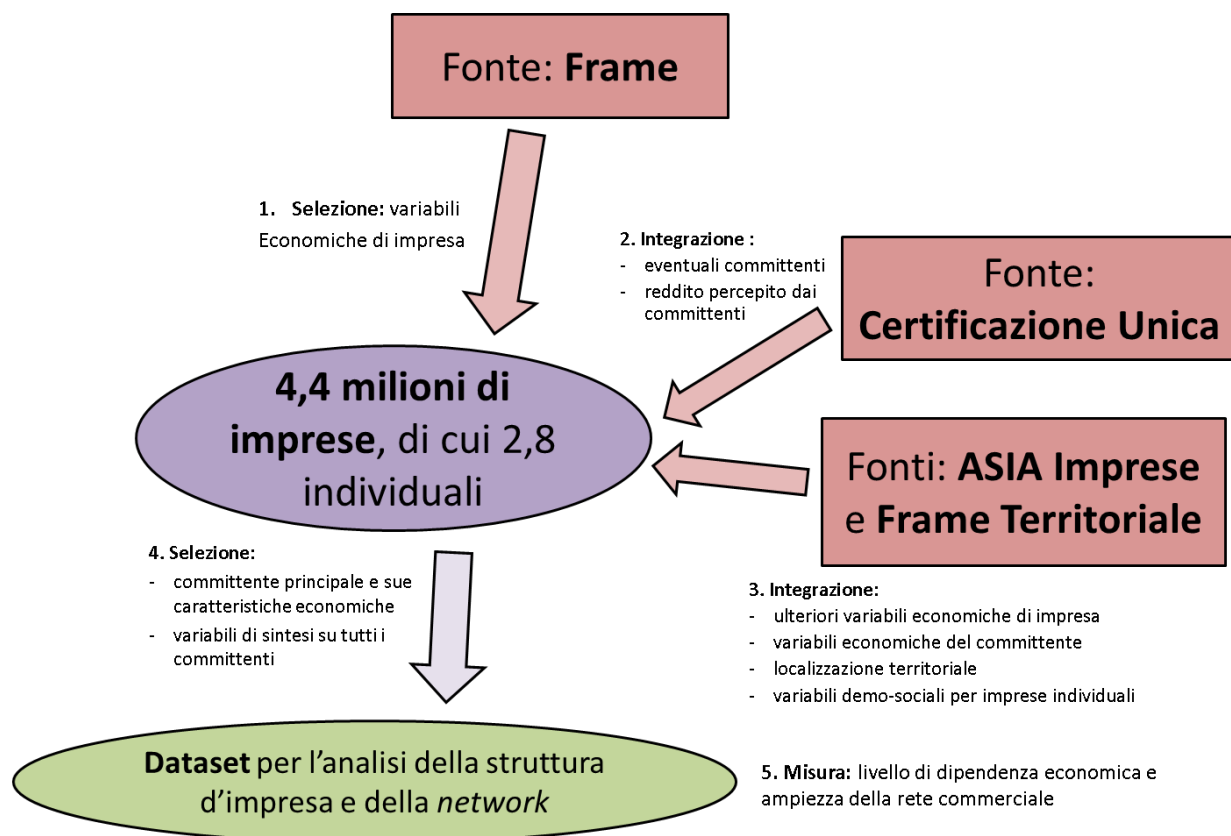
⁵ Corrispondente grosso modo all'anno solare, prolungato sino al 12 gennaio per consentire al sostituto d'imposta di effettuare i conguagli.

occasionali, utilizzo di brevetti, cessioni di diritti d'autore, redditi da locazione/sublocazione con intermediario ecc.). I dati contenuti sono anagrafici, natura fiscale, previdenziale. Data la natura fiscale-contributiva dell'adempimento, la copertura è totale.

La Base dati integrata su Istruzione e Titoli di Studio. La Base dati integrata su Istruzione e Titoli di Studio (BIT), realizzata nel Servizio Fonti amministrative e integrazione dei registri dell'Istat e rivolta all'utenza interna dell'Istituto, la Base Dati fornisce microdati di fonte amministrativa sui percorsi di istruzione e sui titoli di studio, integrati in modo centralizzato e automatizzato, utili per l'aggiornamento del grado di istruzione della popolazione, per il Registro tematico dell'Istruzione e per studi longitudinali. Si compone di dati annuali, dal 2011 in poi, sullo stato di chi è in un percorso di studi, dalla scuola primaria al dottorato di ricerca, e di dati di dettaglio dei titoli conseguiti. Attualmente integra informazioni quantitative di 15 dataset, coprendo più del 95% della popolazione in istruzione. La copertura aumenta nel tempo grazie alla disponibilità di nuove fonti ed è in via di miglioramento la tempestività degli output.

Le fonti espresse, sono state integrate al fine di realizzare un set di microdati, utile all'analisi delle caratteristiche socio economiche della popolazione di riferimento composta, come detto, dalle imprese individuali attive sul territorio italiano durante gli anni 2015 e 2016. Tali unità sono state estratte dalla fonte Frame e corrispondono a circa la metà delle imprese italiane. A queste, attraverso l'integrazione della fonte Certificazione Unica, sono state agganciate le informazioni relative ai propri sostituti di imposta, ovvero gli eventuali committenti. E inoltre, l'utilizzo della fonte Asia ha permesso da un lato di caratterizzare ulteriormente la popolazione di riferimento, dall'altro di qualificare i datori di lavoro individuati. La figura sottostante mostra sinteticamente il processo di integrazione delle fonti.

Figura 1 – Processo di integrazione delle fonti.



Di seguito sono elencate le principali variabili che costituiscono il tracciato record del dataset realizzato, e le loro fonti di provenienza.

La base di dati realizzata per gli anni 2015 e 2016 si compone, per ognuna delle annualità, di quasi 4,4 milioni imprese attive, delle quali circa 2,8 appartenenti alla categoria dei soggetti definiti titolari di P.Iva individuale. Oggetto dei risultati che verranno esposti sono le Partite Iva individuali con fatturato, distinte

per tipologia di rapporto con i datori di lavoro: imprese monocommittenti, pluricommittenti, operanti esclusivamente nel mercato Business-to-Consumer.

Tabella 1 – Data set integrato: variabili

Variabile	Fonte di dati
CODICE UNITA (CODICE FISCALE anonimizzato)	Frame
TASSO DI FEMMINILIZZAZIONE	ASIA
ETÀ MEDIA PER ADDETTO	ASIA
TITOLO DI STUDIO MEDIO PER ADDETTO	ASIA (Bit)
DINAMISMO ECONOMICO TERRITORIALE	Frame TERRITORIALE
SPECIALIZZAZIONE TERRITORIALE	Sistemi Locali del Lavoro
COMUNE	ASIA
REGIONE	ASIA
PROVINCIA	ASIA
RIPARTIZIONE	ASIA
ADDETTI	ASIA
DIPENDENTI	ASIA
ATECO	ASIA
ETÀ D'IMPRESA	ASIA
FATTURATO	Frame
VALORE AGGIUNTO	Frame
NUMERO COMMITTENTI	CU
FATTURATO B2B	CU
FATTURATO 5 COMMITTENTI PRINCIPALI	CU
CODICE FISCALE COMMITTENTE PRINCIPALE	CU
AMPIEZZA NETWORK COMMERCIALE	CU
ATECO COMMITTENTE PRINCIPALE	ASIA
ADDETTI COMMITTENTE PRINCIPALE	ASIA
FATTURATO COMMITTENTE PRINCIPALE	ASIA

2.2 Record linkage

Effettuata una ricognizione delle fonti necessarie, si è provveduto alla integrazione fisica delle informazioni (record linkage). Il record linkage è un processo di integrazione di dati provenienti da fonti diverse e mira ad individuare record di informazioni riferiti alle medesime unità statistiche, ma residenti fisicamente in archivi diversi (integrazione di fonti), gestendo la presenza di duplicati (de-duplicazione). L'identificazione dell'unità in archivi di diversa natura (amministrativa e statistica) avviene attraverso chiavi comuni, presenti nei vari file (le chiavi possono essere anche essere solo parzialmente coincidenti). Ne segue che la complessità del problema di record linkage, pur dipendendo da molteplici aspetti, è principalmente legata alla qualità degli identificatori su cui vengono ordinate le operazioni di *join*, da cui dipende sia la corretta integrazione dei dati, sia la possibilità della presenza di record doppi (consistenza dei dati garantita da *unique constraints* sugli identificatori). In estrema sintesi dalla correttezza degli identificatori acquisiti dipendono l'assenza di univocità o la presenza di errori che impedisce l'integrazione fisica dell'informazione in un unico record.

L'uso di tecniche di record linkage nei vari processi di produzione è ormai diffuso da diversi anni e consente la creazione di basi di dati in grado di studiare a livello di microdato (individuo, famiglia, impresa ecc.) lo studio dei profili associativi (studio delle distribuzioni congiunte) di variabili che costituiscono fenomeni multivariati complessi oggetto di analisi più disparati, sia cross section che longitudinale. Molteplici i campi di applicazione, che interessano ormai ogni tipo di studio applicativo (ricerca socio-economica e

demografica, marketing ecc.). Il record linkage è un processo complesso a causa dei numerosi aspetti di natura diversa che lo compongono. Se negli archivi da abbinare sono presenti identificatori univoci allora il problema non ha una grande complessità; in generale però, per analizzare dati privi di identificatori univoci o con identificatori univoci affetti da errore, sono richieste sofisticate procedure statistiche; soluzioni informatiche non banali sono necessarie per gestire e trattare grandi moli di dati, mentre i vincoli legati al tipo di applicazione che si intende effettuare possono comportare la soluzione di complessi problemi di programmazione lineare.

In linea di principio, è possibile individuare le seguenti come le principali fasi che costituiscono un processo di record linkage:

1. Preparazione dei dati di input (pre-processing);
2. Selezione degli attributi identificativi comuni (variabili di matching);
3. Scelta della funzione di confronto;
4. Riduzione dello spazio di ricerca delle coppie candidate all'abbinamento;
5. Scelta del modello di decisione;
6. Selezione degli abbinamenti univoci;
7. Valutazione dei risultati del record linkage e validazione delle stime.

Ciascuna fase è perciò descritta così come realizzata nel presente processo di integrazione.

Fasi 1 e 2. Nel caso in esame, l'integrazione fisica è stata svolta sugli identificativi degli individui, opportunamente anonimizzati al fine di garantire l'assenza di riconoscibilità dei singoli individui. Si tratta perciò di identificativi anonimi, collegati in maniera univoca al codice fiscale degli individui coinvolti. La consistenza dei dati dalle principali fonti di input è una qualità che caratterizza gli insiemi di partenza. Di seguito una rapida analisi di ciascuna di esse. Nel caso di tutte le fonti statistiche derivate da archivi amministrativi, i dati sono stati trattati in maniera da ovviare ai problemi di sovra-sotto copertura: tutti i dati presentano perciò univocità di individui (righe) e rappresentano in maniera esaustiva l'universo di riferimento. Il dati identificativi sono rappresentati da un codice anonimizzato, collegato al codice fiscale, assegnato attraverso pretrattamento in ISTAT all'interno del sistema di integrazione dei dati amministrativi (SIM)⁶. Il SIM supporta trasversalmente i processi di produzione dell'istituto, essendo il *repository* dei dati amministrativi acquisiti dall'Istat, in cui i dati vengono mappati e trattati al fine di evitare fenomeni di duplicazione, fisica o semantica. Il medesimo trattamento coinvolge anche le fonti che popolano la BIT e i suoi output. L'archivio CU è di fonte fiscale, ossia Agenzia delle Entrate che da diversi anni fiscali controlla la consistenza dei codici fiscali inseriti nei dichiarativi con quelli presenti in Anagrafe Tributaria, provvedendo alla eliminazione delle comunicazioni con codici fiscali non presenti/invalidi, costringendo perciò i contribuenti al reinvio dei dichiarativi fino a quando non siano formalmente corretti. Il database di produzione prevede l'acquisizione in una forma verticalizzata, dato che la natura dell'informazione fiscale si presenta tipicamente sotto forma di matrici "sparse". La qualità dei database di input prevede delle operazioni di pre-processing limitate, in particolare lo scarto delle dichiarazioni sostituite da successive inviate dai contribuenti e loro sistemazione in tabelle di dati orizzontali (del tipo unità x variabili). La vera difficoltà è di natura computazionale, rappresentata dalla particolare forma delle CU, fornite in una tabella verticalizzata (i dati fiscali sono sempre raccolti in tale forma a causa della loro natura di matrici estremamente sparse). Le *query* su un insieme estremamente importante in termini di cardinalità (diversi miliardi di record) prevede modalità di interrogazione basate su particolari indici partizionati al fine di creare tabelle idonee all'uso statistico dell'informazione (individui x variabili).

La selezione delle variabili di matching si esaurisce nella scelta dell'identificativo anonimo legato al codice fiscale, dato che questa variabile selezionata consente di identificare univocamente le unità della popolazione di interesse. Si può ritenere inoltre, per quanto detto in precedenza, che essa sia (relativamente) non affetta da errori, mancate risposte, mancanza di stabilità nel corso del tempo (National Statistics 2004a, 2004b); i problemi legati alla privacy sono gestiti grazie al processo di anonimizzazione descritto.

⁶ I dati vengono considerati micro poiché si riferiscono alle unità: Individui, Unità economiche, Luoghi, ovvero le unità di base della statistica ufficiale. L'aggettivo integrato si riferisce al processo di integrazione ed in particolare all'integrazione delle unità di base. Oltre alle unità di base, che allo stato attuale comprendono le tre tipologie di Individui, Unità economiche e Luoghi, anche le relazioni tra le unità dello stesso tipo o di diverso tipo costituiscono uno specifico interesse per l'analisi dei fenomeni statistici.

Fasi 3, 4, 5 e 6. Data la tipologia e qualità dei dati descritti (che ha già assolto in via autonoma il *pre-processing* dei dati), la funzione di distanza prescelta è un modello deterministico esatto, senza riduzione dello spazio di ricerca delle coppie candidate all'abbinamento, dato che i dataset di input sono già univoci negli individui, ossia le righe dei file da integrare (riduzione 1:1 per modello deterministico già presente nei dati). Il portato delle condizioni precedenti è che gli abbinamenti sono tutti univoci, per cui la valutazione, seppure effettuata, è pleonastica.

Fase 7. Dato che gli abbinamenti effettuati sono univoci, i maggiori problemi da valutare ai fini della qualità riguardano i mancati abbinamenti e le discrasie fra le informazioni contenute nelle due fonti. Rispetto alle variabili contenute negli archivi o nei data set estratti da archivi acquisiti centralmente, deve essere condotta in primo luogo un'analisi concettuale sulla corrispondenza con le variabili obiettivo statistico, per prevenire errori di specificazione. Per quanto riguarda la coerenza concettuale, esiste una quasi totale corrispondenza concettuale fra le variabili obiettivo del presente studio e variabili amministrative da CU (e segnatamente la copertura della fonte rispetto all'universo e la corrispondenza delle variabili qualificanti il rapporto di lavoro in termini di capitale umano). Non sono perciò evidenziate particolari problematiche in questo senso, essendo il maggiore problema teorico costituito dalla eventuale presenza di discrasia fra condizioni lavorative dichiarata e amministrativa (errori nei dati amministrativi, nelle dichiarazioni rese dall'intervistato, eventualmente riconducibile a lavoro nero con relativa problematica di misurazione). Il focus si sposta perciò direttamente sul calcolo delle misurazioni di qualità sui dati, ad esempio l'entità dei dati mancanti nelle variabili di interesse. Qui la qualità del dato amministrativo è altissima, possedendo le variabili usate un ruolo centrale nel processo amministrativo ed essendo perciò sottoposte a validazione durante il processo acquisitivo. Non risultano allo stato dei fatti dati mancanti. Nel campione in esame, una ampia percentuale di popolazione mostrava coerenza fra variabile statistica rilevata e la corrispondente variabile amministrativa. Nel caso dei dati sulla fornitura, non esiste allo stato attuale la possibilità di verificare altrimenti i dati relazionali (i rapporti commerciali di committenza) fra le imprese. Almeno fino a quando non saranno disponibili i dati relativi alle fatture emesse/ricevute (cosiddetto "Spesometro"), l'unico modo di verificarne la correttezza è ricreare delle variabili di posizione (il fatturato complessivo e le sue porzioni B2B e B2C) e confrontarlo con i dati desunti dalle dichiarazioni IVA e dalle altre fonti strutturali. Sono le coerenze di cui si diceva in precedenza; altro elemento di grande affidabilità di questi dati è la natura di redditi "certificati", che vengono utilizzati per la compilazione delle denunce dei redditi: ne segue che tutte le variabili utilizzate, che hanno una importanza capitale nel volume delle imposte da pagare a conguaglio, sono sicuramente fra le variabili *core* della fonte, soggette a controllo incrociato fra le due parti coinvolte nella certificazione dei redditi.

Si può assumere perciò con ragionevole sicurezza sia (1) la totale copertura della popolazione statistica obiettivo da parte delle fonti amministrative utilizzate, come (2) la piena coincidenza fra la variabile amministrativa e variabile statistica di interesse (punto B.9 del manuale sulle Linee guida per la qualità dei processi statistici che utilizzano dati amministrativi; ISTAT, 2016). Si sono perciò valutate le coerenze inter-fonte fra variabili, verificando le coerenze definitorie e sciogliendo le discrasie secondo (1) un concetto di gerarchia fra le fonti e (2) verificando che le stime ottenute dall'integrazione fossero coerenti con i dati di altre fonti amministrative e/o con fonti statistiche interne o esterne all'Istituto, ovvero la coerenza dei risultati rispetto a rapporti che possono essere considerati pressoché costanti. La prevalenza in termini di gerarchia è stata assegnata alle fonti statistiche ISTAT per ovvi motivi di pulizia e coerenza dei dati.

2.3 Integrazione dati: Output quality

Una volta realizzato un processo di integrazione dati coerente con i principi e le linee guida da seguire al fine di produrre, in modo efficiente, statistiche di elevata qualità, è necessario misurare la qualità del risultato ottenuto. Ai fini della misurazione della qualità delle statistiche, l'Istat ha adottato la definizione della qualità originariamente rilasciata da Eurostat nel 2004 e recepita dalla nostra legislazione (Gazzetta Ufficiale 13 ottobre 2010, n. 240). La qualità delle statistiche prodotte e diffuse viene definita come "il complesso delle caratteristiche di un prodotto o di un servizio che gli conferiscono la capacità di soddisfare i bisogni impliciti o espressi" deve essere valutata secondo i seguenti 9 criteri.

1. **Pertinenza.** Il grado con cui le esigenze degli utenti sono soddisfatte in termini di completezza dell'informazione prodotta.
2. **Accuratezza.** Il grado di vicinanza tra la stima e il valore vero che la statistica intende misurare.

3. Attendibilità. La vicinanza del valore della stima iniziale e i successivi valori rilasciati sulla stessa stima.
4. Tempestività. E' il periodo di tempo che intercorre tra l'evento o il fenomeno che i risultati descrivono e il momento in cui gli stessi vengono resi disponibili.
5. Puntualità. Il periodo di tempo tra la data effettiva del rilascio dei dati e quella pianificata.
6. Accessibilità. La facilità di ottenimento dei dati.
7. Chiarezza. La facilità con cui gli utenti vengono messi in grado di capire i dati.
8. Comparabilità (nel tempo e nello spazio). Misura quanto le differenze nel tempo e tra aree geografiche siano dovute a variazioni reali e non a differenze di natura strumentale (concetti statistici, strumenti di misurazione e procedure).
9. Coerenza. Misura l'adequatezza delle statistiche ad essere combinate in una molteplicità di fini.

La misurazione quantitativa diretta può essere effettuata solo per alcuni criteri (tempestività e comparabilità). Per le altre dimensioni spesso è solo possibile esprimere delle valutazioni di natura qualitativa. "L'approccio alla misurazione della qualità risulta perciò in un compromesso, in cui alle poche misurazioni dirette si affiancano delle misurazioni indirette" (ISTAT, 2016). Gran parte dei requisiti sopra esposti sono maggiormente calzanti con esigenze di statistica ufficiale e solo parzialmente cogenti con le esigenze di una pubblicazione scientifica come la presente. Problemi legati a criteri come pertinenza, tempestività, comparabilità, puntualità, accessibilità sono laterali nel nostro contesto.

Altri problemi sono comunque rilevanti, ma difficilmente calcolabili. Il principale è costituito dall'accuratezza delle stime (vicinanza tra la stima e il valore "vero"). Da un punto di vista quantitativo, misure di accuratezza possono essere ottenute stimando l'errore attraverso metodi di "bootstrap re-sampling", utilizzabile in situazioni di integrazione micro e macro (WP 6 ESSnet "Use of Administrative and Accounts Data in Business Statistics"; Deliverable 6.3., 2011). In questo caso timidi tentativi sono stati posti in essere per definire una misura del *mean square error* delle stime, non giunti però ad un livello tale da essere presentato. Riteniamo però che questa possa essere la strada più facilmente percorribile per avere una misura dell'Errore Quadratico Medio, ossia dell'accuratezza delle stime.

Laddove non sia possibile definire precisamente una misura dell'errore quadratico medio, una valutazione della qualità dei risultati può essere effettuata analizzando le diverse componenti di errore che impattano sull'accuratezza delle stime. Seguendo quanto viene abitualmente per statistiche di indagine, verranno considerate le possibili componenti di errore legate alle (1) unità e alle (2) variabili.

2.3.1 Errori sulle unità o popolazione

Non corrispondenza concettuale tra popolazione statistica e popolazione amministrativa

Il problema della non corrispondenza tra popolazione statistica e amministrativa è stato affrontato in precedenza trattando dell'integrazione fisica e armonizzazione dei dati da confronto inter-fonte. Sussistono una serie di rilievi che possono essere avanzati, il maggiore dei quali è costituito dal volume delle relazioni e dagli importi delle transazioni: ma di fatto è necessario aspettare il possesso di altre fonti per sciogliere in maniera definitiva il nodo. Rimane la certezza che la natura di redditi certificati che posseggono le CU, è una assicurazione forte sulla qualità dei dati.

Errori di selezione

Gli errori di selezione hanno origine dalla discrepanza tra l'insieme dei dati di interesse accessibile (o osservabile) e quello concretamente acquisito. In questo caso il problema potrebbe essere rappresentato da una incompleta acquisizione; nel caso delle CU l'integrazione con altri dati fiscali e statistici serve a valutare (anche) i problemi di sotto-sovracopertura della fonte (gli errori di selezioni si concretizzano in dati mancanti e duplicazioni, magari legati a ritardi o duplicazioni nella trasmissione dei dati).

Errori di linkage o errore di abbinamento

Nell'integrare archivi diversi di dati amministrativi tra di loro e con dati di indagine, i principali errori che si commettono sono i) falsi non abbinamenti e ii) falsi abbinamenti. È evidente che la qualità dei risultati dell'abbinamento dipende fortemente dalla qualità delle chiavi di abbinamento utilizzate. La qualità delle chiavi di linkage (il codice identificativo anonimo delle imprese) è valutata di altissima qualità per le considerazioni già esposte, ancorché non esente da errori.

2.3.2 Errori sulle variabili

Errori di specificazione

Così come nel caso di indagini dirette gli errori di specificazione derivano dalla non corrispondenza tra obiettivi conoscitivi di indagine e concetti rilevati attraverso i quesiti del questionario; nell'ambito dell'utilizzo dei dati amministrativi, questi errori sono riferiti a discrepanze tra il concetto obiettivo teorico statistico e quello amministrativo. È forse il tipo di errore che ha il maggior impatto sulla pertinenza delle statistiche prodotte e può causare errori di accuratezza, in particolare sulla componente della distorsione. Gli errori di specificazione sono stati perciò gestiti con estrema attenzione essendo la tipologia di errore con il maggior impatto in termini di pertinenza e distorsione delle statistiche prodotte. Il fine è, naturalmente, che esista la massima corrispondenza fra obiettivi conoscitivi e concetti rilevati attraverso confronti con i dati amministrativi e soluzione delle incoerenze per gerarchia di fonte, ponendo grande enfasi sul trattamento automatizzato dei dati al fine di prevenire errori materiali di derivazione (vedi oltre, errori di processo e classificazione).

Errori di misurazione

Gli errori di osservazione possono verificarsi nella fase di raccolta (errori di misurazione in senso stretto). Anche questo tipo di errore è stato gestito con l'integrazione e confronto con i dati delle altre fonti al fine di verificare la coincidenza delle informazioni raccolte.

Errori di processo e classificazione

Errori di processo e di classificazione sono stati affrontati attraverso la gestione automatizzata dei dati, al fine di evitare errori materiali nel trattamento dei dati o di errata classificazione (ricorso a metadati organizzati in forma di schema relazionale in Oracle).

L'insieme delle fasi di lavoro descritte in questi paragrafi, coerenti con le attuali *best practices*, ha consentito di delineare un processo di integrazione dati, idealmente strutturato in tre parti (la qualità dell'input, del processo produttivo e dell'output). Risultato finale del processo è stata la produzione di un dataset integrato, valutato e validato seguendo i criteri che rappresentano degli standard per base di dati statistiche e amministrative.

3 Le classi di variabili individuate: imprenditorialità e capitale umano

Una delle caratterizzazioni delle imprese individuali oggetto di studio è il raggruppamento in tre grandi classi di lavoratori autonomi, sostanzialmente dipendente dal settore d'attività economica d'appartenenza: commercianti, professionisti (privi o meno di albo/ordine) ed imprese. Ed ognuna di queste categorie può avere o meno la proprietà di essere artigiana. Risulta immediato percepire come queste categorie, nella realtà imprenditoriale italiana, rappresentino contenitori caratterizzati da notevole eterogeneità. La linea di demarcazione che si cerca di aggiungere alle sottopopolazioni sopra individuate è l'idea di un *continuum* autoimpiego-piena imprenditorialità che caratterizza i lavoratori autonomi come figure intermedie fra i lavoratori subordinati e le imprese propriamente dette. Dotati infatti di una imprenditorialità, avvalendosi o meno di altre risorse umane per lo svolgimento delle proprie attività professionali, i lavoratori autonomi individuali possono essere efficacemente rappresentati da un set di variabili così classificate:

- A. capitale umano, la cui dotazione influisce sulla produttività e che può essere espressa in termini analoghi a teorie e prassi che formalizzano la categoria dei lavoratori dipendenti;
- B. imprenditorialità, che caratterizza invece i lavoratori autonomi in maniera più simile alle imprese e che può essere definita facendo ricorso ad approcci consolidati dal punto di vista fiscale (segnatamente gli Studi di Settore dell'Agenzia delle Entrate, che individuano una serie di variabili ritenute direttamente correlate con la possibilità di generare maggiori utili).

Tale classificazione consente di definire analiticamente il potenziale produttivo e di esprimere sinteticamente l'autonomia aziendale tramite un continuum di valori che vede ai suoi estremi il lavoratore quasi-integrato verticalmente in altra azienda committente (cosiddetta "finta Partita IVA") oppure il lavoratore davvero dotato di autonomia e quindi compiutamente impresa. Le variabili che aiutano a caratterizzare nei due sensi il lavoratore autonomo sono contenute nello schema sottostante (Tabella 2).

Tabella 2 – Profilo degli indipendenti: variabili di imprenditorialità e capitale umano

AREA	
IMPRENDITORIALITÀ	CAPITALE UMANO
1. Età dell'impresa	8. Attività svolta (Settore ATECO)
2. Territorio	9. Titolo di Studio (più elevato conseguito)
3. Committenza	10. Età
4. Internazionalizzazione	11. Sesso
5. Volume di Fatturato	12. Nazionalità
6. Dinamismo economico territoriale	
7. <i>Network</i> commerciale	

Gli elementi di imprenditorialità sono caratteristiche d'impresa ritenute influenti nel definire la capacità di produrre ricavi e quindi valore aggiunto. In particolare:

1. Età dell'impresa: variabile proxy importante nel definire la capacità di produrre fatturato e reddito da parte dell'impresa, perché correlata con la possibilità di consolidare nel tempo conoscenze, relazioni di fornitura/clientela e quindi sulla possibilità di fare business.
2. Territorio di principale attività: la localizzazione territoriale delle imprese influisce sulla capacità di produrre ricavi e coglie le differenze qualitative tra i comuni in termini di sviluppo socio-economico a prescindere dalla loro dimensione geografica e/o demografica.
3. Committenza: tipologia e apporti rispetto al profitto complessivo d'impresa.
4. Internazionalizzazione: è la capacità di esportare e segnala una maggiore organizzazione ed efficienza aziendale (necessarie ad affrontare mercati esteri e disciplinati in maniera differente), nonché un livello di qualità di prodotto/servizio superiore che si accompagna solitamente a una maggiore produttività/redditività aziendale.
5. Volume di fatturato: variabile di scala che individua significative differenze a livello organizzativo e produttivo
6. Dinamismo economico territoriale.
7. *Network* commerciale: considerata una rete commerciale di un'impresa composta da se stessa, le imprese committenti e da tutti le imprese ad esse connesse tramite un rapporto di committenza, l'ampiezza della rete è misurata come il numero minimo di legami tra l'impresa allo studio e l'impresa committente più distante (Figura 1).

Le variabili del secondo gruppo riguardano invece caratteristiche tradizionalmente sfruttate nelle analisi dei differenziali retributivi sul mercato del lavoro che individuano variabili di stock di capitale umano. Nello specifico:

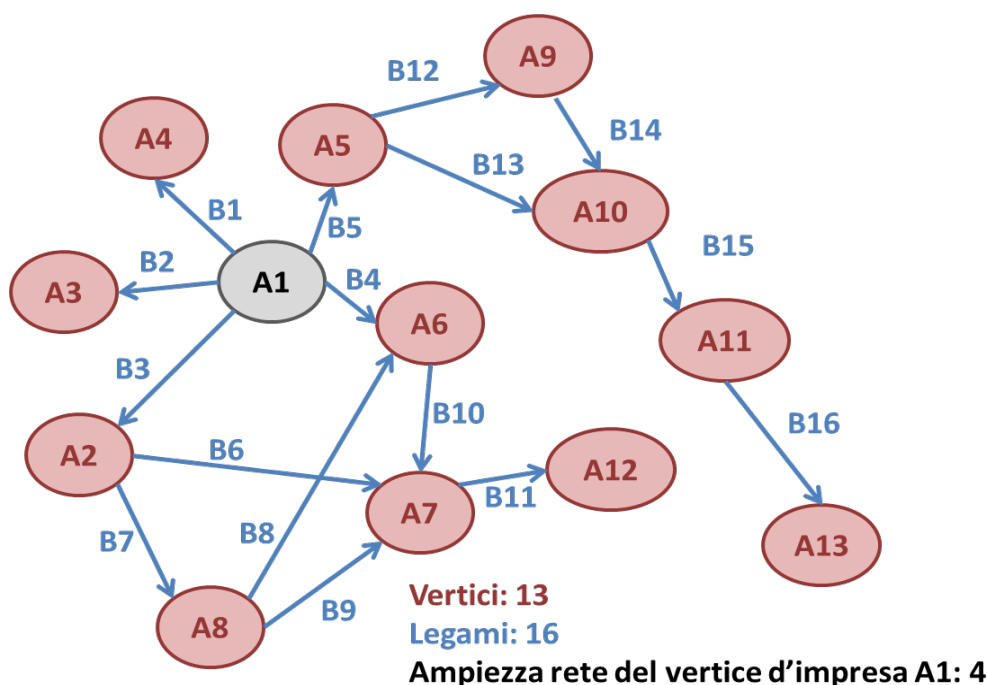
8. Attività svolta (ATECO): raccordata alla classificazione delle professioni rilasciata dall'International Labour Organization (ILO), consente di definire il livello di skill espresso da ciascun lavoratore autonomo.
9. Titolo di studio più elevato conseguito: variabile chiave a livello imprenditoriale e produttivo, rappresenta un dato fondamentale in termini di educazione formale all'interno dell'impresa.
10. Età anagrafica del lavoratore individuale: si ritiene una variabile proxy della complessiva esperienza lavorativa accumulata dal lavoratore durante tutta la sua carriera lavorativa.
- 11./12. Le ultime due variabili, prettamente socio-demografiche, non rappresentano fattori direttamente connesse alla produttività, bensì individuano sottopopolazioni tipicamente rilevanti in termini sociali e di disparità retributive.

L'insieme di queste variabili rappresenta, ad avviso degli scriventi, una proposta analitica in grado di rappresentare in maniera adeguata il fenomeno del lavoro autonomo, peculiarità tutta italiana che dovrebbe trovare una adeguata rappresentazione come fenomeno socio-economico: è di fatto una ampissima categoria dove sono presenti unità economiche caratterizzate da fortissima eterogeneità, rappresentabili come un *continuum* logico che spazia dall'autoimpiego a realtà aziendali evolutissime.

4 Analisi di network

Uno dei grandi *asset* informativi della base dati realizzata è da individuarsi nella struttura delle relazioni di committenza che lega le imprese individuali. Tali soggetti sono infatti sottoposti a un tipologia di ritenuta di acconto sui redditi professionali, certificata poi nelle CU, che consente di ricreare la rete di relazioni *trade*, ossia di natura commerciali, intervenuta con tutti gli altri operatori economici. Segnatamente si contano oltre 17 milioni di relazioni commerciali, che scendono a 10 milioni circa se ci si concentra sulla sottopopolazione obiettivo (l'insieme delle relazioni commerciali intervenute fra le imprese individuali e tutti gli altri soggetti IVA, pubblici o privati). Questo insieme di relazioni è per sua natura rappresentabile attraverso la teoria matematica dei grafi.

Figura 2 – Rappresentazione grafica del network di relazioni trade



L'analisi delle reti cliente/fornitore tramite la teoria matematica dei grafi ha un duplice scopo: (a) ereditare le definizioni e proprietà già note per alcune classi di problemi grafali (es. problemi di flusso) e (b) disporre di una valutazione di complessità degli algoritmi basata su una solida letteratura. Come in ogni attività di ricerca in cui si mettono insieme contesti eterogenei (rif. reti del valore vs. grafi), è necessario valutare quali sono le relazioni di equivalenza tra i concetti di base dei due approcci. In questo lavoro, riteniamo che un concetto di base della teoria dei grafi, utile nella nostra analisi delle reti cliente/fornitore, sia quello di componente connessa. Per componente connessa si indica l'insieme di tutti e soli gli elementi del grafo (vertici) che sono connessi (i.e. collegati da un cammino di soli elementi della componente) e che non hanno alcun legame con il resto della rete. Una componente connessa rappresenta nella rete del valore l'insieme delle coppie (cliente, fornitore) che caratterizzano una economia locale e/o nazionale a seconda della localizzazione delle singole aziende. La ricerca delle componenti connesse è un problema facile, secondo la definizione data in teoria della complessità, e quindi gli algoritmi presenti in letteratura ben si prestano all'analisi di reti aventi milioni di nodi (i.e. clienti e fornitori). E' anche vero che le reti da analizzare risentono della presenza di legami non sempre necessari e quindi si ritiene che si possa avere bisogno di tecniche che semplifichino le reti considerate, facendo emergere solo i legami economici rilevanti per l'analisi considerata e quindi definire meglio le componenti connesse implicite nella rete. In termini grafali,

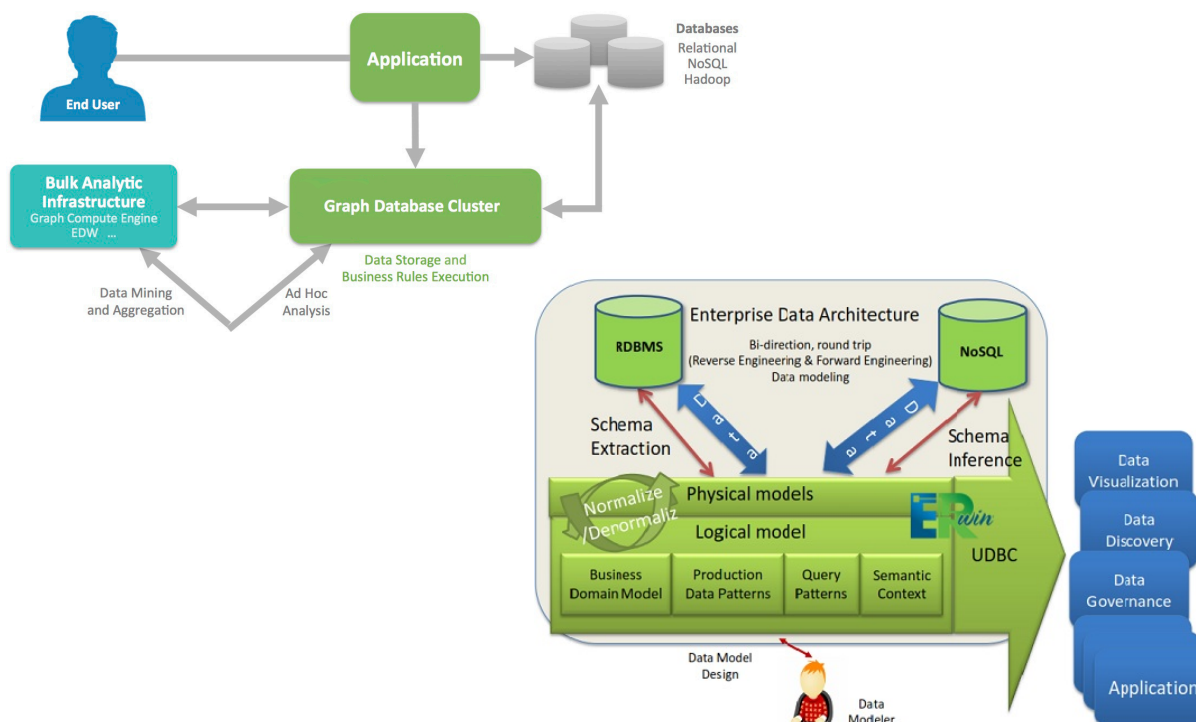
possiamo dire che a partire da una rete G otteniamo una rete G' con lo stesso insieme di nodi ma con un sottoinsieme di spigoli. Si osserva che più piccolo è l'insieme degli spigoli risultante e maggiore la frammentazione del grafo G' in componenti connesse. Un indice di una buona frammentazione è dato dalla percentuale di nodi isolati su l'insieme totale. Questa percentuale può avere diverse chiavi di lettura: valori bassi indicano sia una insufficiente cancellazione di spigoli e sia una rete fortemente connessa che non si riesce a separare con facilità. Analogamente, valori alti indicano sia una eccessiva cancellazione di spigoli ovvero che la rete considerata rappresenta legami economici deboli. Il punto di equilibrio tra questi due aspetti implica significative considerazioni economiche ed un approccio algoritmico iterativo in cui si arriva alla configurazione ottimale per approssimazioni successive.

Una parte importante del lavoro è quindi costituita dal passare dalle coppie di relazioni suddette, a una struttura dei dati organizzata in nodi (imprese) e grafi orientati (le relazioni commerciali) che legano le imprese. In altre parole, una parte importantissima del lavoro è la costruzione da una struttura di database opportunamente progettata per gestire cardinalità importanti di dati, dalla natura fortemente poliedrica. E' necessario definire basi di dati che risultino altrettanto flessibili, pur garantendo scalabilità e prestazioni in presenza di grosse moli di dati e strutture pesanti da interrogare attraverso le normali modalità. La modalità prescelta è l'integrazione fra database SQL e noSQL.

L'affermazione ed evoluzione di soluzioni Big Data sta trasformando il mondo delle tecnologie di database. Disponibilità di dati e sviluppo di piattaforme in grado di gestire queste moli di dati, rendono implicita la convergenza tra mondo relazionale e non relazionale, tra piattaforme SQL e noSQL ed è questa la scommessa su cui si gioca la *leadership* futura nel mondo delle piattaforme di database. Nel caso in esame si sono esaminate le implementazioni pratiche di strategie di integrazione di database SQL/noSQL proposte da Oracle, basata sulla disponibilità di connettori che abilitano la capacità di lavorare con linguaggi noSQL e ambienti Hadoop. In moltissimi casi, il modello relazionale rimane l'opzione di carattere più generale, particolarmente utile nella misura in cui la struttura rigida che DB SQL richiedono non è necessariamente un problema, ma costituisce un vantaggio: è questo il caso in cui si devono modellare dati molto strutturati, come quelli normalmente gestiti negli registri ISTAT o proveniente da fonte fiscale. In questi casi, la duttilità del NoSQL non appare un requisito fondamentale; la gestione di questi dati avverrebbe attraverso database OLAP (On-Line Analytical Processing), strutturate in tradizionali star / snowflake schemas, opportunamente indicizzati e strutturati attraverso viste materializzate in grado di ottimizzare l'analisi interattiva e veloce di grandi quantità di dati. La costruzione della base di dati prevede la gestione della dimensione temporale (dataset versione *long*) e lo sviluppo degli oggetti di analisi (proiezione dei dati di basi) attraverso *materialized views*, abilitate al *query rewrite*, al fine di garantire prestazioni (scalabilità) al sistema.

Laddove invece esista la presenza di dati strutturati, ma risultino molto collegati tra loro – situazione aggravata da una loro gran quantità – le consuete forme di interrogazione SQL rischiano di non essere lo strumento ottimale: sarebbe da prediligere la navigazione tra oggetti sfruttando i riferimenti tra i vari nodi di informazione (nodi e grafi orientati). La struttura a grafo si mostra estremamente comoda ed efficiente nel trattare strutture dati che possono essere rappresentate con naturalezza da un grafo poiché sono esse stesse dei grafi (ad esempio file XML o strutture di reti come questa in esame). L'esplorazione di queste strutture risulta in genere più veloce rispetto a un database a tabelle dato che la ricerca di relazioni fra nodi è un'operazione primitiva e non richiede molteplici join innestate su tabelle diverse. Ogni nodo contiene l'indice delle relazioni entranti e uscenti da esso, quindi la velocità di attraversamento del grafo non risente delle dimensioni complessive del database (tabelle) ma solo della densità dei nodi attraversati. Esistono inoltre delle implementazioni già pronte per le operazioni più comuni sui grafi (la ricerca del cammino minimo tra nodi, il calcolo del diametro della rete ecc.), nonché la possibilità di far collassare le informazioni in tabelle interrogabili tramite *query* SQL in ricerche complesse (per esempio basate su confronti matematici tra i campi delle *tuple*). L'utilizzo di appositi database a grafi come NEO4J e Apache Cassandra, integrati con database Oracle è la base necessaria per studiare rete di relazioni, più o meno estese e strutturate.

Figura 3 – Integrazione Database SQL e noSQL: gestione dati e possibilità di analisi



Una volta definito il database SQL-noSQL è stato possibile utilizzare le possibilità informative di una simile base dati integrata. Le possibilità informative sono davvero ampie e in questa sede appena sfruttate. Trattasi di tutte le analisi conosciute in letteratura sotto il nome di network analysis. Lo scopo principale dell'analisi di network è quello di individuare e analizzare legami (*ties*) tra gli individui (*nodes*) ed evidenziare le proprietà di rete nel loro complesso: semplicemente la presenza di nodi isolati e la cardinalità della rete, la distanza massima della rete di appartenenza, il tasso di saturazione dei legami in una rete e la sua conseguente struttura; coesione, centralità, la ricerca di sottoreti specifiche (gruppi, egonet) o di somiglianze fra reti (equivalenza strutturale, automorfica e regolare). Le applicazioni sono davvero infinite e in termini futuri potrebbero qualificare notevolmente l'impresa in termini di appartenenza alle catene del valore locali/globali e il suo ruolo con problemi di sviluppo locali che, allo stato attuale sono appena abbozzati.

Nel caso presente ci si è limitati a una piccola applicazione, mutuando una logica da social media. Si sono dapprima definiti i clienti diretti di ogni impresa e calcolato il peso del cliente più importante in termini di fatturato, così come dei primi 5 committenti: questi dati sono fondamentali in termini di analisi della dipendenza. Si è poi declinata la rete *trade* in termini analoghi a quanto avviene nei social media con amici, amici degli amici, amici degli amici degli amici ecc.: in questo caso trattasi però dei già citati clienti diretti; dei clienti dei clienti, i clienti dei clienti diretti di ciascuna impresa; da lì i clienti dei clienti dei clienti e così via. Si è poi testato econometricamente il ruolo sulla produttività di questi livelli di relazioni in cui è inserita l'impresa, così come mostrato nel successivo paragrafo 5.2.

5 Analisi dei dati

Come anticipato nell'introduzione, è stata svolta una descrizione delle imprese individuali, con riguardo alla posizione settoriale e territoriale: a livello uni e bivariato, sia a livello di analisi multivariata. Infine, attraverso una modellazione della produttività è stata sintetizzata l'analisi svolta con i precedenti passi al fine di mettere in luce come le variabili proposte a descrivere un profilo dell'impresa individuale (segnatamente come l'ubicazione territoriale e il sistema di relazioni commerciali in cui è inserita, influiscano sulla profittabilità e sulla capacità imprenditoriale).

5.1 Analisi descrittiva semplice e multivariata

Semplici statistiche descrittive sono state svolte su ognuna delle variabili ottenute dal data set integrato. Le più significative riguardano senz'altro l'incidenza intesa come frequenza relativa del numero delle imprese e del valore aggiunto sul totale nazionale.

Tabella 3 – Partite Iva individuali per categoria di lavoro autonomo, ripartizione geografica e anno. Val. assoluti in migliaia

Anno 2015				
Ripartizione	Commerciante	Professionista	Impresa	TOTALE
Nord-ovest	257	216	314	788
Nord-est	196	140	229	564
Centro	202	164	218	584
Sud e isole	258	132	194	584
ITALIA	1.026	705	1.049	2.779

Anno 2016				
Ripartizione	Commerciante	Professionista	Impresa	TOTALE
Nord-ovest	258	224	314	796
Nord-est	196	144	228	568
Centro	202	169	217	588
Sud e isole	259	135	195	589
ITALIA	1.028	728	1.048	2.804

Altre statistiche di una certa rilevanza riguardano la distribuzione territoriale di queste imprese (Tabella 3), così come la produttività apparente del lavoro che tali imprese esprimono (Tabella 4).

Tabella 4 – Valore aggiunto medio per addetto per ripartizione geografica e anno delle Partite Iva individuali distinte per categoria di lavoro autonomo

Anno 2015				
Ripartizione	Commerciante	Professionista	Impresa	TOTALE
Nord-ovest	19.508	36.106	28.002	26.822
Nord-est	20.772	33.523	27.915	26.300
Centro	17.430	31.303	24.097	23.228
Sud e isole	15.062	25.105	21.137	19.047
ITALIA	17.767	31.633	25.231	23.581

Anno 2016				
Ripartizione	Commerciante	Professionista	Impresa	TOTALE
Nord-ovest	19.753	35.884	28.323	27.032
Nord-est	21.220	33.290	28.601	26.744
Centro	17.652	31.011	24.258	23.349
Sud e isole	15.391	24.483	21.135	19.098
ITALIA	18.070	31.287	25.493	23.761

Anche le altre variabili mostrano un deciso interesse, in specie l'aspetto occupazionale, quello relazionale e infine la collocazione sul territorio. Si nota in particolare che la stragrande maggioranza di queste imprese, oltre l'80%, sono costituite da un singolo imprenditore. Minoranza sono quindi i casi di realtà aziendali in cui il titolare è accompagnato da coadiuvanti o dipendenti.

Tabella 5 – Numero medio di addetti d'impresa per ripartizione geografica delle Partite Iva individuali distinte per categoria di lavoro autonomo

Anno 2016				
Ripartizione	Commerciante	Professionista	Impresa	TOTALE
Nord-ovest	1,53	1,13	1,58	1,44
Nord-est	1,63	1,13	1,65	1,51
Centro	1,53	1,11	1,65	1,45
Sud e isole	1,55	1,11	1,69	1,50
ITALIA	1,56	1,12	1,64	1,47

Oltre la metà delle aziende italiane nel suo complesso è quindi costituita da micro imprese con un solo titolare. La domanda spontanea è se l'aspetto dimensionale è contemperato dall'aspetto relazionale: come a dire che quello che non è assolto dalla dimensione interna (la scala dimensionale, l'occupazione) è svolto dalla dimensione esterna, ovvero dalla cooperazione fra imprese. Anche in questo caso la risposta è negativa, infatti la maggior parte delle imprese individuali costituisce un nodo isolato (55,7% del totale), per cui opera in totale solitudine. Per la maggior parte di queste imprese non esiste il supporto di altri soggetti economici: ne segue un severo limite alla produttività e alle possibilità di business che non può seguire via d'espansione né interne né esterne. Sono caratteristiche comuni a un numero elevatissimo di imprese, che si delineano quindi come soggetti economici dotati di una forte marginalità (più sulla polarità autoimpiego che imprenditorialità, nei termini espressi nel paragrafo 3). Il carattere di isolamento presenta caratteristiche diverse in base a *pattern* di specificità settoriale. Le imprese sono le più isolate, seguite molto da vicino dai commercianti. I professionisti rappresentano invece la tipologia di impresa maggiormente connessa, per ovvie specificità lavorative che riguardano da vicino il tipo di servizio prestato.

Tabella 6 – Numero medio di addetti d'impresa per ripartizione geografica delle Partite Iva individuali distinte per categoria di lavoro autonomo

TIPO IMPRESA	%		P25	P50	Media	P75	P90
	isolato	collegato					
Professionista	12,4	87,6	1	2	2,23	4	4
Commerciante	68,0	32,0	0	0	0,71	1	4
Impresa	72,8	27,2	0	0	0,48	1	1
TOTALE	55,7	44,3	0	0	1,00	1	4

Onde inquadrare il fenomeno da una prospettiva più ampia, si è proceduto ad effettuare una analisi fattoriale sotto forma di una Analisi delle Corrispondenze Multiple (ACM), i cui risultati sono rappresentati nella Figura 4. E' qui possibile apprezzare il ruolo delle variabili descritte in precedenza nell'ambito di una interpretazione unitaria, consentendo di capire quali tendenze di fondo danno una struttura al fenomeno complesso multidimensionale. Esistono due principali tendenze di fondo, che rappresentano una ampia porzione della variabilità lineare del fenomeno complesso⁷. Il primo di questi, l'asse principale rappresentato sulle ascisse, costituisce l'asse della produttività aziendale, un *continuum* che segnala un progressivo crescendo dello sviluppo aziendale. La progressione della produttività del lavoro è accompagnata dal crescere monotono della scolarizzazione, che è fortemente connessa con il valore aggiunto per addetto. Si noti un analogo andamento dell'ampiezza delle rete commerciale, la cui netta dicotomia fra essere isolati ("no clienti") e avere qualunque ordine di clienti rispecchia l'alta quota di imprese individuali isolate che esistono in Italia. Comunque inteso, esiste una forte evidenza circa la connessione fra produttività e ampiezza della rete *trade*. Se l'aspetto relazionale svolge un ruolo di grande importanza, altrettanto può dirsi di quello territoriale, che costituisce la seconda tendenza di fondo del fenomeno complesso. La progressione

⁷ Si ricorda che l'inerzia non è stata rivalutata, per cui trattasi di un risultato di tutto rispetto in termini di variabilità lineare rappresentata dal modello, essendo tale statistica, nel caso della ACM una interpretazione "pessimistica" del fenomeno (Benzecri, 1979).

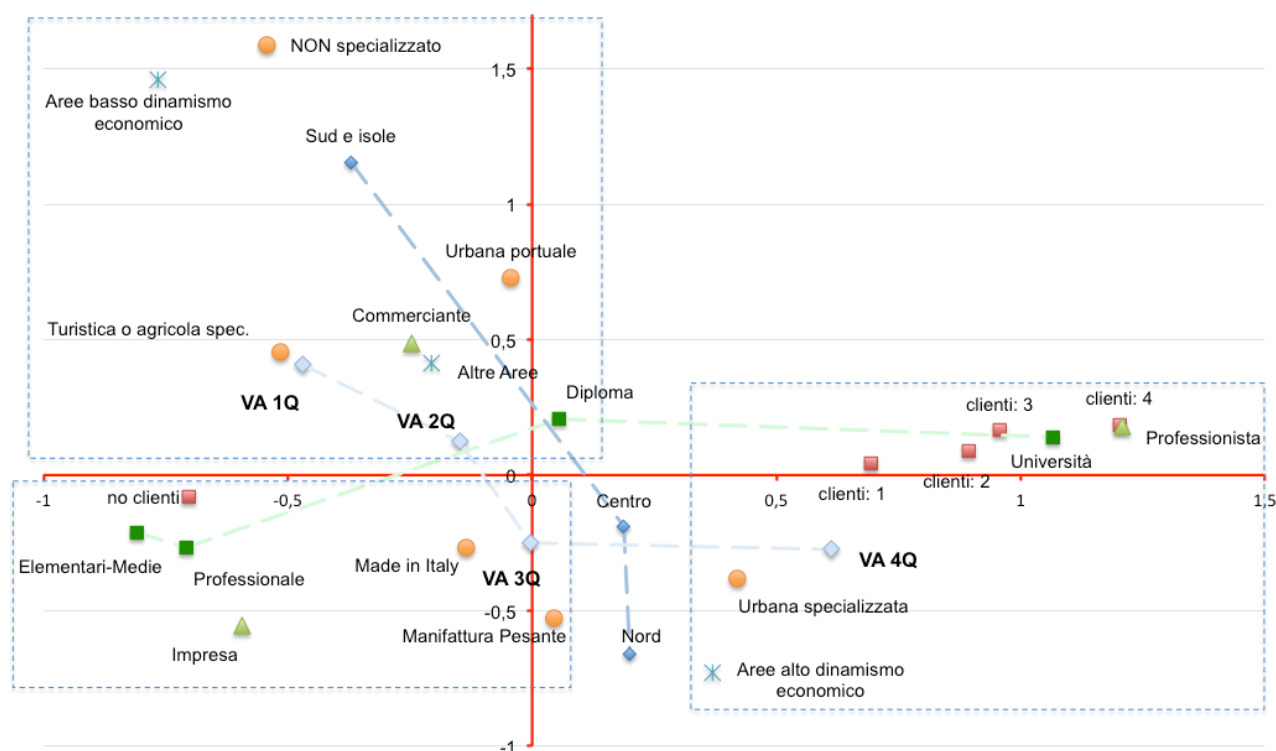
è da sud verso nord, ma soprattutto segna l'incidere da aree a scarsa specializzazione (sovra rappresentate nel meridione) ad aree ad alta specializzazione, in cui spiccano aree industriali (distrettuali e non) ed urbane mono/pluri specializzate.

Interpretati gli assi, si sono individuati 3 nuclei fattuali, che sono interpretabili come tipologie di imprese statisticamente rilevanti (indicati dai rettangoli tratteggiati di colore blu). Partendo dall'alto troviamo il primo gruppo, in prevalenza commercianti, caratterizzato da basso valore aggiunto, mercato di riferimento costituito principalmente da aree non specializzate a basso dinamismo economico, prevalentemente collocate al sud. Il diploma è il titolo di studio caratteristico.

Il secondo nucleo fattuale, quello in cui prevalgono le imprese *tout-court*, è caratterizzato invece da buoni livelli di produttività, ubicazione dell'attività aziendale in aree di specializzazione industriale (distrettuali e non), Prevalentemente collocati al centro-nord (segnatamente nord-est). Il titolo di studio segnala una scolarizzazione medio-bassa.

Sempre la collocazione al centro-nord contraddistingue infine l'ultimo nucleo, quello costituito dalle unità economiche a maggiore produttività, collocate in aree ad alto dinamismo economico (soprattutto urbane mono-pluri specializzate), alti livelli di produttività, ampia rete commerciale in cui sono inseriti. Prevalgono i professionisti e il titolo di studio è quello universitario.

Figura 4 – Analisi multivariata dei dati



I tre nuclei fattuali non sono ovviamente costituiti esclusivamente dalle rispettive tipologie di imprese individuali caratteristiche, anche se ciascuna delle tre li popola in misure superiori alla media. Senza passare per una cluster analysis, proviamo così a valutare il profilo associativo delle variabili di capitale e imprenditoria rispetto alla produttività, testandone gli effetti marginali in un semplice modello regressivo.

5.2 Analisi di regressione

I risultati descritti nei precedenti paragrafi evidenziano come le variabili selezionate per la caratterizzazione dei lavoratori individuali sono correlate rispetto alle variabili di performance abitualmente utilizzate in ambito economico (valore aggiunto per addetto). Segnalano cioè delle partizioni del collettivo fortemente esplicative rispetto alla capacità di generare ricavi ed alla produttività di impresa. È possibile sintetizzare

tutto ciò attraverso una semplice modellizzazione, nella quale viene misurata l'influenza delle diverse variabili scelte per caratterizzare l'agente della produttività del lavoro autonomo individuale.

Come prima evidenza bisogna ricordare l'incipit del presente lavoro: l'estrema polverizzazione che caratterizza la realtà imprenditoriale italiana. Empiricamente essa comporta che un semplice modello di regressione, a fini puramente descrittivi, possiede una bassa capacità predittiva ($R^2 \approx 0,08$), pur in presenza di *trend* statisticamente significativi nella sottostante nuvola dei punti. Tali *trend* statistici sono quelli mostrati con le precedenti analisi descrittive: tutte le variabili considerate nel modello sono significative all'1%. In Figura 5 sono rappresentati gli effetti marginali per tutto il collettivo e successivamente per ciascuna delle figure professionali che caratterizza in maniera precipua i nuclei fattuali. Gli effetti dei profili associativi sulla produttività vengono qui specificati e differenziati per categoria di lavoratore autonomo, consentendone il confronto diretto.

Figura 5 – Impatto delle caratteristiche di imprenditorialità e capitale umano rispetto alla produttività attesa (logaritmo). Valori di regressione (anno 2015).

Parametro	Stima			
	Totale	Commercianti	Professionisti	Imprese
intercetta	9,1168	8,0498	8,3828	9,4474
Età dell'impresa	0,0233	0,0374	0,0388	0,0139
Età dell'impresa (eff. quadratico)	-0,0002	-0,0002	-0,0009	-0,0001
Tasso di femminilizzazione	-0,0047	-0,0050	-0,0018	-0,0054
Aree alto dinamismo economico	0,1850	0,1570	0,1742	0,1464
Aree basso dinamismo economico	-0,0652	0,0583	-0,1283	-0,1084
Altre Aree	0,0000	0,0000	0,0000	0,0000
Urbana portuale	0,0368	-0,0258	0,0556	0,0492
Manifattura Pesante	0,1774	0,0737	0,1162	0,2441
Made in Italy	0,2386	0,1425	0,1223	0,2984
Turistica o agricola spec.	0,2504	0,3776	0,0558	0,2154
Urbana specializzata	0,1673	0,0454	0,1223	0,1852
NON specializzato	0,0000	0,0000	0,0000	0,0000
Artigiano: no	-1,1586	-0,3367	0,3940	-0,3120
Artigiano: si	0,0000	0,0000	0,0000	0,0000
clienti: 4	1,6235	1,6265	0,8913	0,9350
clienti: 3	1,6014	2,1277	0,6998	0,6558
clienti: 2	1,5594	2,0341	0,7165	0,6296
clienti: 1	1,4204	1,7166	0,6151	0,7613
no clienti	0,0000	0,0000	0,0000	0,0000
Esportatore: no	-0,1750	-0,5062	0,4183	-0,1061
Esportatore: si	0,0000	0,0000	0,0000	0,0000
Università	1,1132	1,0929	0,3052	0,9526
media superiore	0,4215	0,5783	0,1664	0,1718
professionale	0,4997	0,8091	0,2020	0,3484
elementare-media inferiore	0,0000	0,0000	0,0000	0,0000

Importante il ruolo svolge altresì l'età anagrafica dell'impresa (una sorta di *tenure* del lavoratore autonomo), che come variabile *proxy* della capacità competitiva rappresenta una crescente produttività legata all'esperienza professionale (rilevante soprattutto per commercianti e professionisti).

Riguardo l'area del capitale umano, il titolo di studio manifesta una progressione crescente clamorosa sulla produttività del lavoratore autonomo. A parità di altre condizioni, rispetto agli imprenditori con scuola dell'obbligo, a una formazione superiore è associato il livello di produttività atteso del +49%, +42% e +111% (rispettivamente per istituto professionale, diploma e istruzione universitaria). Si noti come il ruolo dell'istruzione, seppur presente, risulti molto meno marcato per i professionisti (proprio perché caratterizzati da alfabetizzazione media più elevata).

Considerazioni analoghe, ma ancora più esplosive riguardo l'inserimento in una rete *trade*: l'effetto è assai significativo nell'inserimento in una rete commerciale (abbandono della condizione di "isolamento") e tende a scemare mano a mano che ci si allontana dai committenti B2B diretti. L'effetto sulla produttività si esaurisce (diventa non significativo) passato il quarto *step* di distanza. L'effetto sulla produttività può essere inteso come il combinato degli effetti di maggiori opportunità di business legati ad una rete più ampia; così come al fatto che orientarsi ad una clientela B2B presuppone un livello di organizzazione e di qualità del prodotto di maggior livello rispetto alla mera vendita al consumatore finale.

Coerente con la precedente analisi multivariata anche il ruolo esercitato dal territorio. Quasi scontato il segno del dinamismo economico territoriale e interessante il segno positivo per i commercianti (+5% nelle aree a basso dinamismo), che fa supporre maggiori utili legati a scarsa concorrenza dei mercati distributivi in aree marginali. Interessante il ruolo delle aree specializzazione, dove spicca il ruolo delle aree metropolitane ed industriali (con particolare riferimento alle imprese in senso stretto).

Infine alcune considerazioni sulle caratteristiche di capitale umano non direttamente produttive. Trattasi delle variabili al punto 11./12. del paragrafo 3. Variabili prettamente socio-demografiche, non direttamente connesse alla produttività, ma che individuano sottopopolazioni tipicamente rilevanti in termini sociali e di disparità retributive. Questa evidenza è confermata anche dalla presente analisi. Nello specifico la nazionalità non costituisce, a parità di altre condizioni, un elemento discriminante (variabile non significativa, omessa perciò dal modello), mentre *pattern* di svantaggio importante sono da segnalare con rispetto al genere. A parità delle altre condizioni, questa variabile mostra valori medi di produttività attesa per le donne inferiori di circa il 18% nel caso dei professionisti, ma che possono arrivare al 50% per le altre categorie di lavoratori (sono in effetti i professionisti la categoria in cui sono sovra rappresentate le donne fra i lavoratori autonomi). E' questa una evidenza di grande importanza: il gender pay gap è infatti un tema molto sentito per quello che riguarda i lavoratori dipendenti, ma evidentemente il ruolo del pregiudizio e delle concezioni di genere svolge un ruolo significativo nel determinare situazioni di svantaggio di genere anche sul versante del lavoro indipendente.

6 Conclusioni

Il presente lavoro intende ha proposto una descrizione dei lavoratori autonomi presenti nel settore privato in Italia, fornendo un quadro di analisi in grado di caratterizzare una tipologia di impresa socialmente molto rilevante nonché tipica del sistema economico italiano. Tali imprese rappresentano una parte significativa del sistema economico italiano, sia in termini di numerosità sia in relazione alla quota di valore aggiunto nazionale.

Il quadro d'analisi proposto è un sistema informativo costituito da nodi (le imprese) e relazioni (i rapporti commerciali) che consente la massima ampiezza dell'informazione adatta a definire il target di analisi. In particolare, le caratteristiche posizionali (ad es. struttura e profittabilità di impresa, localizzazione) e le informazioni del network in cui è inserito il soggetto (espresso da relazioni commerciali), congiuntamente alle informazioni di contesto economico, permettono di ottemperare i più svariati ambiti conoscitivi, sia a livello di impresa (profitto, capacità di sopravvivenza, competitività) sia a livello di sviluppo locale (ampiezza e densità delle reti, ricadute economiche sul territorio, resilienza). Le maggiori evidenze emerse dall'analisi empirica sono state le seguenti:

- gli imprenditori individuali rappresentano un mondo estremamente variegato;
- l'estrema polverizzazione implica grande dispersione nella nuvola dei punti, ma consente comunque di trovare trend significativi;
- gli aspetti territoriali, correttamente specificati, sono importanti e articolati e caratterizzano in maniera specifica i gruppi di lavoratori autonomi;
- la caratterizzazione "mista" imprenditoriale/capitale umano sembra una modellizzazione valida, da perfezionare su diversi punti.

7 Abstract in inglese

The purpose of this work is to provide a comprehensive representation of the business networks in which the self-employed workers are inserted, a socio-economic aggregate of great importance in the national context which represents more than 45% of the total active enterprises registered in the Statistical Archive of Companies Active - ASIA and 7% of total added value.

The emphasis of this work focuses in particular on the characteristics of local business networks, ie the commercial network (transactions) in which the self-employed workers are included, tracing an ideal framework that ranges from lonely players ("free beaters") to economic actors that constitute important nodes of important economic fabrics. By exploiting the analytical tools of the network analysis, clusters of relationships are defined for territorially compact areas (Local Job Systems) and the membership of the network is investigated in terms of greater profitability / efficiency of the individual company.

Cardinality of the network; depth, density and extension of relationships; ownership of the nodes (turnover, company age, location, levels of education) allow an innovative representation of the problem, while providing important insights for regional policies

8 Bibliografia

- [1] Becker G.S. (1964), "Human Capital" – ed. USA: Prentice-Hall.
- [2] Benzécri, J.P. (1979). Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. Cahiers de l'Analyse des Données, 4, 377–378.
- [3] Biagi F. e M.L. Parisi (2012), "Are ICT, human capital and organizational capital complementary in production? Evidence from the Italian manufacturing sector", JRC Working Papers JRC75890, Sevilla
- [4] ISTAT, Linee guida per la qualità dei processi statistici che utilizzano dati amministrativi - Versione 1.1. Roma Agosto 2016.
- [5] Istat (2018), Rapporto sulla competitività dei settori produttivi, <https://www.istat.it/storage/settori-produttivi/2018/Rapporto-competitivita-2018.pdf>.
- [6] National Statistics (2004a). National Statistics code of practice – protocol on Data Matching. Office for National Statistics, London.
- [7] National Statistics (2004b). National Statistics code of practice – protocol on Statistical Integration. Office for National Statistics, London.