XXXIX CONFERENZA ITALIANA DI SCIENZE REGIONALI

La geografia del capitale umano: disparità territoriali e sviluppo locale

Paola Bosso[[1]](#footnote-1), Stefano De Santis[[2]](#footnote-2), Federica Pintaldi[[3]](#footnote-3)

**ABSTRACT**

Lo scopo del presente lavoro è fornire una mappa censuaria, definita a livello di provincia/comune sulla dotazione di capitale umano presente nel sistema economico italiano e di indagare le sue relazioni con i livelli di sviluppo, dinamicità e produttività. A questo fine è stato necessario definire un articolato data set di informazioni, ottenuto come combinazione di diversi fonti di informazioni (in parte di natura amministrativa, in parte di natura statistica) convenientemente integrate fra di loro. E’ così possibile esaminare la relazione fra capitale umano e livelli di sviluppo, in linea sia con la recente teoria economica, così come nel rispetto della necessità di definire in maniera statisticamente coerente gli aggregati di studio. La realizzazione della mappatura del capitale umano richiede l’incrocio di dati amministrativi disponibili su ciascuna posizione lavorativa (a partire dell’incrocio dei dati della social security con quelli previdenziali/fiscali delle certificazioni dei redditi), con dati statistico/amministrativi sui titoli di studio, nonché dati sulla condizione lavorativa proveniente dalla indagine delle Forze di Lavoro. E’ stato così possibile definire un quadro censuario del capitale umano (basato sull’approccio del capitale umano usato da EUROSTAT per valutare i differenziali retributivi – pay gap – sul mercato del lavoro) ad elevatissimo dettaglio territoriale. La stima della dotazione di capitale umano è stata altresì utilizzata per indagare le sue relazioni con i complessivi livelli di sviluppo economico-sociale nei cluster territoriali individuati.

# Introduzione

Scopo del presente lavoro è definire una mappa che restituisca le dotazioni dotazione di capitale umano presente nel sistema economico italiano e di indagare le sue relazioni con i livelli di sviluppo, dinamicità e produttività. E’ perciò un lavoro che possiamo immaginare come idealmente diviso in 2 parti.

Una prima fase in cui si descrive posta la costruzione della base dati mediante integrazione di fonti di natura statistica ed amministrativa, in maniera da ottenere un dataset coerente con gli obiettivi conoscitivi. Tutta l’enfasi è perciò posta sulle fonti di dati, la procedura di integrazione fisica e armonizzazione statistica, la qualità della base dati integrata costruita. Lo scopo finale è sfruttare l’impianto teorico ed empirico utilizzato da Eurostat per la valutazione dei differenziali retributivi, per effettuare una stima delle dotazioni di capitale umano a livello provinciale/comunale.

Esaurita la fase tecnica, la seconda fase sposta invece il focus su un obiettivo di analisi. Più precisamente lo scopo è sfruttare la mappa delle dotazioni di capitale umano per indagare il nesso esistente con i complessivi livelli di sviluppo economico-sociale, suggerendo in questo modo delle spiegazioni riguardo le diverse traiettorie di sviluppo possibili nei cluster territoriali individuati.

# 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.

## Fonti di dati

Rilevazione Continua sulle Forze di lavoro.

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 dei redditi diversi da parte del sostituto di imposta, da lavoro dipendente, da pensione, da lavoro autonomo o da 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”[[4]](#footnote-4). 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 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.

## Record linkage

Effettuata una ricognizione delle fonti necessarie, si è provveduto alla integrazione fisica delle informazioni (record linkage) e successiva armonizzazione probabilistica delle unità campionarie, mediante calibrazione delle pionarie che risultano dal precedente passo di integrazione rispetto a totali di popolazione (noti da fonte amministrativa, ovvero da stime campionarie).

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-deduplicazione). 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.

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 dell’Indagine sulle Forze di Lavoro, l’estrazione del campione avviene direttamente dalle Liste Anagrafiche Comunali, che vengono pretrattate in ISTAT all’interno del sistema di integrazione dei dati amministrativi (SIM)[[5]](#footnote-5). 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 (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 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”.

Tabella –

Fonte: Istat,

Figura –

## Metodologia di integrazione fisica e armonizzazione statistica

## Integrazione e qualità della base dati costruita

# Dotazioni di capitale umano e i livelli di sviluppo economico-sociale: un’analisi spaziale

# Conclusioni

# Abstract in inglese

The purpose of this work is to provide a census map, defined at the province / municipality level on the human capital endowment present in the Italian economic system and to investigate its relations with the levels of development, dynamism and productivity. To this end, it was necessary to define an articulated data set of information, obtained as a combination of different sources of information (partly of an administrative nature, partly of a statistical nature) conveniently integrated with each other. It is thus possible to examine the relationship between human capital and levels of development, in line with recent economic theory, as well as in respect of the need to define the study aggregates in a statistically coherent manner. The mapping of human capital requires the intersection of administrative data available on each job position (starting from the intersection of social security data with the social security / income tax certificates), with statistical / administrative data on educational qualifications , as well as data on the working condition from the Labor Force survey. It was thus possible to define a census of human capital (based on the approach of human capital used by EUROSTAT to evaluate the pay gap - pay gap - on the labor market) with a very high territorial detail. The estimate of the human capital allocation was also used to investigate its relations with the overall levels of economic-social development in the identified territorial clusters.

# Bibliografia

1. Becker G.S. (1964), “Human Capital" – ed. USA: Prentice-Hall.
2. 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
3. Istat (2018), Rapporto sulla competitività dei settori produttivi, <https://www.istat.it/storage/settori-produttivi/2018/Rapporto-competitivita-2018.pdf>.

1. Istat, V. Balbo, Roma, [paola.bosso@istat.it](mailto:paola.bosso@istat.it) [↑](#footnote-ref-1)
2. Istat, V. Balbo, Roma, [sdesantis@istat.it](mailto:sdesantis@istat.it) [↑](#footnote-ref-2)
3. Istat, V. Balbo, Roma, [pintaldi@istat.it](mailto:pintaldi@istat.it) [↑](#footnote-ref-3)
4. Corrispondente grosso modo all’anno solare, prolungato sino al 12 gennaio per consentire al sostituto d’imposta di effettuare I conguagli. [↑](#footnote-ref-4)
5. 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 [↑](#footnote-ref-5)