

La profilazione delle persone disponibili al lavoro in Piemonte

Il ruolo dei dati amministrativi e dei dati territoriali locali

Monica Bosia
Agenzia Piemonte Lavoro

Marianna Carlone
Università degli Studi di Torino

Elisa Floredan
Agenzia Piemonte Lavoro

Francesca Prever
Università degli Studi di Torino

Il lavoro analizza il potenziale dell'integrazione tra dati amministrativi individuali e indicatori territoriali per il miglioramento dei modelli di profiling dei Servizi per l'impiego. Utilizzando dati delle schede anagrafico-professionale piemontesi e un modello logit di riferimento Anpal, si mostra che l'introduzione di variabili dei Sistemi locali del lavoro e l'utilizzo di tecniche di machine learning migliorano la capacità predittiva e la sensibilità nella previsione del mancato reinserimento occupazionale, con implicazioni per le politiche attive del lavoro.

Profiling people available for work in Piedmont. The role of administrative data and local territorial data
This paper analyses the integration of individual administrative data and local labour market indicators to improve profiling models used in employment services. Using Piedmont SAP data and a benchmark Anpal logit model, we show that including Local Labour Systems variables and machine learning techniques enhances predictive performance and sensitivity in identifying individuals at risk of non-employment, with relevant implications for active labour market policies.

DOI: 10.53223/Sinappsi_2026-02-10

Citazione	Parole chiave	Keywords
Bosia M., Carlone M., Floredan E., Prever F. (2026), La profilazione delle persone disponibili al lavoro in Piemonte. Il ruolo dei dati amministrativi e dei dati territoriali locali, <i>Sinappsi</i> , XVI, n.2, pp.145-160	Dati amministrativi Mercato del lavoro Occupabilità	<i>Administrative data Labour market Employability</i>

Introduzione

Il progressivo spostamento dell'analisi economica dai dati campionari ai dati amministrativi rappresenta uno dei cambiamenti strutturali più rilevanti nella ricerca empirica contemporanea. Nelle principali riviste internazionali – quali *American Economic Review* e *Quarterly Journal of Economics* – la quota di articoli basati su microdati d'indagine si è drasticamente ridotta tra il 1980 e il 2010, mentre è cresciuta in misura speculare la produzione fondata su archivi amministrativi (Chetty 2012). Anche nel contesto italiano tale dinamica è evidente: una larga parte degli studi empirici sul mercato del lavoro utilizza oggi basi dati amministrative, in particolare quelle dell'Inps, a fronte di un ricorso relativamente più contenuto all'Indagine sulle Forze di lavoro dell'Istat (Leombruni 2024).

Questo mutamento non è soltanto quantitativo,

ma riflette una trasformazione più profonda nelle esigenze conoscitive della ricerca economica e sociale. La crescente attenzione verso l'eterogeneità individuale, le micro-dinamiche e gli effetti distributivi delle politiche – in particolare nell'ambito delle politiche del lavoro e del welfare – alimenta una domanda strutturale di microdati. La valutazione causale degli interventi pubblici richiede infatti informazioni granulari sui percorsi individuali, non sempre osservabili attraverso indagini campionarie. In questa prospettiva, i microdati assumono le caratteristiche di un bene pubblico, il cui valore scientifico dipende anche dalla possibilità di accesso e replicabilità delle analisi (Trivellato 2019).

Il tema dell'accesso e dell'integrazione delle fonti informative si colloca, tuttavia, all'incrocio tra esigenze di ricerca e tutela della riservatezza.

La possibilità di collegare archivi amministrativi a popolazioni statisticamente ben definite costituirebbe una base solida per l'inferenza, ma – anche alla luce delle normative sulla protezione dei dati – tale ancoraggio non è sempre disponibile ai ricercatori. Nel contesto italiano, l'utilizzo degli archivi amministrativi per la ricerca ha radici consolidate ma ha conosciuto una nuova accelerazione grazie, da un lato, all'introduzione delle Comunicazioni Obbligatorie (COB) che tracciano in modo sistematico avvisi, trasformazioni e cessazioni dei rapporti di lavoro e, dall'altro, alla crescente attenzione per le disuguaglianze e i determinanti sociali delle condizioni economiche e occupazionali (Leombruni e Ardito 2026). Tutto ciò ha stimolato diverse iniziative volte a promuovere l'uso dei dati amministrativi su occupazione, welfare, istruzione e salute a fini di ricerca.

Il database COB contiene infatti informazioni estremamente ricche, con variabili relative al lavoratore – quali dati demografici di base e livello di istruzione – e informazioni molto dettagliate sul rapporto di lavoro, tra cui la tipologia contrattuale, il contratto collettivo applicato, una classificazione occupazionale molto fine e il settore di attività. La raccolta e la gestione di tali comunicazioni avvengono attraverso sistemi informativi regionali o tramite il Sistema nazionale. In Piemonte, ad esempio, le COB sono gestite attraverso la piattaforma GE.CO (Gestione online Comunicazioni Obbligatorie), mentre altre regioni utilizzano direttamente il portale nazionale del Ministero del Lavoro. La disponibilità di informazioni tempestive e riferite all'intera popolazione ha favorito una rapida espansione degli studi sul mercato del lavoro italiano, in particolare a livello regionale. Molte regioni italiane consentono oggi l'accesso scientifico ai propri archivi COB regionali, rendendo possibile la produzione di analisi dettagliate e aggiornate sulle dinamiche dei mercati del lavoro locali.

In tale scenario si inserisce il presente lavoro, che esplora l'impiego integrato di dati amministrativi provenienti da Agenzia Piemonte Lavoro (APL) e di informazioni territoriali dell'Istat a livello di Sistemi locali del lavoro (SLL) per finalità di profiling. La profilazione viene effettuata al momento del rilascio della Dichiarazione di immediata disponibilità (DID), l'atto formale con cui un individuo si dichiara immediatamente disponibile al lavoro e che costituisce il punto di ingresso nel sistema di osservazione amministrativa utilizzato nello

studio. A seguito della DID, i soggetti sono tenuti a sottoscrivere un Patto di servizio personalizzato (PSP) e a partecipare alle politiche attive del lavoro, attraverso la partecipazione a misure definite con il supporto del Centro per l'impiego. Tali interventi comprendono un insieme di servizi (LEP – Livelli essenziali delle prestazioni) calibrati in funzione del livello di occupabilità e di fragilità risultante dalla profilazione quantitativa e dall'assessment qualitativo (Anpal¹ 2022).

L'obiettivo è verificare se l'utilizzo di flussi amministrativi ancorati a contesti locali consenta di migliorare la capacità predittiva e discriminante dei modelli di profiling rivolti ai soggetti in cerca di occupazione oggi utilizzati a livello nazionale. L'ipotesi di fondo è che una maggiore aderenza al contesto territoriale effettivo in cui si svolgono le transizioni lavorative possa cogliere meglio eterogeneità e dinamiche locali, rafforzando la qualità informativa della profilazione e, conseguentemente, il potenziale supporto alle politiche attive del lavoro.

Il presente articolo si sviluppa secondo il seguente schema. La sezione 1 illustra il ruolo e le caratteristiche dei dati amministrativi, con particolare attenzione al quadro normativo italiano. La sezione 2 ricostruisce l'evoluzione del profiling nei Servizi per l'impiego, a livello internazionale e nazionale, e descrive il modello Anpal. La sezione 3 espone la strategia empirica e confronta il modello nazionale con la replica su dati regionali piemontesi, attraverso l'analisi degli effetti marginali su profili rappresentativi. La sezione 4 estende l'analisi introducendo indicatori di contesto territoriale a livello di Sistema locale del lavoro, valutandone il contributo alla capacità predittiva. La sezione 5 approfondisce l'impiego di tecniche di machine learning (Random Forest e Boosting). L'ultima sezione conclude discutendo le implicazioni di policy e le prospettive di sviluppo.

1. I dati amministrativi come risorsa per la ricerca e le politiche pubbliche

Caratteristiche, potenzialità e quadro normativo

I dati di natura amministrativa costituiscono un capitale informativo strategico per la progettazione, l'attuazione, il monitoraggio e la valutazione delle politiche pubbliche, in quanto sono caratterizzati da elevata copertura informativa, granularità dei dati e aggiornamento continuo. Le principali fonti comprendono i registri civili, utilizzati per le analisi

1 Agenzia nazionale per le Politiche attive del lavoro (Anpal) (2022), *Delibera n. 5 del 9 maggio 2022, Approvazione strumenti per la profilazione quantitativa e qualitativa nell'attuazione dell'assessment, standard dei servizi e costi standard, Banca Dati del Lavoro*. <https://bdl.sviluppopolavoroitalia.it/bdds/ViewScheda.action?product=NORMALAVORO&uid=b3b7658c-3c87-44c9-b85b-ea1eef776f33&title=scheda>.

demografiche e della struttura della popolazione, i dati fiscali, impiegati per lo studio delle dinamiche economiche, e, soprattutto per quanto qui rileva, i registri sull'impiego, che consentono di analizzare il funzionamento del mercato del lavoro e di valutare l'efficacia delle politiche occupazionali. L'impulso del Legislatore è stato determinante nella costruzione di questo patrimonio informativo, inizialmente finalizzato a migliorare l'accesso ai servizi pubblici e a semplificare i procedimenti amministrativi attraverso l'uso delle tecnologie digitali. Un passaggio fondamentale è rappresentato dalla legge n. 59/1997² (c.d. legge Bassanini), cui ha fatto seguito l'adozione del Codice dell'amministrazione digitale (CAD) nel 2005³, che ha definito il quadro giuridico per la digitalizzazione della pubblica amministrazione e per la condivisione dei dati tra enti.

Il Codice ha inoltre introdotto infrastrutture e principi volti a favorire l'interoperabilità dei sistemi informativi pubblici, tra cui la Piattaforma digitale nazionale dati (art. 50-ter CAD) e i principi del *digital first* e del *once only*⁴. In questo contesto si colloca anche il ruolo dell'Agenzia per l'Italia digitale⁵, responsabile della definizione delle linee guida tecniche per la gestione, la conservazione e l'accesso ai dati della Pubblica amministrazione, secondo standard che ne garantiscano integrità, sicurezza e tracciabilità.

In linea con i principi dell'open government, orientati a garantire trasparenza e integrità nei processi decisionali pubblici⁶, le amministrazioni sono tenute a rendere disponibili in formato open data le informazioni di interesse pubblico prodotte nell'ambito dell'attività amministrativa. Questo quadro normativo, rafforzando l'accountability delle istituzioni, favorisce anche l'accesso ai dati per finalità di ricerca. Ulteriori disposizioni regolano l'utilizzo dei dati amministrativi a fini statistici e scientifici: il D.Lgs. n. 322/1989⁷ prevede l'uso prioritario degli archivi amministrativi per statistiche ufficiali, mentre il Regolamento (UE) 2016/679⁸ e la Direttiva Comstat n. 11/2018 disciplinano la tutela dei dati personali e le modalità di accesso ai microdati anonimizzati per scopi di ricerca.

Dati amministrativi: qualità, limiti e valore strategico per le politiche pubbliche

Pur essendo raccolti nell'ambito dell'attività amministrativa, tali dati possono essere efficacemente utilizzati anche per finalità statistiche e di ricerca, consentendo analisi basate su archivi informativi già esistenti. La registrazione sistematica delle informazioni sull'intero universo di riferimento consente di disporre di basi dati molto estese, che rendono possibili analisi dettagliate su popolazioni e territori e riducono la dipendenza da rilevazioni campionarie, pur richiedendo specifiche attenzioni in termini di qualità e copertura delle informazioni. La qualità dei dati può infatti variare in funzione delle modalità di raccolta e aggiornamento, e la presenza di errori di imputazione, informazioni mancanti o non aggiornate può influenzare l'affidabilità delle analisi statistiche e delle evidenze empiriche. Per questo motivo assumono particolare rilevanza i protocolli di verifica e di valutazione della qualità come quelli adottati nell'ambito del Sistema Statistico Nazionale e dell'Istat, che prevedono criteri standardizzati per la misurazione di dimensioni quali rilevanza, accuratezza, completezza, tempestività, coerenza e accessibilità dei dati⁹.

Nonostante l'ampiezza e la granularità, i dati amministrativi presentano limiti nella copertura informativa: non intercettano fenomeni sociali per i quali non è necessario né obbligatorio rivolgersi al pubblico servizio. Per questo, è spesso opportuno integrarli con fonti campionarie o qualitative, adottando disegni di ricerca misti che consentano una comprensione più ampia di fenomeni complessi (Istat 2014).

In Piemonte come in altre regioni d'Italia (Bazzoli *et al.* 2018) l'integrazione di fonti amministrative e campionarie è stata utilizzata per leggere le trasformazioni del mercato del lavoro a livello territoriale, analizzare i percorsi individuali di accesso alle politiche attive, i fabbisogni di competenze espressi dal tessuto produttivo, e orientare di conseguenza la programmazione di interventi formativi e di accompagnamento al lavoro più mirati. In questo senso, l'analisi dei dati amministrativi si configura come strumento di raccordo tra sistema

2 L. 15 marzo 1997 n. 59.

3 D.Lgs. 7 marzo 2005 n. 82 (e successive modificazioni).

4 Il principio *digital first* implica che i servizi pubblici siano erogati prioritariamente in modalità digitale; il principio *once only* stabilisce che le informazioni personali debbano essere comunicate una sola volta allo Stato.

5 Per le linee guida sulla conservazione, si veda: Agenzia per l'Italia digitale (2021), disponibile su: www.agid.gov.it/piattaforme/conservazione.

6 Cfr. Dipartimento della Funzione Pubblica (2024), disponibile su: open.gov.it.

7 D.Lgs. 6 settembre 1989 n. 322.

8 Regolamento (UE) 2016/679 del Parlamento europeo e del Consiglio 27 aprile 2016.

9 Sistan, La qualità dei dati amministrativi per la statistica, <https://www.sistan.it/index.php?id=424>.

del lavoro, formazione e servizi alle persone e alle imprese. Gli archivi amministrativi non sono solo un supporto digitale, ma una risorsa strategica: la loro interpretazione amplifica le opportunità di incontro tra domanda e offerta di lavoro e abilita modelli decisionali *evidence based*, rafforzando sia il monitoraggio e la valutazione delle politiche pubbliche, sia i processi decisionali individuali.

I dati amministrativi impiegati nel modello

I dati amministrativi impiegati nello studio derivano dalla Scheda anagrafico-professionale (SAP), che costituisce la principale unità operativa primaria del sistema informativo regionale¹⁰. La SAP rappresenta il fascicolo digitale del lavoratore e raccoglie in modo integrato le informazioni anagrafiche, la biografia professionale completa, il percorso formativo e la storia occupazionale del cittadino, inclusa l'eventuale partecipazione a misure di politica attiva. L'archivio della SAP è articolato in diverse sezioni tra loro collegate: dati anagrafici e di contatto, titoli di studio e formazione, storia lavorativa, stato occupazionale e amministrativo e sezione dedicata alle politiche del lavoro. Quest'ultima comprende, tra le altre informazioni, lo stato di disoccupazione, la Dichiarazione di immediata disponibilità (DID), il Patto di servizio personalizzato (PSP), le azioni concordate con il Centro per l'impiego nell'ambito delle politiche attive del lavoro (ad esempio percorsi collegati al Programma GOL o alla NASpl) e gli esiti della profilazione del lavoratore.

La SAP è alimentata automaticamente dalle Comunicazioni Obbligatorie (COB) e dalle attività dei Centri per l'impiego per quanto concerne le politiche attive del lavoro. I dati vengono aggiornati automaticamente grazie a un sistema di cooperazioni applicative tra archivi regionali e nazionali: il sistema informatico regionale intercetta il codice fiscale del lavoratore e aggiorna automaticamente le informazioni presenti nell'archivio. Gli operatori dei Centri per l'impiego accedono a tali informazioni attraverso il Sistema informativo lavoro Piemonte (SilP).

Profiling – literature review

Negli ultimi decenni, i Servizi pubblici per l'impiego (SPI) hanno assunto un ruolo sempre più centrale nell'implementazione delle politiche attive del lavoro, in un contesto caratterizzato da risorse pubbliche limitate, elevata eterogeneità dei disoccupati e crescenti aspettative in termini di efficacia degli interventi. In questo scenario, il profiling si è affermato come uno strumento chiave di

supporto decisionale, volto a migliorare il targeting delle misure di attivazione e a orientare l'allocazione dei servizi in funzione delle caratteristiche individuali dei beneficiari. In senso generale, il profiling consiste nel processo di raccolta e utilizzo sistematico di informazioni sui disoccupati per valutarne le prospettive di reinserimento lavorativo e definire, di conseguenza, l'intensità e la tipologia degli interventi più appropriati. L'obiettivo principale è duplice: da un lato, individuare tempestivamente le persone a maggior rischio di disoccupazione di lunga durata; dall'altro, evitare un utilizzo inefficiente delle risorse pubbliche su individui che presentano elevate probabilità di reinserimento autonomo nel mercato del lavoro. In tal modo, il profiling rappresenta uno strumento preventivo, volto a ridurre la persistenza della disoccupazione e a migliorare l'efficacia complessiva delle politiche attive (Desiere *et al.* 2019). La letteratura e la pratica internazionale distinguono tradizionalmente tre principali approcci al profiling nei Servizi per l'impiego: il profiling basato su regole amministrative (*rule-based profiling*), quello fondato sul giudizio degli operatori (*caseworker-based profiling*) e il profiling statistico. Il primo utilizza criteri predefiniti, quali età, livello di istruzione o durata della disoccupazione, per segmentare i disoccupati in gruppi di utenza e determinare l'accesso ai servizi. Questo approccio è ampiamente diffuso grazie alla sua semplicità, trasparenza e facilità di implementazione, ma tende a trascurare l'eterogeneità individuale all'interno dei gruppi e può riflettere priorità politiche piuttosto che il rischio effettivo di esclusione dal mercato del lavoro. Il profiling basato sul giudizio del *caseworker* attribuisce invece un ruolo centrale alla valutazione discrezionale degli operatori dei SPI, spesso supportata da strumenti di *assessment* qualitativi o quantitativi. Tale approccio consente di incorporare informazioni difficilmente formalizzabili, come la motivazione o specifiche barriere individuali, ma presenta criticità legate alla soggettività delle valutazioni, alla variabilità tra operatori e ai maggiori costi di implementazione, data la forte dipendenza dalle risorse umane disponibili. Il profiling statistico si propone come un'alternativa più standardizzata, basata sull'utilizzo di modelli quantitativi per stimare la probabilità di esiti occupazionali sfavorevoli, in particolare il rischio di disoccupazione di lunga durata. Rispetto agli approcci tradizionali, esso consente di trattare i beneficiari come individui piuttosto che come membri di gruppi omogenei, favorendo una maggiore personalizzazione dei servizi. Numerosi Paesi hanno progressivamente adottato strumenti di profiling statistico, spesso integrandoli con il giudizio

10 Agenzia Piemonte Lavoro, *Scheda anagrafica e professionale (Sap)*, disponibile su: <https://agenziapiemontelavoro.it/scheda-informativa/scheda-anagrafica-e-professionale-sap/>.

degli operatori, in un'ottica di complementarità piuttosto che di sostituzione (Loxha e Morgandi 2014).

Le esperienze internazionali mostrano alcune best practice ricorrenti nella progettazione e nell'uso dei sistemi di profiling. In primo luogo, l'efficacia del profiling dipende in misura cruciale dalla qualità e dall'integrazione dei dati amministrativi disponibili, che devono essere aggiornati e sufficientemente ricchi da cogliere le principali dimensioni dello svantaggio occupazionale (Bollens 2022). In secondo luogo, i sistemi più efficaci sono quelli inseriti in quadri istituzionali coerenti, in cui il profiling è parte integrante del processo di presa in carico e di definizione dei percorsi personalizzati. Infine, emerge l'importanza di un utilizzo flessibile dei risultati del profiling, che supporti il lavoro dei caseworker senza vincolarlo rigidamente, preservando margini di adattamento alle specificità individuali (Desiere *et al.* 2019). Nella maggior parte dei casi, tali modelli si basano su tecniche statistiche relativamente consolidate e trasparenti, utilizzate per stimare la probabilità di reinserimento lavorativo o, in alternativa, il rischio di disoccupazione di lunga durata entro un determinato orizzonte temporale. Esempi emblematici sono rappresentati dal *Worker Profiling and Reemployment Services* negli Stati Uniti, dal *Job Seeker Classification Instrument* in Australia e dal *Work Profiler* nei Paesi Bassi, tutti accomunati dall'uso di informazioni amministrative individuali relative alla storia lavorativa, al livello di istruzione e alle caratteristiche sociodemografiche. Un'ulteriore best practice riguarda l'aggiornamento periodico dei modelli e la loro integrazione nei sistemi informativi dei SPI. Poiché le relazioni tra caratteristiche individuali ed esiti occupazionali possono variare nel tempo in funzione del ciclo economico e delle trasformazioni del mercato del lavoro, i modelli di profiling statistico vengono regolarmente ricalibrati sulla base dei dati più recenti (O'Connell *et al.* 2009).

Nel contesto italiano, l'adozione di strumenti di profiling si è sviluppata con un certo ritardo rispetto ad altri Paesi europei. Per lungo tempo, il sistema delle politiche attive del lavoro si è basato prevalentemente su criteri amministrativi e su un ruolo limitato dei Servizi per l'impiego. Un primo cambiamento significativo è avvenuto con il programma Garanzia Giovani, che ha introdotto un modello di profilazione volto a classificare i beneficiari in fasce di svantaggio sulla base di caratteristiche individuali, fornendo un supporto

strutturato all'orientamento e alla definizione dei percorsi di politica attiva. Un ulteriore passo in avanti è stato compiuto nell'ambito delle riforme avviate con il *Jobs Act*, che ha ridefinito il ruolo dei Servizi per l'impiego e posto l'attivazione al centro delle politiche del lavoro. In questo quadro, la Dichiarazione di immediata disponibilità (DID) e la successiva profilazione dei disoccupati sono diventate passaggi fondamentali per l'accesso alle misure di sostegno e per la stipula dei patti di servizio personalizzati (decreto legislativo n. 150/2015¹¹). L'introduzione della DID online a partire dal 2017 ha inoltre contribuito a migliorare significativamente il patrimonio informativo disponibile. Più recentemente, l'Agenzia nazionale per le politiche attive del lavoro¹² (Anpal) ha formalizzato un modello di profilazione quantitativa a supporto dell'attuazione del Programma garanzia occupabilità lavoratori (GOL), basato esclusivamente su variabili individuali osservabili al momento della sottoscrizione della DID, relative alle caratteristiche anagrafiche, formative, lavorative e familiari del disoccupato (Anpal 2022). Questo modello rappresenta oggi il principale riferimento nazionale per la profilazione nei Servizi per l'impiego e costituisce il punto di partenza per ulteriori sviluppi e applicazioni a livello territoriale.

2. Trasferibilità del modello Anpal: un confronto tra scala nazionale e regionale

Dati del modello

Al fine di valutare la trasferibilità del modello di profiling adottato a livello nazionale e di verificare se l'utilizzo di informazioni riferite al contesto locale possa migliorarne la capacità predittiva, è stato stimato un modello logit su dati amministrativi piemontesi analogo a quello proposto da Anpal (2022). Formalmente, la probabilità di mancato reinserimento occupazionale è modellata come:

$$P(Y_i = 1 | X_i) = \frac{e^{X_i' \beta}}{1 + e^{X_i' \beta}}$$

dove:

- $Y_i=1$ se l'individuo i non risulta occupato nei sei mesi successivi alla sottoscrizione della DID, $Y_i=0$ altrimenti;
 - X_i è il vettore delle variabili esplicative osservate al momento della DID;
 - β è il vettore dei coefficienti da stimare.
- Le variabili esplicative incluse riproducono,

11 D.Lgs. 14 settembre 2015 n. 150.

12 L'Agenzia nazionale per le politiche attive del lavoro (Anpal) è stata soppressa con decorrenza dal 1° marzo 2024. Le relative funzioni sono state attribuite al Ministero del Lavoro e delle politiche sociali, mentre il personale del comparto ricerca è confluito nell'INAPP. Anpal Servizi S.p.A. ha assunto la denominazione di Sviluppo Lavoro Italia S.p.A.

per quanto consentito dai dati disponibili, quelle utilizzate nel modello nazionale di Anpal. Esse comprendono, tra le principali: genere, età, titolo di studio, esperienza lavorativa recente e condizione occupazionale nell'anno precedente.

La tabella A1 in Appendice riporta le stime dei coefficienti relativi alle variabili individuali comuni ai due modelli considerati. Il modello nazionale (Anpal 2022) considera come variabile dipendente il mancato reinserimento a 12 mesi dalla DID e stima un modello logit su un campione di circa 3,4 milioni di DID registrate in Italia nel periodo 2018-2019. Il modello regionale fa invece riferimento al mancato reinserimento a 6 mesi dalla DID, su un campione di 163.433 DID sottoscritte in Piemonte nel 2022. La scelta dell'orizzonte di 6 mesi deriva da un vincolo empirico: l'estensione a 12 mesi avrebbe ridotto in modo significativo la numerosità campionaria, a causa della minore disponibilità di follow-up completi nei dati amministrativi piemontesi, compromettendo la stabilità delle stime e la rappresentatività per alcuni sottogruppi.

La comparabilità tra le due specificazioni è parziale, in quanto alcune covariate presenti nel modello nazionale non sono osservabili nei dati regionali. In particolare, non sono disponibili le informazioni relative alla partecipazione a tirocini e alla qualifica professionale prevalente nei 24 mesi precedenti la DID. Inoltre, il trattamento dell'età differisce tra i due modelli. Nel modello nazionale l'età entra come variabile continua (con trasformazioni funzionali per cogliere non linearità) per l'effetto principale, ed è utilizzata anche in forma categoriale (classi di età) per le interazioni con altre variabili. Nel modello piemontese, invece, l'età è disponibile e viene utilizzata esclusivamente in forma categoriale, sia per l'effetto principale sia per le interazioni. Tali differenze nella disponibilità e nella definizione delle covariate possono influenzare la comparabilità dei coefficienti stimati tra i due modelli. In particolare, l'assenza di alcune informazioni presenti nel modello nazionale può determinare fenomeni di variabile omessa, che suggeriscono cautela nell'interpretazione delle differenze osservate tra i coefficienti.

Successivamente, il modello è stato esteso introducendo variabili di contesto riferite al Sistema locale del lavoro, con l'obiettivo di verificare se l'integrazione tra caratteristiche individuali e informazioni territoriali consenta di migliorare la capacità discriminante del profiling. In una fase ulteriore, l'analisi è stata affiancata dalla stima di modelli predittivi più flessibili, al fine di valutare se approcci di machine learning possano fornire performance superiori rispetto alla specificazione logistica utilizzata nel modello nazionale.

Confronto tra modelli

Poiché i modelli logit sono stimati su popolazioni differenti, la comparabilità diretta dei coefficienti, riportati nella tabella A1 in Appendice, può risultare limitata da differenze nella varianza non osservata. Inoltre, non essendo disponibili i microdati utilizzati per la stima del modello nazionale, non è stato possibile calcolare effetti marginali medi pienamente comparabili tra le due popolazioni. Il confronto deve pertanto essere interpretato in termini descrittivi e relativi, privilegiando il segno, la significatività e i pattern complessivi degli effetti. Coerentemente con la finalità del lavoro, orientata alla valutazione della capacità di profiling e di identificazione dei soggetti a maggior rischio di mancato reinserimento occupazionale, i risultati non vengono interpretati in termini causali. Per le dimensioni di maggiore interesse sono stati tuttavia calcolati effetti marginali per profili rappresentativi (MER, *Marginal effects at Representative values*), che consentono di valutare come la probabilità predetta di mancato reinserimento vari tra contesti territoriali per tipologie comparabili di individui.

Sulla base di questa impostazione, si confrontano di seguito gli effetti marginali calcolati per profili rappresentativi nei due contesti (tabella 1).

Nel complesso, il modello stimato sui dati piemontesi riproduce in larga parte le regolarità empiriche individuate da Anpal a livello nazionale, ma emergono differenze significative nell'intensità – e talvolta nel segno – degli effetti, che riflettono sia la diversa scala territoriale sia specificità strutturali del mercato del lavoro piemontese.

L'analisi degli effetti marginali con riferimento alla classe under 24 evidenzia un gradiente crescente dell'età per un uomo con le altre caratteristiche fissate al valore di riferimento in entrambi i modelli. Nel modello piemontese la penalizzazione cresce in modo progressivo e relativamente regolare, passando da circa 1 punto percentuale per i 25-29 anni a oltre 12 punti percentuali per gli over 50. Nel modello nazionale, invece, l'effetto è quasi nullo per i 25-29 anni ma aumenta rapidamente a partire dai 30 anni, raggiungendo circa 21 punti percentuali per gli over 50. L'età avanzata rappresenta dunque un fattore di rischio in entrambi i contesti, ma la penalizzazione risulta più intensa e più rapidamente crescente a livello nazionale.

Una dinamica analoga emerge con riferimento al genere. Nel modello piemontese l'effetto diretto dell'essere donna è contenuto (circa +1 punto percentuale in termini di MER), mentre nel modello nazionale risulta sensibilmente più marcato (circa +9 punti percentuali). Considerando la significatività e la struttura dei coefficienti stimati, nel contesto regionale le differenze di genere sembrano

manifestarsi principalmente attraverso interazioni con altre caratteristiche individuali, piuttosto che configurarsi come uno svantaggio generalizzato. Tuttavia, i MER calcolati per una donna di 30-39 anni con figli indicano che, anche tenendo conto dell'interazione tra genere e condizione familiare, in Piemonte l'effetto aggiuntivo della maternità è sostanzialmente nullo (inferiore a un punto

percentuale), mentre a livello nazionale raggiunge circa 4-5 punti percentuali. Il vincolo familiare appare dunque più penalizzante nel quadro nazionale, verosimilmente anche in relazione a differenze territoriali nei servizi di conciliazione e nelle strutture del mercato del lavoro, che possono attenuare o amplificare l'impatto della maternità sulle prospettive di reinserimento occupazionale.

Tabella 1. Effetti marginali su profili rappresentativi (MER) per il modello piemontese e nazionale¹³

Variabile/Profilo	Piemonte (pp)	Italia (pp)
<i>Genere e famiglia (30-39 anni)</i>		
Donna	+1,17	+8,67
Figli	-0,39	-1,20
Donna con figli	+0,72	+4,50
<i>Classi di età (rif. Under 24)</i>		
25-29	+1,20	+0,03
30-39	+2,64	+6,10
40-49	+6,31	+14,18
Over 50	+12,18	+20,92
<i>Esperienza lavorativa recente</i>		
40-49 con esperienza	-41,67	-9,20
Over 50 con esperienza	-42,54	-10,32
<i>Durata presenza in Italia (30-39 anni)</i>		
Fino a 12 mesi	-2,59	-0,56
Da 1 a 2 anni	-5,61	-1,52
Oltre 2 anni	-1,40	+1,42
Nato in Italia	+14,02	-2,0

Nota: Effetti marginali calcolati per profili rappresentativi (MER). I valori sono espressi in punti percentuali (pp) e indicano la variazione della probabilità di mancato reinserimento occupazionale associata alla caratteristica considerata; valori positivi corrispondono a una maggiore probabilità di mancato reinserimento. La variabile dipendente identifica la mancata occupazione successiva alla DID (1 = non occupato; 0 = occupato). Nel modello piemontese l'esito è osservato a 6 mesi dalla DID (N = 163433 DID sottoscritte in Piemonte nel 2022), mentre nel modello nazionale Anpal a 12 mesi dalla DID (circa 3,4 milioni di DID sottoscritte in Italia nel biennio 2018-2019).

Fonte: elaborazioni degli Autori; per il modello nazionale, calcoli propri a partire dai coefficienti riportati in Anpal (2022)

13 Gli effetti marginali sono calcolati variando una sola caratteristica rispetto a un profilo di riferimento. Il profilo di riferimento e le variazioni considerate sono specificati per ciascuna sezione della tabella. In tutte le sezioni, le variabili non esplicitamente menzionate sono fissate ai rispettivi valori di riferimento del modello.

- *Genere e famiglia (30-39 anni)*. Profilo di riferimento: uomo di 30-39 anni, cittadino italiano, senza figli, senza esperienza lavorativa recente.
 - Donna: confronto tra donna e uomo (a parità di età, senza figli e senza esperienza lavorativa recente).
 - Figli: confronto tra uomo con figli e uomo senza figli (a parità di età e senza esperienza lavorativa recente).
 - Donna con figli: confronto tra donna con figli e uomo senza figli (a parità di età e senza esperienza lavorativa recente).
- *Classi di età*. Profilo di riferimento: uomo di età inferiore a 25 anni (under 24), cittadino italiano, senza figli, senza esperienza lavorativa recente. Ogni riga mostra l'effetto di appartenere alla classe di età indicata rispetto al profilo di riferimento.
- *Esperienza lavorativa recente*. Profilo di riferimento: uomo di età inferiore a 25 anni (under 24), con esperienza lavorativa nei sei mesi precedenti la DID. Le righe riportano l'effetto dell'appartenenza alla classe di età indicata, in presenza di esperienza lavorativa recente, rispetto al profilo di riferimento.
- *Durata presenza in Italia (30-39 anni)*. Profilo di riferimento: uomo di 30-39 anni, cittadino italiano, senza figli, senza esperienza lavorativa recente. Le righe riportano l'effetto delle diverse durate di presenza in Italia (in mesi/anni) rispetto alla cittadinanza italiana (categoria di riferimento).

Le differenze tra i modelli risultano particolarmente evidenti con riferimento alla durata della presenza in Italia. Sia i coefficienti sia i MER indicano che, rispetto alla cittadinanza italiana, una presenza relativamente recente è associata a una maggiore probabilità di reinserimento occupazionale. Tuttavia, l'intensità dell'effetto è decisamente più marcata in Piemonte. Per un uomo di 30-39 anni, essere presente in Italia da 1 a 2 anni riduce la probabilità di mancato reinserimento di circa 5,6 punti percentuali nel modello regionale, contro 1,5 punti percentuali nel modello nazionale; effetti più contenuti emergono per permanenze inferiori a un anno. Inoltre, mentre nel modello piemontese la categoria "Nato in Italia" è associata a un aumento di circa 14 punti percentuali della probabilità di non reinserimento, nel modello nazionale l'effetto è di segno opposto e di entità modesta. Tale divergenza suggerisce che la variabile catturi nel contesto regionale dinamiche specifiche, plausibilmente legate a meccanismi di selezione migratoria: una quota rilevante di lavoratori stranieri entra infatti nel Paese con un impiego già attivo o con una forte pressione a reinserirsi rapidamente, anche in funzione del rinnovo del permesso di soggiorno.

La differenza più marcata tra i due modelli riguarda tuttavia l'esperienza lavorativa recente. I coefficienti piemontesi risultano molto più elevati in valore assoluto e i MER confermano un impatto estremamente forte: per un uomo di 40-49 anni, aver lavorato nei sei mesi precedenti la DID riduce la probabilità di mancato reinserimento di oltre 40 punti percentuali; un effetto analogo si osserva per gli over 50. Nel modello nazionale, dove l'esperienza è misurata su un orizzonte di dodici mesi, l'effetto si attesta intorno ai 9-10 punti percentuali. La diversa finestra temporale contribuisce certamente a spiegare lo scarto, ma il risultato indica anche che nel contesto piemontese la continuità occupazionale recente rappresenta un segnale particolarmente forte di occupabilità, soprattutto per le classi di età mature.

Per alcune dimensioni, quali il livello di istruzione e il settore di occupazione precedente, il numero elevato di categorie rende meno agevole una sintesi tramite effetti marginali su profili rappresentativi. Per tali dimensioni si fa pertanto riferimento ai coefficienti stimati, riportati nella tabella A1 in Appendice. Per quanto concerne il capitale umano, entrambi i modelli evidenziano che livelli di istruzione più elevati sono generalmente associati a una minore

probabilità di disoccupazione di medio periodo. Nel modello piemontese per i titoli universitari, sebbene i coefficienti non siano statisticamente significativi, si osserva un'inversione di segno, a differenza dell'andamento decrescente atteso. Queste evidenze appaiono controintuitive in ragione dei limiti informativi dei dati amministrativi regionali con i quali è stimato il modello. Infatti, nonostante la cooperazione applicativa consenta l'aggiornamento dei dati a livello nazionale e interregionale, ciascuna regione può accedere esclusivamente alle informazioni di propria pertinenza, secondo il principio della minimizzazione dei dati¹⁴. Ne deriva che un individuo laureato che, successivamente alla DID, trasferisca residenza e domicilio in un'altra regione – come avviene frequentemente nel caso di studenti fuori sede che rientrano nella regione d'origine – e vi trovi occupazione, non risulterà occupato nei dati piemontesi. Questo può determinare una sovrastima della disoccupazione per tale categoria di titolo di studio. Per le categorie con titolo di studio non superiore al diploma, tale meccanismo dovrebbe essere meno rilevante, dal momento che la mobilità interregionale associata ai percorsi formativi interessa soprattutto gli individui che intraprendono studi universitari.

Con riferimento al settore di occupazione precedente, entrambi i modelli mostrano che l'aver maturato un'esperienza lavorativa in specifici comparti è generalmente associato a una minore probabilità di mancato reinserimento rispetto ai soggetti privi di precedenti esperienze lavorative. Nel modello piemontese emerge una maggiore omogeneità nel segno dei coefficienti tra i diversi settori, suggerendo che il semplice mantenimento di un legame recente con il mercato del lavoro costituisca un fattore protettivo rilevante, indipendentemente dal comparto di provenienza. Le interazioni di genere confermano inoltre che alcuni settori presentano dinamiche differenziate tra uomini e donne, con effetti talvolta più accentuati nel contesto regionale.

In sintesi, il confronto tra modello nazionale e modello piemontese indica che le determinanti fondamentali della disoccupazione di medio periodo sono largamente condivise, ma l'intensità degli effetti varia sensibilmente a livello territoriale. Il modello piemontese mette in luce un ruolo particolarmente forte della storia lavorativa recente e di alcune caratteristiche specifiche del contesto locale, mentre nel modello nazionale emergono penalizzazioni più

14 Per la Regione Piemonte i casi accessibili sono:

- (i) cittadino residente/domiciliato in Piemonte con COB in azienda localizzata in Piemonte;
- (ii) cittadino residente/domiciliato in altra regione con COB in azienda localizzata in Piemonte;
- (iii) cittadino residente/domiciliato in Piemonte con COB in azienda localizzata in altra regione.

Tabella 2. Coefficienti stimati per le variabili di contesto selezionate

Variabile	Coefficiente	Significatività
<i>Indipendenti delle imprese individuali per 100 indipendenti</i>	0,0459	***
<i>Indipendenti delle microimprese per 100 indipendenti</i>	-0,1332	***
<i>Dipendenti delle imprese ad alta tecnologia e ad alta intensità di conoscenza per 100 dipendenti</i>	0,0241	***
<i>Indipendenti per 100 residenti (25-34 anni)</i>	0,5093	***
<i>Indipendenti per 100 residenti (totale)</i>	-0,4751	***
<i>Indipendenti 25-34 anni per 100 indipendenti</i>	-0,4547	***
<i>Addetti 25-34 anni per 100 addetti</i>	0,1766	***
<i>Addetti con cittadinanza straniera per 100 addetti (25-64 anni)</i>	0,0192	**
<i>Femmine per 100 indipendenti (25-64 anni, istruzione terziaria)</i>	0,0074	*
<i>Indipendenti per 100 residenti (25-64 anni)</i>	-0,1023	

Nota: Coefficienti stimati mediante modello logit. La specificazione include anche le variabili individuali discusse nel paragrafo 4 (non riportate in tabella). Le variabili di contesto sono state selezionate mediante procedura di best subset selection basata sul criterio AIC. Per la definizione della variabile dipendente e del campione si rimanda alla nota della tabella 1. Gli asterischi indicano il livello di significatività statistica: *** $p < 0,01$; ** $p < 0,05$; * $p < 0,10$.

Fonte: elaborazione degli Autori su dati Agenzia Piemonte Lavoro

marcate legate a età, genere e condizione familiare. Nel complesso, i risultati suggeriscono che le politiche di reinserimento dovrebbero essere non solo personalizzate rispetto al profilo individuale, ma anche calibrate sulle specificità strutturali dei mercati del lavoro regionali.

3. Modello con variabili di contesto

Per approfondire il ruolo delle caratteristiche territoriali nel determinare le probabilità di reinserimento occupazionale, è stato stimato un ulteriore modello logistico che integra, alle variabili individuali già considerate, un insieme di indicatori strutturali riferiti ai Sistemi locali del lavoro (SLL). Gli indicatori territoriali sono tratti dal Sistema informativo sui SLL dell'Istat e derivano dall'integrazione di più registri statistici ufficiali (Istat 2025). In particolare: gli addetti, dipendenti e indipendenti provengono dal Registro statistico delle imprese attive (ASIA); le caratteristiche individuali degli occupati (età, cittadinanza, istruzione) derivano dal Registro integrato degli occupati (RBI); i residenti sono tratti dalla popolazione residente (Anagrafe e bilancio demografico). Gli indicatori utilizzati nel modello (ad esempio, "addetti stranieri per 100 addetti" o "indipendenti per 100 residenti") sono pertanto costruiti da Istat mediante elaborazioni su registri amministrativi integrati¹⁵.

L'obiettivo è verificare se, a parità di

caratteristiche sociodemografiche e professionali, il contesto produttivo locale incida in modo significativo sulla probabilità di non rientrare nel mercato del lavoro entro sei mesi dalla sottoscrizione della DID. Mentre il confronto precedente valutava la robustezza dei risultati nel passaggio dalla scala nazionale a quella regionale, questa estensione verifica l'eterogeneità interna al Piemonte. In questa prospettiva, il profiling non può essere inteso come uno strumento neutrale rispetto al territorio, ma deve essere sensibile alle specificità strutturali dei diversi mercati locali, al fine di orientare in modo più mirato le politiche attive.

Per la selezione delle variabili di contesto è stato adottato il metodo della *best subset selection*, utilizzando come criterio di scelta l'*Akaike Information Criterion* (AIC). L'AIC è stato preferito ad altri criteri in quanto, selezionando le specificazioni con migliore capacità predittiva fuori campione, risulta particolarmente adatto per finalità predittive (Akaike 1974; Burnham e Anderson 2002). L'applicazione della procedura ha portato alla selezione di 10 indicatori territoriali, i cui coefficienti stimati sono riportati in tabella 2. Rispetto al modello con sole variabili individuali, il modello arricchito con gli indicatori SLL presenta un AIC sensibilmente più basso, a conferma del miglioramento della capacità predittiva.

L'inclusione di tali variabili evidenzia che il contesto produttivo locale esercita un effetto

15 Per le definizioni di addetto, dipendente e indipendente si rimanda a Istat, Glossario statistico, <https://www.istat.it/classificazioni-e-strumenti/glossario/>.

autonomo sulla probabilità di non reinserimento. In particolare, una maggiore quota di addetti stranieri è associata a una più elevata probabilità di non reinserimento, il che potrebbe riflettere una maggiore competizione nei segmenti occupazionali a bassa qualificazione nei territori caratterizzati da una forte presenza di lavoratori stranieri. Analogamente, una maggiore incidenza di dipendenti in imprese *hightech* e *knowledgeintensive* è positivamente associata alla probabilità di non rientro: ciò può indicare un disallineamento tra la struttura della domanda locale – orientata verso competenze elevate – e il profilo medio degli individui che rilasciano la DID, potenzialmente meno qualificati.

Più articolata è l'interpretazione degli indicatori di autoimpiego. Una maggiore quota complessiva di *indipendenti per 100 residenti* presenta un coefficiente negativo, suggerendo che nei territori in cui l'autoimpiego costituisce una componente strutturale consolidata del tessuto produttivo il rischio di esclusione dal mercato del lavoro risulti inferiore. Al contrario, un'elevata incidenza di *indipendenti tra i giovani (25-34 anni)* è positivamente associata alla probabilità di non reinserimento, indicando che quando l'autoimpiego è concentrato nelle coorti più giovani esso può riflettere difficoltà di accesso al lavoro dipendente piuttosto che una scelta imprenditoriale consolidata. In questo senso, l'autoimpiego appare 'protettivo' quando esprime una struttura economica matura, ma può essere associato a maggiore vulnerabilità quando rappresenta una strategia residuale di ingresso nel mercato del lavoro.

Si precisa che l'interpretazione dei coefficienti delle variabili di contesto è intrinsecamente complessa e non può essere intesa in senso causale, trattandosi di associazioni stimate in un modello con finalità predittive su dati osservazionali. Nel complesso, i risultati confermano che il reinserimento occupazionale non dipende esclusivamente dalle caratteristiche individuali, ma è fortemente influenzato dalla configurazione produttiva e demografica dei mercati locali. Ne discende una chiara implicazione di policy: gli strumenti di profiling dovrebbero incorporare in modo sistematico indicatori territoriali, così da orientare l'allocazione delle politiche attive non solo in funzione del profilo individuale, ma anche delle opportunità e dei vincoli propri dei diversi contesti locali. Un approccio multilivello – nazionale, regionale e subregionale (SLL) – appare pertanto essenziale per garantire interventi più efficaci e coerenti con le specificità dei mercati del lavoro.

4. Approcci predittivi avanzati: Random Forest e Boosting

Poiché l'obiettivo dell'analisi non è soltanto interpretativo ma anche operativo – ossia volto

alla costruzione di uno strumento di profilazione utile a orientare le politiche attive – l'attenzione si estende oltre la regressione logistica verso approcci maggiormente orientati alla previsione. In questa prospettiva, accanto al modello logistico sono stati stimati modelli di machine learning, in particolare Random Forest e Boosting, con l'obiettivo di massimizzare la capacità predittiva anche a costo di una minore interpretabilità.

Un primo elemento che emerge dal confronto riguarda il livello complessivo di accuratezza (*accuracy*), intesa come quota di osservazioni correttamente classificate dal modello rispetto al totale dei casi considerati (tabella 3). Il modello Anpal stimato su dati nazionali presenta un'accuratezza pari al 67,3%, mentre i modelli stimati su dati piemontesi (logistico, con variabili di contesto, Random Forest e Boosting) si collocano su valori compresi tra il 63% e il 64,5%. Tale differenza può essere ricondotta a diversi fattori. In primo luogo, la diversa definizione della variabile dipendente: il modello Anpal considera un orizzonte temporale di 12 mesi, mentre nel caso piemontese l'esito è osservato a 6 mesi dal rilascio della DID. Un orizzonte più lungo potrebbe intercettare condizioni di disoccupazione più strutturate e persistenti, potenzialmente più facili da prevedere. Al contrario, dinamiche di breve periodo sono maggiormente influenzate da fattori congiunturali e dunque meno facilmente anticipabili.

Un ulteriore elemento riguarda la diversa disponibilità informativa. Alcune variabili incluse nel modello nazionale – in particolare quelle relative allo svolgimento di tirocini e alla qualifica professionale prevalente nei 24 mesi precedenti la DID – non trovano corrispondenza nei dati regionali e non sono quindi state incluse nel modello piemontese. Si tratta di informazioni potenzialmente rilevanti per cogliere aspetti della storia lavorativa recente, la cui assenza può incidere sulla capacità predittiva complessiva. Un'ulteriore differenza riguarda il trattamento dell'età. Nel modello nazionale essa è specificata come variabile numerica ed è oggetto di trasformazioni funzionali, consentendo di stimare in modo flessibile eventuali effetti non lineari. Nei dati piemontesi, invece, l'età è disponibile solo in forma categoriale, con conseguente perdita di informazione. La discretizzazione implica infatti che individui appartenenti alla stessa classe siano trattati in modo omogeneo e non consente di intercettare variazioni marginali interne all'intervallo considerato, incidendo potenzialmente sulla performance predittiva.

Se si considera la *sensibilità (sensitivity)* – ovvero la quota di soggetti rimasti disoccupati a sei mesi dal rilascio della DID e correttamente identificati

Tabella 3. Performance predittiva dei modelli di profiling

Metodo	Accuracy out-of-sample	Sensitivity	Specificity
Modello dati nazionali	67,63%	55,83%	76,76%
Modello dati piemontesi (senza var. di contesto)	63,56%	41,63%	80,32%
Modello dati piemontesi con denominazione SLL	64,12%	43,76%	79,68%
Modello dati piemontesi Best Subset Selection	64,14%	43,74%	79,72%
Modello dati piemontesi Forward Stepwise Selection	64,10%	43,71%	79,68%
Modello dati piemontesi Random Forest	64,00%	45,76%	77,93%
Modello dati piemontesi Boosting	64,50%	47,57%	77,44%

Nota: Per la definizione della variabile dipendente e del campione si rimanda alla nota della tabella 1. L'accuratezza misura la quota di osservazioni correttamente classificate; la sensibilità la quota di soggetti non occupati correttamente identificati come Weaker; la specificità la quota di soggetti Work Ready correttamente identificati.

Fonte: elaborazioni degli Autori per il modello piemontese; Anpal (2022) per il modello nazionale

dal modello – il modello Anpal raggiunge un valore del 55,8%. Ciò significa che poco più della metà dei soggetti effettivamente in stato di disoccupazione di lungo periodo viene classificata correttamente come “Weaker forte” (Anpal 2022). Pur rappresentando un risultato rilevante, questa performance evidenzia anche un ampio margine di miglioramento, in particolare se l'obiettivo è intercettare tempestivamente i soggetti che richiedono misure di sostegno più strutturate e personalizzate.

Nei modelli stimati su dati piemontesi la sensibilità è più bassa, in linea con la minore accuratezza complessiva. Il modello logistico più simile a quello Anpal a livello di predittori (senza variabili di contesto) presenta una sensibilità del 41,6%. Tuttavia, già l'introduzione della sola variabile di appartenenza all'SLL determina un incremento di circa due punti percentuali (sensibilità pari al 43,8%), pari a un incremento di circa il 5%. Un miglioramento analogo si osserva includendo le variabili di contesto selezionate a livello di SLL. Questo risultato suggerisce che la dimensione territoriale, anche in forma sintetica, contribuisce in modo significativo alla capacità di intercettare i soggetti più fragili.

Il passaggio a modelli di machine learning rafforza ulteriormente tale evidenza. Il modello Random Forest ottiene una sensibilità del 45,8%, mentre il Boosting raggiunge il 47,6%, con un incremento di circa il 14% rispetto al modello logistico senza variabili di contesto. Per questi ultimi modelli l'aumento dell'accuratezza complessiva è invece più contenuto. Considerando che, in un'ottica di policy,

può essere prioritario ridurre il numero di soggetti fragili non riconosciuti, il miglioramento della sensibilità rappresenta un risultato particolarmente rilevante. In questa prospettiva, l'adozione di strumenti predittivi più flessibili rispetto al modello logistico tradizionale potrebbe rafforzare l'efficacia del sistema di profilazione.

Per quanto riguarda la *specificità* (*specificity*) – ovvero la quota di soggetti “Work-Ready” correttamente identificati dal modello – i modelli piemontesi mostrano valori generalmente superiori rispetto al modello nazionale Anpal. Il modello logistico più semplice, senza variabili di contesto, supera l'80% di specificità. Questo aspetto non è secondario: una buona capacità di riconoscere i soggetti “Work-Ready” consente di evitare l'allocazione inefficiente di risorse per accompagnare individui che presentano già elevate probabilità di reinserimento autonomo.

Nel complesso, i risultati mostrano un trade-off tra interpretabilità e performance predittiva. Tale aspetto assume particolare rilevanza nel contesto delle politiche attive del lavoro, dove gli strumenti di profiling orientano decisioni che riguardano l'impiego di risorse pubbliche e l'individuazione dei percorsi di accompagnamento al lavoro. In questo contesto, la trasparenza del modello rappresenta una caratteristica importante, poiché consente di ricostruire e comunicare i fattori associati al rischio di mancato reinserimento occupazionale. Il modello logistico rimane quindi uno strumento trasparente e facilmente comunicabile, la cui performance

predittiva migliora ulteriormente quando integra variabili di contesto territoriale. L'adozione di algoritmi di machine learning, come Random Forest e Boosting, consente invece di migliorare la capacità di intercettare i soggetti a maggiore rischio di esclusione prolungata dal mercato del lavoro. In un sistema di politiche attive orientato all'efficacia e alla personalizzazione degli interventi, tale miglioramento può tradursi in un impatto concreto sull'allocazione delle risorse e sulla qualità dei servizi offerti.

Conclusioni

Il presente lavoro ha esplorato le potenzialità dell'integrazione tra dati amministrativi individuali e informazioni territoriali per il profiling delle persone disponibili al lavoro, con l'obiettivo di migliorare la capacità predittiva dei modelli attualmente utilizzati a livello nazionale. L'analisi, condotta sui dati piemontesi della Scheda anagrafico-professionale (SAP), ha preso avvio dalla replica del modello logistico Anpal, confermando le principali regolarità empiriche già note, ed evidenziando al contempo differenze significative nella rilevanza dei vari predittori dello stato di disoccupazione, riconducibili alle specificità strutturali del mercato del lavoro regionale.

L'estensione del modello con variabili di contesto riferite ai Sistemi locali del lavoro (SLL) evidenzia il ruolo rilevante delle condizioni socio-economiche locali nella probabilità di mancato reinserimento, al netto delle caratteristiche individuali. Indicatori quali la quota di personale con cittadinanza non italiana, l'incidenza di dipendenti in imprese ad alta tecnologia e la struttura dell'autoimpiego contribuiscono a spiegare variazioni territoriali nella vulnerabilità delle persone disoccupate. L'utilizzo di tecniche di selezione delle variabili ha consentito di individuare un insieme ristretto di indicatori territoriali associati a un miglioramento della capacità predittiva del modello.

Il ricorso a modelli di machine learning – Random Forest e Boosting – ha ulteriormente incrementato la capacità di intercettare i soggetti a maggior rischio di esclusione prolungata dal mercato del lavoro (sensibilità). In un'ottica operativa, questo miglioramento può tradursi in una più efficace allocazione delle risorse pubbliche, orientando le politiche attive verso i beneficiari che necessitano di interventi più strutturati e personalizzati, da bilanciare però con la minore interpretabilità e trasparenza di questi modelli rispetto a modelli statistici tradizionali.

Nel complesso, i risultati suggeriscono che i sistemi di profiling non possono prescindere dalla dimensione territoriale. Un approccio multilivello – nazionale, regionale e subregionale – appare essenziale per garantire che gli strumenti di

profilazione siano sensibili alle specificità dei mercati locali, contribuendo così a una presa in carico più mirata ed efficiente. Le evidenze emerse supportano la tesi secondo cui l'integrazione sistematica di indicatori di contesto nei modelli di profilazione può rafforzare il supporto alle decisioni dei Servizi per l'impiego, in coerenza con l'obiettivo di personalizzazione degli interventi proposto dal Programma GOL.

I risultati non evidenziano una maggiore qualità informativa dei dati piemontesi rispetto a quelli nazionali, che risultano anzi più completi, come mostrato in Appendice. Il miglioramento delle performance predittive è riconducibile all'introduzione di variabili territoriali (SLL), più che a differenze nella struttura dei dati individuali. Ne consegue che un'estensione del modello nazionale che includa analoghi indicatori di contesto appare tecnicamente realizzabile.

Al contempo, il confronto tra i coefficienti dei modelli nazionale e regionale suggerisce che l'effetto delle caratteristiche individuali possa variare a seconda del contesto territoriale, aprendo la possibilità di sviluppare specificazioni regionali del modello. Questa ipotesi merita ulteriori approfondimenti, anche alla luce dell'elevata numerosità campionaria disponibile a livello nazionale.

Il lavoro presenta alcune limitazioni. In primo luogo, la diversa finestra temporale utilizzata per definire l'esito (6 mesi su dati piemontesi, 12 mesi nel modello Anpal su dati nazionali) limita la comparabilità diretta delle performance. In secondo luogo, l'impossibilità di disporre di alcune variabili presenti nel modello nazionale (tirocini, qualifica professionale pregressa) e la differente specificazione dell'età potrebbero aver influenzato i risultati.

Ciononostante, lo studio fornisce evidenze empiriche a sostegno di un'evoluzione del profiling verso modelli ibridi, capaci di integrare informazioni individuali e territoriali con tecniche statistiche flessibili, senza rinunciare a un'adeguata trasparenza. Future ricerche potrebbero estendere l'analisi ad altre regioni italiane, adottando un disegno comparativo che consenta di generalizzare i risultati.

Appendice

La tabella A1 riporta i coefficienti stimati per i due modelli (nazionale Anpal (2022) e regionale senza variabili di contesto – dati forniti da APL - Agenzia Piemonte Lavoro), relativamente alle variabili individuali comuni. I modelli stimano la probabilità di mancato reinserimento lavorativo successivo alla sottoscrizione della DID. La variabile dipendente è pari a 1 se l'individuo non risulta occupato dopo la DID, 0

altrimenti. Nel modello piemontese l'esito è osservato a 6 mesi dalla DID (N = 163433 DID sottoscritte in Piemonte nel 2022); nel modello nazionale Anpal l'esito è osservato a 12 mesi dalla DID (circa 3,4 milioni di DID sottoscritte in Italia nel 2018-2019).

Nella tabella emergono anche le differenze tra le variabili esplicative considerate, tenendo conto che, per il modello piemontese, si è cercato di ricondurre – per quanto possibile – le variabili regionali a quelle

del modello Anpal. Si ricorda inoltre che nel modello piemontese mancano completamente le informazioni relative alla partecipazione a tirocini e alla qualifica professionale prevalente nei 24 mesi precedenti la DID, variabili invece presenti nel modello nazionale.

Le categorie di riferimento sono indicate tra parentesi. Gli asterischi denotano i livelli di significatività statistica (*p<0,10; **p<0,05; ***p<0,01).

Tabella A1. Coefficienti completi del modello nazionale (Anpal) e del modello piemontese (APL)

ANPAL Variabile (cat. rif.)	APL Variabile (cat. rif.)	ANPAL modalità	APL modalità	ANPAL Coef.	APL Coef.
Genere		Donna	Donna	0,357***	0,023
Classi Età (cat. rif. inferiore a 24 anni)		Da 25 a 29 anni	Da 25 a 29 anni	0,061***	0,060**
		Da 30 a 39 anni	Da 30 a 39 anni	0,310***	0,160***
		Da 40 a 49 anni	Da 40 a 49 anni	0,483***	0,390***
		Over 50	Over 50	0,451***	0,847***
Interazione donna e classe d'età (cat. rif. uomo con età inferiore a 24 anni)		Donna - 25 a 29 anni	Donna - 25 a 29 anni	0,083***	0,042
		Donna - 30 a 39 anni	Donna - 30 a 39 anni	0,123***	0,045
		Donna - 40 a 49 anni	Donna - 40 a 49 anni	-0,100***	-0,162***
		Donna - Over 50 anni	Donna - Over 50 anni	-0,130***	-0,209***
Durata presenza in Italia (cat. rif. Cittadinanza italiana)		Fino a 12 mesi	Fino a 12 mesi	-0,111***	-0,113***
		Da 1 a 2 anni	Da 1 a 2 anni	-0,074***	-0,292***
		Oltre 2 anni	Oltre 2 anni	0,072***	-0,067***
		Nato in Italia	Nato in Italia	0,083*	1,194***
Stato occupazionale alla firma		Occupato alla DID	Occupato alla DID	0,200***	0,144***
Precedente esperienza di lavoro		Ha lavorato nei 12 mesi pre DID	Ha lavorato nei 6 mesi pre DID	-0,322***	-1,842***
Interazione classe d'età e precedente esperienza di lavoro nei 12 mesi pre DID	Interazione classe d'età e precedente esperienza di lavoro nei 6 mesi pre DID	Precedente esperienza - 25 a 29 anni	Precedente esperienza - 25 a 29 anni	0,085***	-0,001
		Precedente esperienza - 30 a 39 anni	Precedente esperienza - 30 a 39 anni	-0,125***	-0,056
		Precedente esperienza - 40 a 49 anni	Precedente esperienza - 40 a 49 anni	-0,498***	-0,119***
		Precedente esperienza - over 50 anni	Precedente esperienza - over 50 anni	-0,696***	-0,364***

segue

segue tabella A1

Titolo di studio (cat. rif. Fino licenza elementare)	Titolo di studio (cat. rif. Nessun titolo)	Licenza elementare	Licenza elementare	-	-0,226**
		Licenza media	Scuola dell'obbligo	-0,073***	-0,177*
		Liceo		-0,024***	
		Qualifica professionale	Istruzione professionale	-0,253***	-0,247**
		Istituto professionale		-0,281***	
Istituto tecnico	-0,188***				
Diploma o Triennale	Titolo di studio (cat. rif. Nessun titolo)	Diploma	Diploma	-0,175***	-0,101
		Scienze umanistiche	Laurea I o II livello	-0,356***	0,015
		Scienze sociali		-0,200***	
		Scienze della salute		-0,514***	
		Ingegneria, informatica e trasporti		-0,547***	
		Altro diploma o triennale		-0,239***	
		Scienze umanistiche	-0,558***	Corsi post laurea	0,127
		Scienze sociali	-0,257***		
		Scienze della salute	-0,666***		
		Ingegneria, informatica e trasporti	-0,741***		
Scienze naturali	-0,571***				
Architettura	-0,042**				
Altra laurea magistrale, specialistica, vecchio ordinamento	-0,348***				
Patente		Possedere una patente	Possedere una patente	-0,189***	-0,595***
		Occupato	Occupato	-0,007	-0,410***
		In cerca di nuova occupazione	In cerca di nuova occupazione	-0,331***	-0,322***
		In cerca di prima occupazione	In cerca di prima occupazione	0,066***	-0,012
		Studente	Studente	-0,141***	-0,056*
Settore prevalente 24 mesi precedenti la firma della DID (cat. rif. Non ha lavorato)	Settore prevalente 6 mesi precedenti la firma della DID (cat. rif. Non ha lavorato)	Agricoltura	Agricoltura	-0,823***	-0,410***
		Industria in senso stretto	Industria	0,023**	-0,454***
		Costruzioni	Costruzioni	-0,216***	-0,003
		Commercio	Commercio	0,121***	-0,083
		Alloggio e ristorazione	Alloggio e ristorazione	-0,451***	-0,235***
		Trasporto, immagazzinaggio e altri servizi	Servizi	-0,140***	-0,364***

segue

segue tabella A1

Interazione donna e settore prevalente 24 mesi precedenti (cat. rif. uomo - non ha lavorato)	Interazione donna e settore prevalente 6 mesi precedenti (cat. rif. uomo - non ha lavorato)	Donna - Agricoltura	Donna - Agricoltura	-0,197***	0,157
		Donna - Industria in senso stretto	Donna - Industria	-0,088***	-0,208***
		Donna - Costruzioni	Donna - Costruzioni	0,619***	0,419***
		Donna - Commercio	Donna - Commercio	-0,134***	-0,076
		Donna - Alloggio e ristorazione	Donna - Alloggio e ristorazione	-0,051***	0,006
		Donna - Trasporto, immagazzinaggio e altri servizi	Donna - Servizi	0,014*	0,022
Numero datori cambiati nei 24 mesi precedenti la firma della DID (cat. rif. Non ha lavorato nel periodo)	Numero contratti avuti nei 6 mesi precedenti la firma della DID (cat. rif. Nessuno)	1 datore di lavoro	1 lavoro	0,863***	1,458***
		2 datori di lavoro	2-5 lavori	0,497***	0,792***
		Più di 2 datori di lavoro	5-10 lavori		0,000
			10 o più lavori	0,000	-0,402*
Impegno familiare		Famiglia con figli	Presenza di figli a carico	-0,058***	-0,055**
Interazione donna e impegno familiare		Donna con figli	Donna - Presenza di figli a carico	0,373***	0,096***
Numero componenti famiglia (cat. rif. Unico componente)		2	2	0,020***	0,003
		3	3	-0,063***	-0,043**
		4	4	-0,138***	-0,032*
		5	5	-0,102***	-0,002
		Più di 5	Più di 5	0,009	0,027
Costante				1,650***	1,076***

Fonte: stime degli Autori per il modello piemontese; coefficienti riportati in Anpal (anno) per il modello nazionale

Bibliografia

- Agenzia per l'Italia digitale (AgID) (2021), *Conservazione dei documenti digitali*, Roma, Agid
- Akaike H. (1974), A new look at the statistical model identification, *IEEE Transactions on Automatic Control*, 19, n.6, pp.716-723
- Bazzoli M., Marzadro S., Trivellato U. (2018), *Un'esperienza pilota di integrazione di dati amministrativi e di survey per l'analisi dell'evoluzione delle storie lavorative dei giovani*, FBK-IRVAPP Working Papers n.2018-01, Trento, Fondazione Bruno Kessler-IRVAPP
- Bollens J. (2022), *Evaluatie van de effectiviteit van diverse activerende arbeidsmarkt-maatregelen bij VDAB*, Brussel, VDAB
- Burnham K.P., Anderson D. R. (2002), *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, New York, Springer, 2nd ed
- Chetty R. (2012), *Time trends in the use of administrative data for empirical research*, 34th Annual NBER Summer Institute, Cambridge, Massachusetts
- Desiere S., Langenbucher K., Struyven L. (2019), *Statistical profiling in public employment services: An international comparison*, OECD Social, Employment and Migration Working Papers n.224, OECD Publishing, Paris
- Dipartimento della Funzione Pubblica (2024), *6° Piano d'Azione nazionale per il governo aperto 2024-2026*, Roma
- Istat (2025), *I Sistemi locali del lavoro 2021*, Focus, Roma, Istat

- Istat (2014), *Progetto ARCHIMEDE. I risultati sperimentali e le linee di sviluppo*, Istat Working Papers, Roma, Istat
- Leombruni R., Ardito C. (2026), The long and winding road. The access to administrative data for research purposes in Italy, in Berton F., Paruolo P. (eds.), *Data-Driven Learning in the EU. From Administrative Sources to Public Value*, Cham, Springer
- Leombruni R. (2024), Interviewing administrative records: A conceptual map for the use of big data for economic research, *Statistica Applicata. Italian Journal of Applied Statistics*, 36, n.3, pp.295-326
- Loxha A., Morgandi M. (2014), *Profiling the unemployed. A review of OECD experiences and implications for emerging economies*, Social Protection and labor discussion paper n.SP1424, Washington, World Bank
- O'Connell P. J., McGuinness S., Kelly E., Walsh J. (2009), *National Profiling of the Unemployed in Ireland. Research Report*, Dublin, Economic and Social Research Institute
- Trivellato U. (2019), Microdata for social sciences and policy evaluation as a public good, in Crato N., Paruolo P. (eds.), *Data-Driven Policy Impact Evaluation*, Cham, Springer, pp.27-45

Monica Bosia

mbosia@agenziapiemontelavoro.it

Analista presso Agenzia Piemonte Lavoro – Servizio Monitoraggio, Studi e Ricerche. Laureata in Scienze politiche, sociali e dell'amministrazione presso l'Università del Piemonte Orientale, dal 2016 si occupa di analisi di dati amministrativi, data mining e analisi delle politiche attive del lavoro a fini di monitoraggio e rendicontazione.

Marianna Carlone

marianna.carlone@edu.unito.it

Laureata magistrale in Metodi statistici ed economici per le decisioni presso il Dipartimento di Economia e statistica "Cognetti de Martiis", Università degli Studi di Torino.

Elisa Floredan

efloredan@agenziapiemontelavoro.it

Specialista delle politiche del lavoro presso Agenzia Piemonte Lavoro – Servizio Monitoraggio, Studi e Ricerche. Laureata magistrale in Sociologia presso l'Università degli Studi di Torino, dal 2020 svolge attività di ricerca quali-quantitativa sui mercati del lavoro e sulle politiche attive territoriali, occupandosi della progettazione e della reportistica periodica per il monitoraggio e la definizione degli scenari locali del lavoro.

Francesca Prever

francesca.prever@unito.it

Borsista di ricerca presso il Dipartimento di Economia e statistica "Cognetti de Martiis", Università degli Studi di Torino.