

IL TRATTAMENTO DEI DATI AMMINISTRATIVI A SOSTEGNO DELL'INFORMAZIONE
STATISTICA: ALCUNI ESPERIMENTI BASATI SUI DATI DELLA TOSCANA
NELL'AMBITO DEL PROGETTO TREND.

Luca Faustini¹, Andrea Brancatello², Tommaso Rondinella³

SOMMARIO

TREND è un progetto nato dalla collaborazione tra Istat e CNA che si propone di sfruttare i dati territoriali raccolti a fini amministrativi sulle piccole imprese (fino a 19 addetti) per ottenere stime trimestrali tendenziali [t, t-4], con grande tempestività e dettaglio territoriale, sull'andamento economico dei settori economici di interesse. Il dato di partenza presenta tuttavia i tipici limiti dei data set amministrativi - errori di classificazione e/o normalizzazione, imperfezioni di misurazione (ad es. dati cumulati), dati mancanti nei singoli record - determinando una certa instabilità delle stime. Al fine di contenere queste problematiche è stato necessario da un lato recuperare la maggior informazione possibile dal data set con procedure di *data cleaning*, dall'altro procedere all'imputazione dei dati mancanti. Dopo un'introduzione che descrive il progetto TREND verrà descritto il data set di partenza (§2); il processo di *data cleaning* (§3); l'imputazione dei dati mancanti e la sua diagnosi (§4); una valutazione dei risultati raggiunti (§5); conclusioni e passi successivi (§6).

1. Introduzione

L'attuale processo di modernizzazione dell'Istat pone al centro della ristrutturazione dell'Istituto la creazione di 4 registri - persone, unità economiche, luoghi ed eventi - dove convergeranno in modo sistematico le informazioni provenienti sia dalle indagini che dagli archivi amministrativi (in linea con quanto previsto dalla ESSVision2020 di Eurostat), influenzando l'intero processo di produzione statistica. Il pieno sfruttamento dell'informazione da fonte amministrativa riveste quindi un ruolo di crescente rilevanza per la Statistica Ufficiale. In quest'ottica si inserisce anche il progetto TREND. Nato attraverso una serie di convenzioni stipulate tra Istat e le sedi regionali della Confederazione Nazionale dell'Artigianato e della Piccola e Media Impresa (CNA) di Emilia Romagna, Marche, Toscana e Umbria si propone di sfruttare i dati

¹ ISTAT – RMC, via dell'Agnolo 80, 50122, Firenze (FI), e-mail: faustini@istat.it (corresponding author).

² ISTAT – RMC, via dell'Agnolo 80, 50122, Firenze (FI), e-mail: brancate@istat.it.

³ ISTAT – RMC, via dell'Agnolo 80, 50122, Firenze (FI), e-mail: rondinella@istat.it

territoriali raccolti a fini prevalentemente fiscali sulle piccole imprese (fino a 19 addetti) per ottenere stime trimestrali tendenziali per provincia e settore d'attività. Attraverso l'accordo di partenariato, CNA mette a disposizione i dati di bilancio depositati dalle imprese aderenti presso le sedi provinciali dell'associazione a sei mesi dal periodo di riferimento.

La successiva elaborazione dei dati permette di produrre statistiche sull'andamento dei principali settori economici nelle provincie interessate con grande tempestività. Il data-set per la Toscana, che prendiamo in considerazione in questa sede, contiene informazioni per circa 11mila imprese in ogni trimestre. Contiene tanto variabili identificative, o di stratificazione (provincia, anno, trimestre, partita IVA, Ateco 2007, numero di addetti) che alcune variabili economiche, o di analisi (ricavi, investimenti, retribuzioni, consumi, assicurazioni). Le stime prodotte nell'ambito del progetto riguardano esclusivamente i tassi trimestrali di variazione tendenziale e non i livelli assoluti delle variabili. Il confronto nella struttura dimensionale tra le imprese incluse nel data-set e quelle presenti in ASIA mostra un sostanziale allineamento per i domini di stima individuati (ovvero settore x provincia), avvallando l'ipotesi di correttezza del campione (Palmieri 2013). La fonte utilizzata risente tuttavia di una serie di limiti dovuti alla non omogenea copertura di settori e territori e di imprecisioni nelle variabili di stratificazione (Ateco incompleti, partite iva duplicate o mancanti, provincie di altre regioni) che comportano significative perdite in termini di numerosità campionaria.

In questo lavoro si presentano una serie di procedure di pulizia dei dati e imputazione dei record mancanti al fine di irrobustire l'informazione di base, aumentare la numerosità campionaria e produrre stime più stabili nel tempo, in particolare per quanto riguarda la variabile di stima più rilevante del data-set, ovvero i ricavi. L'intero lavoro è stato realizzato con il software R; per l'imputazione è stato utilizzato il pacchetto AMELIA II.

2. Descrizione del progetto TREND

La procedura di definizione delle stime è effettuata partendo dalla disponibilità dei valori di bilancio delle imprese aderenti al CNA che hanno depositato i propri dati presso l'associazione con lo scopo di ricevere assistenza riguardo una pluralità di servizi (versamento dell'IVA, buste paga, dichiarativi, confidi ecc.). Le imprese sono inizialmente riclassificate secondo quattro classi d'addetti (1, 2-5, 6-19, oltre 20) al fine di produrre stime relative solo a quelle sotto i 20 addetti.

I settori d'attività sono disaggregati fino ad un massimo di 13 settori regionali⁴. Le imprese sono quindi raggruppate per domini d'analisi rappresentati come intersezione tra provincia e settore economico. Ogni dominio è suddiviso in tre strati secondo le diverse classi d'addetti. Il totale degli strati disponibili è 351 costituito dalle nove province (per Siena non sono disponibili i dati di bilancio), dai tredici settori economici e dalle tre classi di addetti. Ogni trimestre le informazioni vengono raccolte secondo un implicito schema di campionamento che può essere descritto come Cross-Sectional Time Series (Baltagi 2005). Al fine di stimare le variazioni tendenziali delle variabili di interesse, il campione di ogni trimestre t è limitato a quelle imprese che compaiono anche al trimestre $t-4$ andando così a determinare dei campioni panel (chiamati nel corso del testo "panel trimestrali") che si modificano di trimestre in trimestre.

Sono escluse dal campione tutte le imprese la cui informazione a corredo non consente una collocazione univoca negli strati o che non permette un aggancio con il trimestre $t-4$. Non risultano quindi eleggibili quelle imprese:

- prive di partita IVA;
- con partita IVA comune ad altra impresa;

⁴ In dettaglio: S_01 (Tessile-Abbigliamento-Pelle-Calzature); S_02 (Legno-mobilità); S_03 (Meccanica); S_04 (Altra manifattura); S_05 (Edilizia); S_06 (Impiantistica); S_07 (Commercio all'ingrosso); S_08 (Commercio al dettaglio); S_09 (Riparazione autoveicoli e motocicli); S_10 (Trasporto e magazzinaggio); S_11 (Servizi di alloggio e ristorazione - servizi turistici); S_12 (Servizi alla persona e alle famiglie); S_13 (Servizi alle imprese).

- senza addetti;
- di altra regione;
- operanti in settori non oggetto dell'osservatorio;
- con valori pari a zero per tutte le voci contabili;
- che hanno valori positivi di ricavi totali solo nell'ultimo trimestre dell'anno (onde eliminare le imprese la cui contabilità viene rilevata solo a fine anno producendo dati cumulati).

Anche l'esame della distribuzione per classe d'addetti dei domini considerati nel data-set risulta non dissimile rispetto alla distribuzione generale dell'archivio ASIA. Perciò anche da questo punto di vista il campione può essere ragionevolmente considerato come rappresentativo del registro permettendo la costruzione di coefficienti di riporto all'universo per ogni strato (provincia x settore x classe d'addetti) basati sul semplice rapporto tra la numerosità dello strato in ASIA e quella nel campione CNA⁵. Qualora uno o più strati del campione non sia popolato, la procedura prevede il collasso del numero di strati e la relativa ridefinizione dei coefficienti di riporto.

A questo punto la stima del tasso di variazione tendenziale di ciascuna variabile per dominio si basa sulla variazione tra il valore espanso e deflazionato al tempo t e il valore espanso e deflazionato al tempo $t-4$ delle imprese presenti in entrambi i trimestri. L'iterazione di tale procedimento produce però due stime diverse dei valori assoluti espansi per ogni trimestre: una relativa al panel trimestrale $[t, t-4]$ e una relativa al panel trimestrale $[t+4, t]$ dovute alla diversa composizione dei due panel trimestrali. Per tale ragione non sono diffusi i valori assoluti dei singoli trimestri ma solamente le stime relative alle variazioni tendenziali (e a numeri indice costruiti sulla base di esse), le quali sono invece uniche. Tale sistema ha dimostrato nel tempo di produrre stime robuste soprattutto a livelli di aggregazione maggiori. Ciò nonostante, a livelli di maggior dettaglio e soprattutto in quelle province dove l'affiliazione a CNA è meno intensa, la perdita di numerosità campionaria insita nella costruzione dei panel trimestrali può essere fonte di instabilità. Nel paragrafo successivo sono quindi descritti alcuni esperimenti messi in atto al fine di massimizzare l'utilizzo dell'informazione disponibile nei dati di partenza attraverso procedure di *data cleaning* e di imputazione di valori mancanti.

3. Descrizione dati e procedure di data cleaning

Il data set TREND della Toscana contiene 262.886 record riferiti a 26 trimestri consecutivi (dal 1° trimestre 2010 al secondo trimestre 2016) e un numero medio di 10.953 imprese uniche a trimestre. Nell'esplorazione iniziale del database sono state individuate due ordini principali di problematiche: quelle legate alle variabili di stratificazione e quelle legate alla distribuzione degli strati.

3.1. Procedure legate alle variabili di stratificazione e di stima

L'intero data set contiene 28 variabili⁶. Di queste sono state utilizzate Codice Provincia Cna, Codice Provincia, Ufficio (Cna), Anno, Trimestre, Codice Comune (a 6 cifre), Partita IVA, Codice Ateco2007, Ricavi, Numero di Addetti. Come già detto, le variabili di stratificazione per il progetto sono rappresentate dal Codice Provincia, dalla Partita IVA, dal Codice Ateco 2007 e dal Numero di Addetti e servono per

⁵ Il coefficiente ha il solo difetto di fare riferimento ad anni diversi a causa del ritardo di circa due anni nella diffusione dei dati ASIA rispetto a quelli CNA.

⁶ Le variabili presenti nel data set sono : Regione, Codice Provincia Cna, Codice Provincia, Ufficio CNA, Anno, Trimestre, Codice (interno a CNA), Comune, Piva, Codice Fiscale, Codice Ateco 2007, Anno Asia, Ricavi, Ricavi Estero, Ricavi Italia, Ricavi Terzi, Ricavi UE, Investimenti, Investimenti in Immobilizzazioni Materiali, Investimenti in Immobilizzazioni Immateriali, Macchinari, Retribuzioni, Consumi, Formazione, Assicurazioni, Numero Dipendenti, Numero Dipendenti Stimato, Numero Indipendenti, Numero Addetti.

definire strati e domini di stima. Per quanto attiene alle variabili di stima, invece, nel presente lavoro è stata presa in considerazione la sola variabile Ricavi, in quanto variabile principale del progetto. Le altre sono state utilizzate come variabili ausiliarie.

Al fine di standardizzare le operazioni è stata in primo luogo effettuata la normalizzazione delle modalità. Le variabili di stratificazione sono state codificate attraverso stringhe di caratteri mentre, per quanto riguarda i Ricavi, sono stati eliminati i decimali. Successivamente si è proceduto, attraverso un processo di imputazione deterministica, al recupero delle informazioni di imprese presenti nel database ma escluse dalle procedure di calcolo a causa di imprecisioni e mancanze nelle variabili di stratificazione. Per quanto riguarda il Codice Provincia, 993 unità erano valorizzate con codice “0”, mentre alcuni Codici Provincia erano inesistenti (es. 490, 480, 581). Attraverso l’uso della variabile Comune – a sei cifre - è stato possibile correggere gli errori di codifica, mentre la distribuzione dei 993 mancanti ha evidenziato un errore sistematico di valorizzazione della variabile Codice Provincia CNA in luogo della variabile Codice provincia nella provincia di Arezzo.

Successivamente sono state rimosse le province non appartenenti alla Regione e la provincia di Siena per la quale i dati contabili non sono attualmente disponibili. Il campo Partita Iva era completamente valorizzato sebbene in alcuni casi fossero presenti dei doppioni trimestrali. Questi, a meno che non fossero in due province diverse, sono stati rimossi dal dataset. Tale scelta è stata motivata dall’esigenza di cercare di preservare la maggior informazione possibile. La variabile Codice Ateco 2007 è quella che ha richiesto il maggior lavoro, data anche la sua rilevanza al fine della successiva classificazione dei settori economici. Da questo punto di vista è comunque utile sottolineare che a seconda del settore, non è necessario che la lunghezza del codice Ateco sia di sei cifre. In generale, infatti, i settori di stima regionali sono basati su un minimo di 2 cifre ad un massimo di 4. Nell’87% dei record registrati il codice era già di 6 cifre, mentre nei restanti 32.853 record (v.Tabella 1) il codice inserito aveva meno di 6 cifre. In particolare 6.454 record avevano lunghezza zero e 23.203 non più di quattro cifre.

Tabella 1 – distribuzione lunghezza dei Codici Ateco 2007

lunghezza Ateco	pre data cleaning		post data cleaning	
	valori assoluti	valori percentuali	valori assoluti	valori percentuali
0	6454	2,5	1824	0,7
1	7	0,0	7	0,0
2	488	0,2	531	0,2
3	5725	2,2	6129	2,3
4	10529	4,0	11493	4,4
5	9650	3,7	10114	3,9
6	230033	87,5	231432	88,5

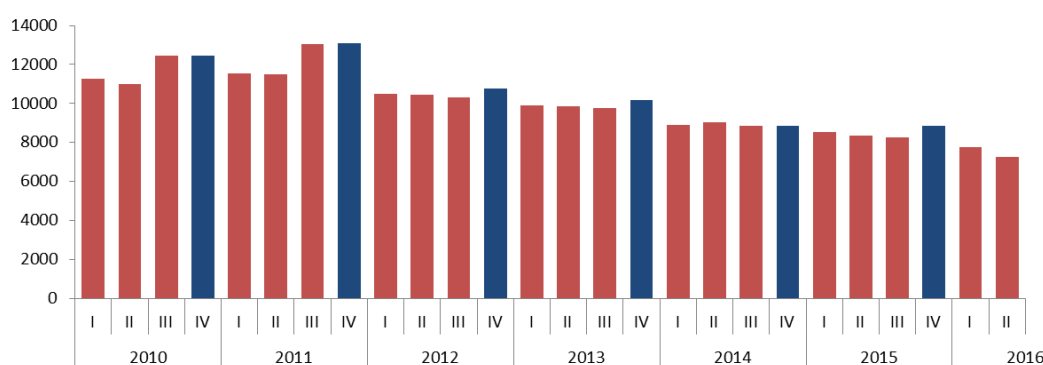
Fonte: nostre elaborazioni su dati CNA

A questi aspetti si aggiunge il fatto che il Codice Ateco non è una variabile statica visto che la singola impresa può, nel tempo, cambiare il proprio settore prevalente d’attività. Per questo motivo la procedura implementata ha cercato di individuare, a seconda della lunghezza del codice inserito, altri record riferiti alla stessa azienda. Nel caso in cui si sia riscontrata una coerenza tra i dati inseriti, il codice più lungo è stato copiato su tutte le celle riferite alla stessa unità. In caso di incoerenza invece, non è stato fatto nulla. La procedura ha permesso di ridurre il numero di record con Ateco mancante di circa 4600 unità. Restano comunque oltre 1800 record per i quali non è stato possibile procedere alla classificazione del settore d’attività. Per quanto riguarda la variabile Numero Addetti, completamente valorizzata, non è stato necessario alcun intervento.

3.2 La distribuzione degli strati

La Figura 1 evidenzia il calo continuo della numerosità dei dati trimestrali nel data set CNA. Un primo aspetto da sottolineare fa riferimento al confronto con i dati di Statistica Ufficiale. Rispetto al dato registrato per la Toscana nell'archivio ASIA⁷ per le imprese 0-9 addetti⁸, tra il 2012 e il 2015 si evidenzia una diminuzione di unità più che proporzionale, 19.2% rispetto al 2,9%, come se fattori diversi rispetto al normale movimento demografico abbiano influenzato la dinamica di entrata e uscita dal data set. Tuttavia, se si prende in considerazione ad esempio il settore delle costruzioni, che nei nostri dati pesa per il 33% circa, si nota che nel registro ASIA, dove il suo peso relativo è del 12%, il calo riscontrato è di circa il 13 punti percentuali (risultato maggiormente in linea con quanto avvenuto nel DB CNA). In pratica, sebbene per i motivi spiegati in nota 8 non sia possibile trarre conclusioni più stringenti, il confronto tra i due insiemi di imprese permette di accettare l'idea che buona parte del movimento riscontrato nel data set CNA sia imputabile all'andamento economico e demografico generale delle imprese toscane. Si può anche notare come la numerosità infra-annuale non segua una distribuzione completamente uniforme, con picchi che si manifestano con maggiore frequenza nel quarto trimestre. Questo comportamento rispecchia una generale irregolarità nella cadenza delle registrazioni, legata alle diverse esigenze di ricorso ai servizi di CNA, e in alcuni casi sottende la presenza di dati cumulati.

Figura 1 – Numerosità dei record per anno e trimestre



Fonte: nostre elaborazioni su dati CNA

La distribuzione completa degli strati, come prodotto cartesiano delle 9 province, dei 13 settori economici e delle 3 classi di addetti, prevede un totale di 351 strati a livello regionale e 39 a livello di singola provincia.

Tuttavia, data la natura non probabilistica del disegno campionario e l'andamento decrescente della numerosità delle imprese, 12 strati non sono popolati. In tabella 3 è riportato l'elenco dei 12 strati vuoti non recuperabili neppure tramite imputazione⁹. Come si vede a livello territoriale la provincia di Massa Carrara è quella che presenta il più elevato numero. È inoltre l'unica provincia in cui manca uno strato della seconda classe di addetti (2-5 addetti). In tutti gli altri casi invece gli strati non popolati fanno riferimento alle imprese di maggiori dimensioni.

⁷ Archivio Statistico delle Imprese Attive (Istat) disponibile presso il data warehouse I.Stat sul sito www.istat.it.

⁸ Non essendo disponibile un dato per le imprese fino a 19 addetti è stato necessario usare il dato che più ci si approssimava.

⁹ Per questo motivo risulta essenziale l'uso della procedura di collasso degli strati per la ridefinizione dei riporti all'universo.

Tabella 2– Lista degli strati mancanti a livello di singola provincia

<i>provincia</i>	<i>settore di attività</i>	<i>classe di addetti</i>
MS	S_01	2
MS	S_01	3
MS	S_02	3
MS	S_04	3
MS	S_08	3
LU	S_09	3
LU	S_12	3
PI	S_07	3
AR	S_07	3
AR	S_09	3
GR	S_07	3
PO	S_08	3

Fonte: nostre elaborazioni su dati CNA

La distribuzione globale degli strati non rende inoltre conto dell’eterogeneità temporale a livello annuale e di singolo trimestre. In tabella 4 è riportata la distribuzione di frequenza degli strati vuoti a livello di singolo anno e trimestre. Dall’osservazione della tabella emerge che all’aumentare del dettaglio temporale il loro numero tende a crescere. Infatti, se gli strati completamente mancanti sono 12, il numero medio di strati mancanti a livello annuale sale a circa 26, mentre a livello trimestrale raggiunge il valore di circa 37.

Tabella 3 – lista degli strati mancanti per anno e trimestre

<i>anno</i>	<i>distrib. annuale</i>	<i>distrib. trimestrali</i>			
		<i>1</i>	<i>2</i>	<i>3</i>	<i>4</i>
2010	20	34	33	37	33
2011	25	39	38	28	37
2012	22	36	33	35	31
2013	23	38	34	38	31
2014	32	41	47	41	42
2015	25	44	42	42	35
2016	40	48	50	0	0

Fonte: nostre elaborazioni su dati CNA

Inoltre, la costruzione tramite record linkage dei panel trimestrali $[t, t-4]$ necessari per procedere alla stima del valore economico dei domini, influenza ulteriormente il processo di stratificazione. I motivi principali sono due: da un lato il “saldo naturale” delle imprese presenti nei campioni trimestrali, in quanto vengono prese solo le unità presenti in entrambi i trimestri; dall’altro la distribuzione degli strati vuoti e pieni nei due trimestri. In pratica, qualora nello strato x del trimestre t siano presenti unità diverse da quelle presenti nel medesimo strato del trimestre $t-4$, lo strato risultante dall’aggancio tra trimestri risulterà vuoto. Allo stesso modo se uno strato è pieno in un trimestre e vuoto nell’altro lo strato finale risulterà vuoto. Poiché la distribuzione annuale rappresenta il numero minimo di strati ottenibili dai dati disponibili nell’anno è stata introdotta una ipotesi di comodo che chiameremo “ipotesi annuale”. In base ad essa un’unità che ha dato un segnale di esistenza in un singolo trimestre dell’anno è considerata esistente per tutto l’anno e quindi per

ogni trimestre in cui non sono presenti registrazioni relative alla singola unità, verranno aggiunti altrettanti record. Il data-set annuale avrà a questo punto una forma rettangolare, con le medesime imprese presenti in tutti e quattro i trimestri. L'ipotesi annuale permette di ridurre i suddetti problemi di record linkage, sia riducendo le problematiche di mancato aggancio tra unità sia minimizzando i problemi di stratificazione incrociata. Avendo posto un vincolo di tipo annuale, dati due anni consecutivi, i quattro panel trimestrali costruibili a partire dal primo trimestre $[t, t-4]$, $[t+1, t-3]$, $[t+2, t-2]$ e $[t+3, t-1]$ hanno tutti la stessa distribuzione degli strati, contribuendo in questo modo a ridurre l'eterogeneità generale delle stime. A titolo di esempio è stato effettuato l'aggancio tra i trimestri 2015_4 e 2014_4 sia nei dati di partenza, sia nei data set rettangolari dei due anni. La tabella 5 riporta l'elenco dettagliato degli strati mancanti prima (48) e dopo l'applicazione (45) dell'ipotesi annuale evidenziando in rosso i 3 strati recuperati nelle province di Pisa e Grosseto.

Tabella 4 – Lista degli strati mancanti nel panel trimestrale [2015_4 , 2014_4] prima e dopo l'applicazione dell' "ipotesi annuale"

<i>db originale</i>			<i>ipotesi annuale</i>			<i>db originale</i>			<i>ipotesi annuale</i>		
<i>provincia</i>	<i>settore</i>	<i>classe addetti</i>	<i>provincia</i>	<i>settore</i>	<i>classe addetti</i>	<i>provincia</i>	<i>settore</i>	<i>classe addetti</i>	<i>provincia</i>	<i>settore</i>	<i>classe addetti</i>
MS	S_01	2	MS	S_01	2	PI	S_09	3	PI	S_09	3
MS	S_01	3	MS	S_01	3	PI	S_13	3	PI	S_13	3
MS	S_02	3	MS	S_02	3	AR	S_01	3	AR	S_01	3
MS	S_03	3	MS	S_03	3	AR	S_02	3	AR	S_02	3
MS	S_04	3	MS	S_04	3	AR	S_04	3	AR	S_04	3
MS	S_07	3	MS	S_07	3	AR	S_05	3	AR	S_05	3
MS	S_08	3	MS	S_08	3	AR	S_06	3	AR	S_06	3
MS	S_09	3	MS	S_09	3	AR	S_07	3	AR	S_07	3
LU	S_02	3	LU	S_02	3	AR	S_08	3	AR	S_08	3
LU	S_04	3	LU	S_04	3	AR	S_09	3	AR	S_09	3
LU	S_05	3	LU	S_05	3	AR	S_10	3	AR	S_10	3
LU	S_06	3	LU	S_06	3	AR	S_11	3	AR	S_11	3
LU	S_09	3	LU	S_09	3	AR	S_12	3	AR	S_12	3
LU	S_10	3	LU	S_10	3	AR	S_13	3	AR	S_13	3
LU	S_12	3	LU	S_12	3	GR	S_07	3	GR	S_07	3
PT	S_06	3	PT	S_06	3	GR	S_08	3	GR	S_08	3
PT	S_08	3	PT	S_08	3	GR	S_09	3	GR	S_11	3
PT	S_09	3	PT	S_09	3	GR	S_11	3			
FI	S_11	3	FI	S_11	3	GR	S_12	3			
LI	S_01	2	LI	S_01	2	PO	S_02	3	PO	S_02	3
PI	S_07	2				PO	S_07	3	PO	S_07	3
PI	S_01	3	PI	S_01	3	PO	S_08	3	PO	S_08	3
PI	S_07	3	PI	S_07	3	PO	S_09	3	PO	S_09	3
PI	S_08	3	PI	S_08	3	PO	S_10	3	PO	S_10	3

Fonte: nostre elaborazioni su dati CNA

4. Data imputation

Effettuato il *data cleaning* delle variabili principali si è passati all'imputazione dei dati mancanti per la sola variabile Ricavi. Tra i vari pacchetti disponibili in R per effettuare la *data imputation* - tra i quali MI (Su et al. 2011, Gelman et al. 2015), MICE (van Buuren et al. 2017, van Buuren et al. 2011), VIM (Templ et al.

2017), imputeTS (Moritz 2017) – è stato scelto di utilizzare il pacchetto AMELIA II (Honacker et al. 2011, Honacker et al. 2016) in quanto particolarmente adatto per l'imputazione delle serie storiche sia panel sia CSTS. Il pacchetto inoltre è molto ben testato e supportato. L'algoritmo che utilizza è definito EMB ovvero è un algoritmo di tipo EM con bootstrap, B, e si basa su due ipotesi fondamentali: la normalità della distribuzione congiunta delle variabili e il rispetto dell'assunzione MAR (Missing At Random) per i dati mancanti. In base a quanto ha scritto Rubin (Rubin 1976) la creazione dei dati mancanti è determinata da tre meccanismi prevalenti definiti come: Missing Completely At Random (MCAR), Missing At Random (MAR) e Not Missing At Random (NMAR). Nel caso di dati MCAR la probabilità che il dato sia "missing" non dipende né dalle altre variabili osservate né da variabili non osservate, ovvero la sua assenza è legata a fattori completamente casuali, come se fosse legata al lancio di una moneta. Nel caso dei dati di tipo MAR la probabilità che il dato sia mancante dipende da altre variabili osservate ma non dalla variabile stessa¹⁰. Nel caso di dati NMAR, che rappresenta il caso più difficile da trattare, la probabilità che il dato sia mancante dipende dalla presenza di variabili non osservate nel data set.

Per chiarire meglio questi ultimi due elementi si farà riferimento all'esempio proposto da Dong e Peng (2013)¹¹. In un college agli studenti del corso X è richiesto di fare un test pre-corso e un test post corso con l'intenzione di comparare i risultati. Tuttavia nei vari anni in cui sono state effettuate le prove è emerso che la probabilità che uno studente partecipi al post-test (variabile di analisi) dipende dal risultato del pre-test (predittore). L'ipotesi sottostante quindi è che la mancanza del dato sul post-test sarà verosimilmente dipendente dall'esito del pre-test. A parità di voto nel pre-test tuttavia, è ragionevole supporre che la probabilità di partecipare al post-test sia totalmente casuale. Questo esempio chiarifica le condizioni in base alle quali può essere giustificato un meccanismo di tipo MAR. Però, sempre in base a quanto scrivono gli autori, affinché si possa parlare effettivamente di MAR è necessario che vengano registrati gli esiti individuali del predittore (pre-test) altrimenti, risultando non-osservata, si ricadrà nel caso NMAR. Poiché il meccanismo MCAR è un caso speciale del meccanismo MAR, l'algoritmo presente in AMELIA II risulta essere adatto ad identificare pattern di imputazione in entrambi i casi (ma non in quello NMAR).

Un ultimo aspetto da sottolineare riguarda la non rettangolarità del data set complessivo, anche a seguito dell'introduzione dell'ipotesi annuale. Il numero di imprese presenti varia infatti di anno in anno. Sebbene nel manuale del pacchetto di R non ci sia nessun riferimento riguardo l'impossibilità di utilizzare AMELIA II in caso di data set unbalanced, la prassi e gli esempi forniti sembrano fare riferimento sempre ad un balanced dataset. Per questo motivo si è preferito costruire un nuovo DB specifico per realizzare l'imputazione dei dati mancanti in cui tutte le imprese risultano presenti in tutti gli anni. Le variabili inserite contengono tutte le variabili di stratificazione – Codice Provincia, Partita Iva, Settore di Attività e Classe di Addetti – più le variabili anno, trimestre e Ricavi (variabile di analisi). Per i record non valorizzati nelle variabili di stratificazione è stata definita una funzione in grado di raccogliere il dato mancante dal record, riferito alla stessa azienda, temporalmente più vicino, sotto l'ipotesi semplificatrice che in un tempo non eccessivamente lungo anche le variabili di tipo dinamico come il settore di attività (dipendente dal codice ATECO) e la classe di addetti possano essere considerate approssimativamente costanti. Al termine del processo di imputazione, per mantenere l'ipotesi annuale descritta nel paragrafo precedente, è stato effettuato un record linkage tra il data set imputato e i data set annuali rettangolari descritti nel paragrafo precedente.

4. 1. *La verifica dell'ipotesi di normalità e di quella MCAR/MAR*

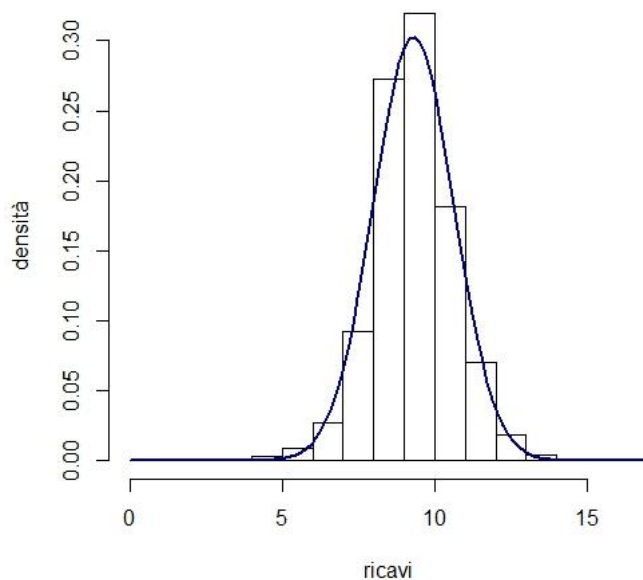
Per quanto riguarda la verifica dell'ipotesi di normalità della variabile Ricavi è noto che, sebbene la sua distribuzione naturale sia asimmetrica, la sua trasformazione logaritmica – con logaritmo naturale – segua approssimativamente una distribuzione di tipo normale. In Figura 2 vengono confrontate la distribuzione dei log-ricavi (istogramma) e la densità normale (linea continua) realizzata con il medesimo valore atteso e la

¹⁰ L'ipotesi MCAR rappresenta inoltre un sotto-caso dell'ipotesi MAR.

¹¹ Traduzione libera

stessa deviazione standard del log. La concordanza dei due profili ha suggerito l'opportunità di effettuare l'imputazione dei ricavi tramite la loro trasformazione logaritmica.

Figura 2 – Normalità della log-distribuzione della variabile ricavi



Fonte: nostre elaborazioni su dati CNA

Per quanto riguarda la verifica di pattern di tipo MCAR o MAR nel data set di imputazione invece, è necessario dire fin da subito che un risultato conclusivo non è stato possibile ottenerlo. Infatti, sebbene sia possibile testare l'assunzione MCAR tramite il test di Little (Little 1988), come è stato più volte sottolineato in letteratura, non esiste un approccio standardizzato che permetta di testare con certezza l'ipotesi MAR (Fielding 2009). In pratica, qualora il test di Little non risulti significativo, non esiste una procedura standardizzata che permetta di escludere la possibilità che le variabili non osservate influenzino il pattern dei dati mancanti. In generale, a parte il ritorno sul campo per cercare di raccogliere le informazioni mancanti direttamente, in letteratura si fa spesso riferimento alla consapevolezza del ricercatore che svolge le analisi – il quale dovrebbe essere in grado di valutare con buona precisione se il processo che ha portato alla formazione del missing sia ascrivibile ad un meccanismo di tipo MAR o meno (Allison 2001) - e all'inserimento di un numero di variabili ausiliare sufficiente a garantire che la dipendenza dalle variabili non osservate sia ragionevolmente bassa (Allison 2001, Dong e Peng 2013). Senza pretesa di completezza un ultimo gruppo di approcci statistici si pongono l'obiettivo di rafforzare la giustificabilità dell'esistenza di un meccanismo di tipo MAR, più che testare l'ipotesi in sè, tra questi c'è l'utilizzo della regressione logistica (Ridout 1993, Faiclough 2002), del test-t per sui gruppi dei rispondenti e dei non rispondenti (Dong e Peng 2013) e del LS-test (Listing and Schlittgen).

Nel data set creato ai fini dell'imputazione, che conteneva una percentuale di missing di circa il 50%, sono stati utilizzati sia il test di Little, il cui esito ha portato a rifiutare l'ipotesi nulla in quanto il p-value raggiunto è sempre stato pressoché zero, sia l'utilizzo della regressione logistica. In base alla presenza o meno di dati mancanti nella variabile ricavi è stata definita una variabile dummy con valore 1 nel caso di dato mancante e 0 nel caso opposto.

Tabella 5 – Risultato della regressione logistica

	variabili	coefficienti	errore standard	Pr(> z)	livello di significatività
	intercetta	-350,9	3,226	< 2e-16	***
tempo	anno	0,1746	0,001603	< 2e-16	***
	trimestre	-0,04233	0,002682	< 2e-16	***
province	046	0,133	0,02082	1,7E-10	***
	047	0,04714	0,01662	0,00457	**
	048	-0,3568	0,01538	< 2e-16	***
	049	-0,2048	0,01766	< 2e-16	***
	050	-0,04678	0,0181	0,00977	**
	051	-0,5311	0,01753	< 2e-16	***
	053	-0,3565	0,01707	< 2e-16	***
	100	-0,1175	0,01778	3,93E-11	***
settore di attività	S_02	0,04467	0,01993	0,02504	*
	S_03	0,1193	0,01731	5,57E-12	***
	S_04	0,1229	0,01755	2,54E-12	***
	S_05	0,1532	0,01355	< 2e-16	***
	S_06	-0,2551	0,01609	< 2e-16	***
	S_07	0,3069	0,02034	< 2e-16	***
	S_08	0,3337	0,01616	< 2e-16	***
	S_09	-0,3008	0,02149	< 2e-16	***
	S_10	-0,2398	0,01853	< 2e-16	***
	S_11	0,761	0,02001	< 2e-16	***
	S_12	-0,2487	0,01511	< 2e-16	***
	S_13	0,303	0,01495	< 2e-16	***
addetti	x5	-0,2625	0,005464	< 2e-16	***

Fonte: nostre elaborazioni su dati CNA

I risultati sono descritti in dettaglio in Tabella 5. La probabilità che un dato sia mancante risulta essere significativamente correlata con gli strati creati, testimoniando che le variabili utilizzate, sebbene come detto non permettano di escludere completamente l'ipotesi NMAR, possono essere considerate ragionevolmente sufficienti per il rispetto dell'ipotesi MAR. Infatti, considerato come baseline lo strato 045_S_01_1, ovvero le aziende appartenenti al settore del Tessile-Abbigliamento-Calzature della provincia di Massa-Carrara di classe dimensionale 1, esistono due province e 7 settori produttivi con coefficienti positivi. Questo risultato può essere integrato con alcune considerazioni di tipo esperienziale. Infatti, la probabilità di ottenere un dato mancante in un trimestre dipende sostanzialmente dal fatto che l'azienda in questione si sia recata presso CNA o meno per usufruire dei servizi offerti dall'ente. In alcuni casi il fenomeno è temporaneo e l'unità riappare con cadenza variabile nei trimestri/anni successivi, in altri casi invece è collegato alla scomparsa dal data set dell'unità. Come già accennato, i motivi possono essere diversi, da fattori di natura economica, a fattori individuali legati alle caratteristiche dell'imprenditore o alla dimensione aziendale fino a fattori di natura opportunistica. Tra questi tuttavia, gli aspetti economici, soprattutto negli anni presi in considerazione, hanno sicuramente svolto un ruolo significativo. Probabilmente non è un caso che le province e i settori dove

la probabilità che si verifichino dei dati mancanti è più elevata siano tra quelle che in Toscana hanno maggiormente risentito dell'effetto della crisi economica.

4.2 Il processo di imputazione

Il processo di imputazione è stato effettuato raggruppando le unità per strato, utilizzando uno spline polinomiale di grado 2 del tempo, differenziato per singolo strato (opzione *intercs*). Coerentemente con le indicazioni di Honecker (2016), sono stati creati in tutto 5 data set completi. Tuttavia, dato l'eccessivo sforzo computazionale, al fine di poter effettuare il lavoro, è stato necessario separare le unità di partenza, attraverso un criterio di continuità geografica e di numerosità, in 4 DB ridotti composti da: 1) Massa-Carrara, Lucca, Pisa e Livorno; 2) Pistoia e Prato; 3) Firenze; 4) Arezzo e Grosseto.

Tabella 6: Esito del processo di imputazione

	output sui 5 data set imputati per ogni area			
	MS-LU-LI-PI	PT-PO	FI	AR-GR
Return code	1	1	1	1
Normal EM convergence	yes	yes	yes	yes
Rows after Listwise Deletion	46590	43265	84995	51500
Rows after Imputation	117156	106028	161018	96798
Patterns of missingness in the data	2	2	2	2
Fraction Missing: indivID	0	0	0	0
Fraction Missing: tempo	0	0	0	0
Fraction Missing: strati	0	0	0	0
Fraction Missing: ricavi	0,6023251	0,5919474	0,4721398	0,4679642

Fonte: nostre elaborazioni su dati CNA

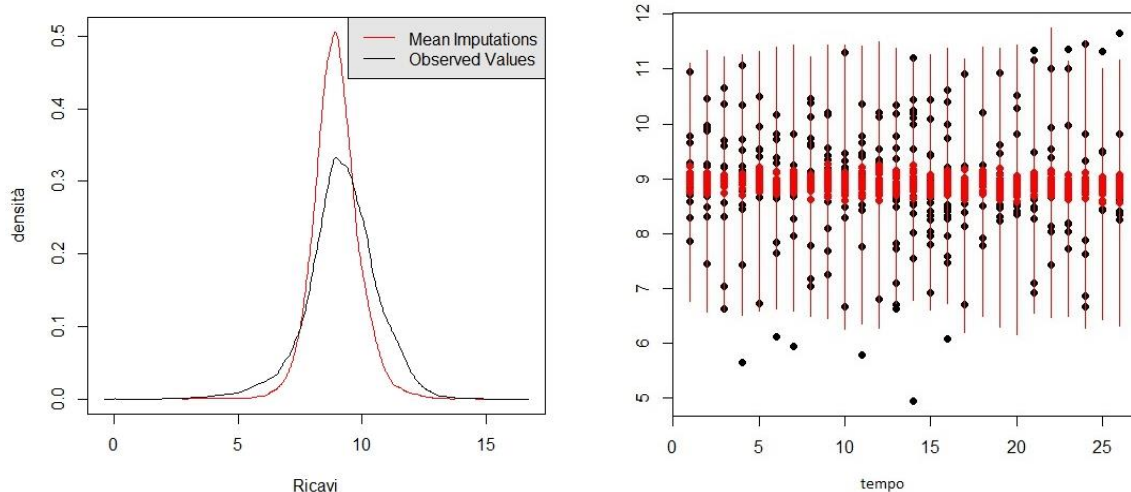
Le variabili inserite sono state: un identificativo unitario relativo alla singola azienda ("indivID" costituito da una stringa composta da Partita IVA e Codice Provincia), il "tempo" come sequenza di trimestri da 1 a 26 per rispettare il formato "long" del data set necessario per il processo di imputazione, gli "strati" e i "ricavi".

La Tabella 6 riassume i principali dati diagnostici del processo di imputazione. Gli ultimi 4 record danno una misura della frazione di dati mancanti per singola variabile e per area di imputazione (nel data set rettangolare) da dove emerge che la zona di Massa-Carrara, Lucca, Livorno e Pisa sia quella con la percentuale più elevata di mancata registrazione dei ricavi¹². Gli altri record invece attestano la correttezza della convergenza dell'algoritmo. Ai fini degli obiettivi del nostro lavoro risulta inoltre interessante il confronto tra il numero di record finali ottenuti con il processo di imputazione rispetto all'approccio listwise deletion, che dà una misura del guadagno numerico ottenuto in questo modo. La diagnostica del processo di imputazione è stata inoltre approfondita grazie all'utilizzo di due funzioni presenti nel pacchetto *Amelia II* che permettono di confrontare la distribuzione dei dati imputati rispetto a quelli originali in termini di densità e nel tempo: "compare.density" e "tscsPlot". In Figura 3 sono rappresentati i risultati ottenuti dall'applicazione delle precedenti procedure diagnostiche per quanto riguarda l'area Massa-Carrara, Lucca, Livorno e Pisa. Nel grafico a sinistra vengono paragonate le distribuzioni di densità dei log-ricavi osservati e della media dei log-ricavi imputati visti nel loro complesso. La distribuzione imputata presenta un profilo molto simile a quello della distribuzione originale, senza alterarne il valore atteso come nelle aspettative dell'algoritmo, ma con una varianza più contenuta. Il comportamento nel tempo conferma le stesse conclusioni. Nel grafico a destra, che nuovamente considera tutti gli strati insieme, sono rappresentati in ascissa i 26 trimestri e in ordinata i valori individuali dei log-ricavi imputati (in rosso) e osservati (scuro). L'ispezione conferma come anche nel tempo l'algoritmo sia riuscito in tutti i trimestri ad imputare i dati

¹² Sebbene l'area di Pistoia e Prato abbia una percentuale di dati mancanti sostanzialmente uguale.

individuali mantenendo un campo di variazione relativamente stretto rispetto ai dati di partenza, con un valor medio che varia di trimestre in trimestre, come desiderato. Il comportamento descritto è riscontrabile anche nelle altre aree geografiche prese in considerazione tuttavia, per motivi di spazio non ne verranno presentati i risultati. Si fa presente comunque che l'area geografica presa in considerazione è quella dove, grazie alla maggior presenza di missing, si registra la maggior instabilità delle stime.

Figura 3 – log-Ricavi imputati e log-Ricavi osservati comparazione delle densità (sx) e nel tempo(dx)



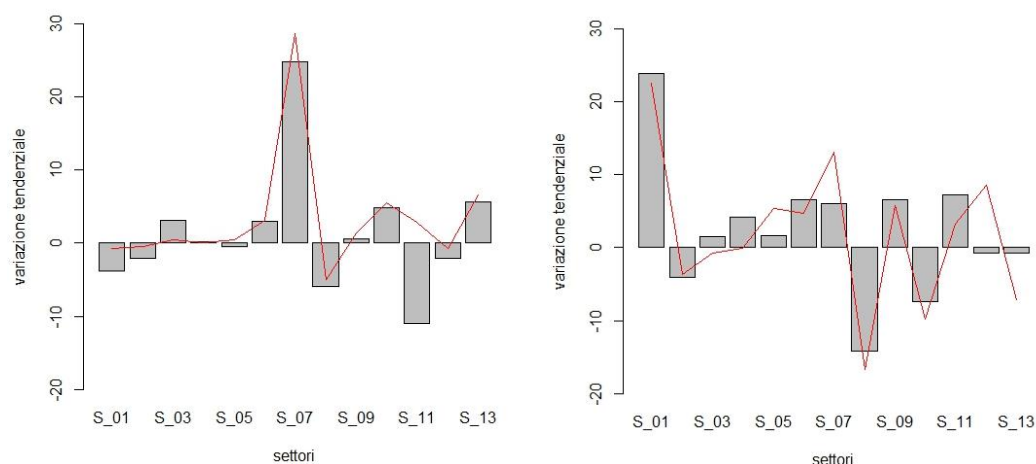
Fonte: nostre elaborazioni su dati CNA

5. Valutazione (approssimata) del lavoro

Terminato il processo di imputazione sono diventati disponibili 5 data set completi di 481.000 unità l'uno che, considerata la natura del lavoro da realizzare e la presenza di una sola variabile di analisi, sono stati accodati in un unico data set rettangolare con chiave di aggancio costituita da "indivID" e "numero di imputazione". Da questo unico insieme di dati sono stati poi ricavati, tramite record linkage, i data set annuali in base all'ipotesi annuale descritta nel paragrafo §3. Il risultato finale è stato quello di avere a disposizione 6 data set rettangolari, sebbene di dimensione diversa, per gli anni 2010, 2011, 2012, 2013, 2014, 2015, più un ulteriore data set rettangolare per il 2016 contenente solo i primi due trimestri (ultimo dato disponibile). In sintesi, il processo ha permesso di costruire un data set completo per il periodo primo trimestre 2010 – secondo trimestre 2016 basato sull'accostamento di una sorta di panel annuali. Come accennato sopra, l'effetto è quello di cercare di contenere il numero degli strati completamente mancanti a seguito dell'aggancio dei trimestri $[t, t-4]$ nella definizione delle variazioni tendenziali stabilizzando i risultati. Per valutare questo secondo aspetto è stata costruita una procedura di stima approssimata delle variazioni tendenziali per dominio¹³. Il motivo dell'approssimazione risiede nel fatto che la stima è stata realizzata utilizzando i valori contabili non deflazionati e senza procedura di collasso. Per questo motivo il confronto è stato fatto tra il quarto trimestre del 2015 e il quarto del 2014 - in un periodo di prezzi alla produzione pressoché costanti - e per le sole province di Firenze e Livorno dove risultava mancante, in entrambi i casi, un solo strato. Con la stessa procedura e per le stesse province, sono state calcolate le variazioni tendenziali sul data set ottenuto dopo la procedura di data cleaning (in questo caso i dati mancanti sono stati posti pari a 0 per poter ottenere le stime).

¹³ Poiché i domini di stima sono rappresentati dall'incrocio tra provincia e settore, a livello della singola provincia i domini di stima corrispondono ai settori provinciali.

Figura 4 – andamento delle variazioni tendenziali approssimate per dominio pre (barre) e post (linea rossa) imputazione. Province di Firenze (sx) e Livorno (dx). Trimestri 2015_4 e 2014_4.



Fonte: nostre elaborazioni su dati CNA

L'andamento dei risultati è riportato in Figura 4 con a sinistra i risultati ottenuti per la provincia di Firenze e a destra quelli ottenuti per la provincia di Livorno. Il grafico a barre rappresenta i valori stimati dal data set pulito ma precedente al processo di imputazione, mentre la linea rossa rappresenta il valore stimato al termine del processo di imputazione. In entrambi i casi nella maggioranza dei settori economici le stime post imputazione evidenziano un andamento più stabile rispetto a quello dei dati di partenza, sebbene con alcune eccezioni come il settore S_07. A Livorno inoltre, in generale l'effetto stabilizzante sembra essere più marcato. In tabella 7 sono riportati i dati puntuali stimati.

Tabella 7 – stime tendenziali approssimate per dominio di stima pre- e post-imputazione. Province di Firenze e Livorno. Trimestri 2015_4 e 2014_4.

	domini di stima provinciali (settori)												
	S_01	S_02	S_03	S_04	S_05	S_06	S_07	S_08	S_09	S_10	S_11	S_12	S_13
Firenze imputato	-0,70	-0,54	0,49	0,09	0,48	3,06	28,70	-4,98	1,44	5,49	2,79	-0,70	6,62
Firenze pre-clean	-3,86	-2,05	3,15	0,15	-0,46	2,94	24,84	-5,88	0,55	4,83	-11,04	-2,13	5,64
Livorno imputato	23,93	-4,11	1,55	4,14	1,67	6,60	6,01	-14,18	6,52	-7,37	7,31	-0,80	-0,73
Livorno pre-clean	22,53	-3,65	-0,70	-0,07	5,36	4,78	13,11	-16,72	5,77	-9,86	3,28	8,54	-7,18

Fonte: nostre elaborazioni su dati CN

Al fine di poter dare una misura dell'effetto globale di contenimento delle stime è stata fatta la sommatoria delle differenze tra i valori assoluti settoriali pre e post imputazione. Il risultato globale che emerge nella provincia di Firenze attesta una riduzione netta di circa 11 punti percentuali mentre per quella di Livorno di circa 28 punti.

6. Conclusioni

Il data set relativo alle caratteristiche delle imprese appartenenti al campione CNA rappresenta una fonte informativa di assoluto interesse che ha favorito la creazione di un osservatorio congiunturale - al momento riferito a Marche, Toscana, Umbria ed Emilia-Romagna - in grado di soddisfare un bisogno informativo proveniente dal territorio in modo efficace e tempestivo. Sempre più infatti è aumentata la consapevolezza degli stakeholders relativa alla necessità di fondare le scelte istituzionali sulle informazioni provenienti dai

dati. A questo aspetto si aggiunga che la popolazione obiettivo del progetto TREND rappresenta una delle parti del tessuto economico nazionale più peculiari e al contempo sfuggenti: la micro e piccola impresa.

Tuttavia, la ricchezza del dato di partenza si scontra da un lato con la natura amministrativa che ne giustifica la creazione ma anche con le necessità e specificità territoriali. Basti pensare come la ripartizione settoriale a livello regionale non possa che rispecchiare le caratteristiche produttive territoriali a loro volta collegate alle tendenze economiche prevalenti. Nel procedere alla realizzazione del progetto sono perciò diventate via via più evidenti sia le potenzialità sia le criticità del lavoro e, in riferimento al presente articolo, dei dati di partenza. Il dato amministrativo infatti, rispetto al dato statistico raccolto appositamente a scopo informativo, è maggiormente influenzato da problematiche legate agli errori non campionari: l'errore di copertura, l'errore di mancata risposta (totale o parziale) e l'errore di misurazione. Il data set in esame soffre di tutti e tre le problematiche e fin da subito l'architettura del progetto TREND è stata costruita al fine di "correggere" le potenziali distorsioni che ne potevano emergere. In questo solco si inserisce il presente lavoro.

L'elevata presenza di mancate risposte sia a livello totale sia soprattutto a livello trimestrale costituisce una criticità che va a toccare il meccanismo centrale dell'intero lavoro: la creazione dei panel trimestrali [t, t-4] fondamentali per procedere alla stima del volume economico settoriale. Tuttavia, la crescente disponibilità di pacchetti informatici affidabili e accessibili ha promosso l'idea di tentare la strada dell'imputazione – per ora per la sola variabile dei "ricavi" e successivamente probabilmente per più variabili – con l'obiettivo di aumentare la disponibilità di record utili per calcolare le variazioni tendenziali. Qualora il processo di imputazione funzioni, il risultato atteso è quello di osservare una minore variabilità delle stime.

L'approccio congiunto di data cleaning, inteso in senso ampio (ovvero sia di pulizia-normalizzazione-codifica delle variabili sia di "data imputation deterministica") e data imputation costituiscono un punto di partenza che si è dimostrato efficace nel confrontarsi con le criticità legate alla forte presenza di dati mancanti. In particolare, sebbene non sia possibile riuscire a colmare con i dati attuali tutti i settori produttivi definiti per la regione, è stato possibile individuare un set minimo di settori mancanti, 12, verso cui tendere.

L'utilizzo dell'ipotesi annuale è servito a mediare tra la realtà del dato originale e le necessità statistiche. L'effetto, come messo in evidenza nel paragrafo 3, è stato quello di ridurre il numero degli strati vuoti dei singoli trimestri e in questo modo ridurre la presenza di strati vuoti nel panel trimestrale, azzerando la variabilità trimestrale tra anni adiacenti. Infatti, attraverso l'ipotesi annuale il comportamento riscontrato tra t e t-4 per un certo trimestre vale anche per tutti gli altri trimestri dell'anno (es. 2015_4 e 2014_4 avranno la stessa struttura di 2015_3 e 2014_3, 2015_2 e 2014_2, 2015_1 e 2014_1). Tuttavia l'ipotesi introdotta, anche in presenza di un efficace sistema di imputazione che permette di azzerare i ricavi mancanti e di stabilizzare le stime, non ha per il momento permesso di raggiungere risultati più consistenti.

Da questo punto di vista alcune strade sono ancora percorribili. Tra queste la prima è rappresentata dall'aggancio con i dati provenienti da ASIA. Finora il lavoro ha teso a valorizzare il contenuto informativo presente nel data set originale. Non è stato però per esempio possibile ricostruire i codici ATECO per diverse unità presenti nel data set. Recuperando quest'informazione potrebbe essere altresì possibile recuperare qualche strato mancante. L'utilizzo di ASIA potrebbe rivestire anche un ruolo rilevante sia per la variabile "ricavi" sia per la variabile "classe di addetti": nel primo caso fornendo informazioni in merito ai limiti dell'intervallo di imputazione da assegnare al singolo record, nel secondo caso fornendo un'informazione di controllo sul numero di addetti che potrebbe cambiarne la distribuzione.

7. Bibliografia

Palmieri R., (eds.) (2013), Studio di fattibilità per un'indagine congiunturale da fonte amministrativa non Sistan sulla micro e piccola impresa, e per l'integrazione di dati proveniente da archivi amministrativi non Sistan nell'ambito della statistica ufficiale. Rapporto finale del Gruppo di lavoro Istat-CNA (PSN STU IST-02344), Istat.

- Baltagi, B.H. (2005) *Econometric Analysis of Panel Data*. 3rd Edition, John Wiley & Sons Inc., New York.
- Su Y.S., Gelman A., et al. (2011) Multiple Imputation with Diagnostics (mi) in R: Opening Windows into the Black Box. *Journal of Statistical Software*, Vol.14, n°2.
- Gelman A., Hill J., Su Y.S. (2015). MI package reference manual. Disponibile: <https://cran.r-project.org/>
- van Buuren S., Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*, Vol.45, n°3.
- van Buuren S., Groothuis-Oudshoorn K., Robitzsch A. (2017). Mice package reference manual. Disponibile: <https://cran.r-project.org/>
- Templ M., Alfons A., Kowarik A., Prantner B. (2017). VIM package reference manual. Disponibile: <https://cran.r-project.org/>
- Moritz S. (2017). imputTS package reference manual. Disponibile: <https://cran.r-project.org/>
- Honaker J., King G., Blackwell M. (2011). AMELIA II a program for missing data. *Journal of Statistical Software*, Vol.45, n°7.
- Honaker J., King G., Blackwell M. (2016). AMELIA II package reference manual. Disponibile: <https://cran.r-project.org/>
- Rubin DB (1976) Inference and missing data. *Biometrika* 63(3):581–592. doi:10.1093/biomet/63.3.581
- Dong, Y., Peng, C.Y.J. (2013). *Principled Missing Data Methods for Researchers*. Springer Plus, 2, 222.
- Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83(404), 1198-1202.
- Fielding, S., P.M. Fayers, and C.R. Ramsay. 2009. “Investigating the Missing Data Mechanism in Quality of Life Outcomes: A Comparison of approaches. *Health and Quality of Life Outcomes* 2009, 7:57 doi:10.1186/1477-7525-7-57
- Allison, P., 2001. *Missing data —Quantitative applications in the social sciences*. Thousand Oaks, CA: Sage. Vol. 136.
- Ridout, M.S., Diggle, P.J. Testing for random dropouts in repeated measurement data. *Biometrics* 1991, 47:1617-1619.
- Fairclough, D.L. 2002. *Design and Analysis of Quality of Life Studies in Clinical Trials* Chapman and Hall; 2002.
- Listing J., Schlittgen R. (1998): Tests if dropouts are missed at random. *Biometrical Journal* 1998, 40:929-935.