Dipartimento di Scienze Statistiche

Dottorato di Ricerca "Scuola di Scienze Statistiche" – 33° Ciclo

Curriculum Statistica Metodologica

Tesi[1]

ANALISI STATISTICA DELL' ECONOMIA SOMMERSA

SUPERVISORI: PROF.SSA GIOVANNA JONA LASINIO

PROF. BERNARDO MAGGI

DOTTORANDA: CECILIA MORVILLO

Anno Accademico 2019/2020

[1] Un sentito ringraziamento va ai miei supervisori per la pazienza, la costanza e la professionalità dimostrate durante il percorso di dottorato, e al Collegio docenti del Dipartimento di Statistica, per le utili indicazioni fornite durante l'incontro del 3 luglio 2020. Un particolare ringraziamento va ai revisori della tesi, Prof. Antonello Maruotti e Prof. Vittorio Nicolardi, per il prezioso contributo fornito per migliorare la qualità della stessa.
Si ringraziano inoltre: Antonio Affuso, Giorgia Marinuzzi e Giulio Palomba (autore delle dispense "Panel data" http://utenti.dises.univpm.it/palomba/Mat/PanelData.pdf).
La tesi è stata discussa il 24.5.2021 di fronte alla Commissione composta da: Prof.ssa Alessandra Luati, Prof. Claudio Agostinelli, Prof. Bruno Scarpa.

INDICE:

INTRODUZIONE

L'elaborato affronta le problematiche riguardanti l'Economia Sommersa attraverso un approccio modellistico. L'obiettivo è fornire un contributo utile alla comunità scientifica economica e statistica che si occupa del tema. Avvalendosi di un esame della letteratura esistente, la tesi propone una metodologia per l'analisi di una problematica ignota. Sono inoltre presenti riflessioni sulla policy implication. La questione tempo è sempre decisiva, ma lo è ancora di più in un periodo come questo, in cui vi è una necessità stringente di una ripresa economica. In estrema sintesi, la filosofia della ricerca è incardinare la tematica dell'Economia Sommersa all'interno di un metodo conosciuto ed efficiente.

Simili studi sono anche indispensabili per supportare l'attività di Governo, che già ampiamente opera sul fenomeno. Infatti, il Decreto legislativo 24 settembre 2015 n. 160, recante disposizioni in materia di stima e monitoraggio dell'evasione fiscale e monitoraggio e riordino delle disposizioni in materia di erosione fiscale[2], ha previsto che, contestualmente alla Nota di aggiornamento del Documento di economia e finanza[3], sia presentato un Rapporto sui risultati conseguiti in materia di misure di contrasto all'evasione fiscale e contributiva. Per la redazione di tale Rapporto, il Governo si avvale della "Relazione sull'economia non osservata e sull'evasione fiscale e contributiva". Tale Relazione è predisposta da una Commissione istituita con Decreto del Ministro dell'economia e delle finanze[4], e fornisce ogni anno una stima ufficiale delle entrate tributarie e contributive sottratte al bilancio pubblico eseguendo una misurazione del divario (gap) tra le imposte e i contributi effettivamente versati e le imposte e i contributi che i contribuenti avrebbero dovuto versare in un regime di perfetto adempimento agli obblighi tributari e contributivi previsti a legislazione vigente.

L'azione per il contrasto all'evasione fiscale [5] è principalmente orientata a stimolare l'adempimento spontaneo dei contribuenti, potenziando lo sfruttamento delle nuove tecnologie in modo da favorire l'acquisizione di informazioni rilevanti per indirizzare controlli mirati ai contribuenti meno affidabili. Tale azione accresce in prospettiva la plausibilità di una riduzione dell'elevato carico fiscale sui contribuenti onesti, rafforzando gli incentivi al lavoro e al doing business. Pertanto, è necessario tener conto dei progressi della tecnologia nella creazione delle base dati, in grado di tracciare le transazioni e incrociare le dichiarazioni fiscali; dei risultati della riflessione economica sugli incentivi a evadere, sulla propensione a colludere nel consumo finale e sulla sostenibilità di mercato di un tessuto produttivo frammentato e marginale; dei risultati ottenuti con le misure di semplificazione amministrativa e di miglioramento del rapporto con il contribuente.

Il presente lavoro è così organizzato:

- Il Capitolo 1 approfondisce il concetto di Economia Sommersa, esamina nel dettaglio le diverse metodologie di stima e fornisce un'analisi della letteratura economica nazionale e internazionale;

[2] In attuazione degli articoli 3 e 4 della Legge 11 marzo 2014, n. 23.

[3] In attuazione dell'articolo 10-bis.1 c. 3 Legge 31 dicembre 2009, n. 196, viene presentato un rapporto sui risultati conseguiti in materia di misure di contrasto all'evasione fiscale e contributiva, distinguendo tra imposte accertate e riscosse, nonché tra le diverse tipologie di avvio delle procedure di accertamento, in particolare evidenziando i risultati del recupero di somme dichiarate e non versate e della correzione di errori nella liquidazione sulla base delle dichiarazioni, evidenziando, ove possibile, il recupero di gettito fiscale e contributivo attribuibile alla maggiore propensione all'adempimento da parte dei contribuenti.

[4] La Commissione è composta da rappresentanti delle seguenti amministrazioni: Ministero Economia e Finanze, INPS, Ministero del Lavoro e delle Politiche Sociali, Università di Roma "La Sapienza", Università degli studi "Roma Tre", ISTAT, ANCI, Agenzia delle Entrate, Banca d'Italia, Conferenza delle Regioni e delle Province autonome, Presidenza del Consiglio dei Ministri, Guardia di Finanza.

[5] Nota di Aggiornamento del Documento di Economia e Finanza del 2019 – Allegato – Relazione sull'economia non osservata e sull'evasione fiscale e contributiva – anno 2019

- Il Capitolo 2 illustra la metodologia utilizzata. In particolare, sono definiti i dati panel e sono mostrati esempi esplicativi, in campo internazionale e italiano, sull'uso degli stessi. Ci si sofferma inoltre sui vantaggi e sugli svantaggi nell'utilizzo di una tale struttura di dati. Il Capitolo termina con una panoramica dei modelli con dati panel attualmente esistenti in letteratura;
- Il Capitolo 3 e il Capitolo 4 esaminano la situazione dell'Economia Sommersa nel caso italiano e nel caso europeo;
- Infine, il Capitolo 5 utilizza una strategia per rispondere al problema di possibile presenza di endogeneità e autocorrelazione seriale, e si pone il problema di capire come un intervento pubblico possa influire sul fenomeno dell'Economia Sommersa. Sono stati quindi introdotti concetti e metodi in grado di studiare gli effetti delle determinanti sull'Economia Sommersa nel tempo, analizzando il comportamento del fenomeno nel breve, medio e lungo periodo.

1 - L'Economia Sommersa (ES)

L'intervento pubblico nel mercato è un fenomeno che accompagna il progresso delle società civili. Le prime tracce storiche, in tema di regolamentazione delle attività economiche e di tassazione, risalgono alle civiltà mesopotamiche, con il codice di Hammurabi[6] (1792-1750 a.C.), e agli antichi Egizi. Questi ultimi sono stati capaci di creare un sistema fiscale nel quale gli scribi svolgevano la funzione di esattori (Viel, 2015; Adams, 2007). Con l'evolversi della società, si è assistito a una progressiva espansione dell'intervento dello Stato nel mercato. Mentre all'inizio l'operatore pubblico aveva soltanto il compito di garantire la sicurezza e il regolare funzionamento del mercato, evitando un coinvolgimento troppo elevato, con il passare del tempo l'operatore pubblico ha cominciato a essere parte attiva. L'autorità pubblica, per conseguire i propri obiettivi, poteva incidere discrezionalmente sugli equilibri spontanei del mercato attraverso interventi che incidevano direttamente nel bilancio pubblico (tassazione, spesa pubblica e indebitamento) e attraverso la regolamentazione, vista come provvedimento in grado di influenzare il comportamento e il livello di benessere degli agenti economici. Gli individui, di fronte alle pretese sempre più esigenti dello Stato, hanno cominciato a valutare le diverse scelte scegliendo ciò che sarebbe stato più vantaggioso. L'individuo, dunque, può decidere di non seguire le imposizioni (fiscali, previdenziali ecc), dettate dallo Stato, cercando di occultare beni, prestazioni e scambi per sfuggire alla normativa e non pagare imposte e dazi (Morlacchi, 2014). L'insieme dei comportamenti scorretti messi in atto dagli individui e dalle imprese per occultare la loro ricchezza e la produzione e lo scambio di beni e servizi, così da non adempiere i propri doveri verso l'operatore pubblico, è definita **Economia Sommersa (ES)**. È questo un fenomeno estremamente complesso che influisce fortemente sullo sviluppo economico, distorcendo e togliendo efficienza al normale funzionamento della concorrenza e del mercato, poiché assorbe risorse altrimenti destinate al bilancio pubblico. Provoca inoltre iniquità, in quanto determina una riduzione del gettito che può pregiudicare la qualità e la quantità dei servizi pubblici offerti a tutti i cittadini, compresi coloro che regolarmente contribuiscono attraverso il pagamento delle imposte. L'Economia Sommersa si esplicita attraverso diverse forme; coglierla non è pertanto affatto semplice e agevole, poiché per definizione essa è non osservabile (Gruppo di lavoro MEF - gennaio/giugno 2011).

1.1 Definizione di Economia Sommersa (ES)

Nel corso del tempo l'attenzione degli economisti si è rivolta alla ricerca di un'idonea definizione di Economia Sommersa che permettesse di definire un'esatta misurazione della stessa. Tra le diverse definizioni troviamo quella secondo cui fanno parte dell'Economia Sommersa[7] "*tutte le attività economiche che contribuiscono al calcolo del prodotto nazionale lordo di un paese, ma non sono ufficialmente registrate*" (Feige, 1989; Schneider, 2004, 2005, 2013; Schneider e Williams, 2013; Schneider e Buehn2016). Smith (1997) definisce l'Economia Sommersa come "*produzione basata sul mercato di beni e servizi, legali o illegali, che sfuggono al rilevamento nelle stime ufficiali del PIL*". Schneider (Schneider e Williams, 2013; Schneider e Buehn, 2016) nei suoi lavori utilizza la seguente definizione: "*L'Economia Sommersa include tutta la produzione legale basata sul mercato di beni e servizi che è deliberatamente nascosta alle autorità pubbliche per evitare: i) il pagamento di tasse (imposte sul reddito o imposte sul valore aggiunto); ii) il pagamento dei contributi di*

[6]Il codice di Hammurabi è una raccolta di 282 disposizioni ordinate dal re Hammurabi di Babilonia. Il corpus legale è suddiviso in capitoli che riguardano varie categorie sociali e di reati, e abbraccia molte delle possibili situazioni dell'umano convivere del tempo, dai rapporti familiari a quelli commerciali ed economici, dall'edilizia alle regole per l'amministrazione del regno e della giustizia. Le leggi sono notevolmente dettagliate, e questo ha fornito un aiuto prezioso agli archeologi, consentendo loro di ricostruire importanti aspetti pratici della società mesopotamica. L'importanza del codice di Hammurabi risiede nel fatto che si tratta di una delle prime raccolte organiche di leggi a noi pervenuta.

[7]Sono stati usati nella letteratura specialistica molti termini: economia "ombra" (shadow), "sotterranea" (underground), "nera/grigia" (black/grey), "non registrata" (unrecorded), "non ufficiale" (unofficial), "informale" (informal), "non osservata" (unobserved), "clandestina" (clandestine), "secondaria" (secondary) e "parallela" (parallel).

sicurezza sociale; iii) certi standard legali del mercato del lavoro, come i salari minimi, l'orario di lavoro massimo, le norme di sicurezza, ecc.; iv) l'adempimento di determinate procedure amministrative, come il completamento di questionari statistici o altre forme amministrative" (Schneider e Williams, 2013). Per chiarire il concetto, l'autore utilizza due tabelle, che forniscono una definizione più ampia di Economia Sommersa (Tabella 1) ed esempi di attività da ritenere inclusi o esclusi dalla definizione di Economia Sommersa (Tabella 2).

Tabella 1: Tassonomia dei tipi di attività dell'Economia Sommersa

Tipo di attività	**Transazione monetaria**		**Transazione non monetaria**	
Attività illegali	Commercio di beni rubati; produzione e spaccio di droga; prostituzione; gioco d'azzardo; contrabbando; frode; ecc.		Scambio di droghe; beni rubati; contrabbando; Produzione o crescita di droghe per uso personale; Furto per uso personale.	
	Evasione fiscale	**Elusione fiscale**	**Evasione fiscale**	**Elusione fiscale**
Attività legali	Reddito non dichiarato di lavoratori autonomi; stipendi, salari e risorse provenienti da lavoro non dichiarato di beni e servizi legali	Agevolazioni e indennità accessoria	Scambio di beni e servizi legali	Lavoro in proprio

Fonte: Schneider e Williams, 2013

Tabella 2: Attività ed Economia Sommersa

Attività	**Fuori o dentro la misura di Economia Sommersa**	**Motivo nel caso di esclusione**
Custodia dei bambini con reddito non dichiarato	dentro	
Spaccio di droga	fuori	Attività illegale
Pagamento in contanti, reddito non dichiarato	dentro	
Lavori di costruzione fatti da proprietario di casa	fuori	Le attività in proprio non sono soggette a tassazione e regolamentazione
Acquisto di sigarette contrabbandato dal Paese dell'UE	dentro	
Produzione contraffatta di un prodotto legale come ad esempio le sigarette	dentro	

Fonte: Schneider e Williams, 2013

Più in generale, l'Economia Sommersa riguarda quell'insieme di attività produttive la cui caratteristica principale è quella di sfuggire all'osservazione, alla regolamentazione e alla rilevazione, sia che comporti transazioni monetarie (produzione e distribuzione), che transazioni non monetarie (autoproduzione, scambio e baratto). Sono quindi sommerse tanto le attività produttive legali, ma svolte in modo irregolare, quanto le attività illegali, per le quali si verifica una violazione della legge (Lucifora, 2003; Dell'Arno, 2003; Dell'Arno e Schneider, 2003).

La definizione più comune, recepita anche dall'ISTAT[8], rappresenta il fenomeno dell'ES come l'insieme di tutte le attività economiche non registrate, che sfugge a ogni rilevazione statistica e ai controlli fiscali. Queste

[8]L'ISTAT definisce l'Economia Sommersa come l'aggregato che include tutte quelle attività che sono volontariamente celate alle autorità fiscali, previdenziali e statistiche. Esso è generato da dichiarazioni mendaci riguardanti sia fatturato e costi delle unità produttive (in modo da generare una sotto-dichiarazione del valore aggiunto), sia l'effettivo utilizzo di input di lavoro (ovvero l'impiego di lavoro irregolare). Ulteriori integrazioni derivano: dalla valutazione delle mance che i lavoratori dipendenti ricevono dai clienti in alcune attività economiche; dai risultati della procedura di riconciliazione delle stime indipendenti dell'offerta e della domanda di beni e servizi; dalla valutazione degli affitti in nero (ISTAT 2015, 2017, 2018).

attività irregolari nascoste rientrano nel calcolo ufficiale del Prodotto Interno Lordo (Schneider ed Enste, 2000).

1.2 L'Economia Sommersa - ISTAT

Da un punto di vista statistico, le attività economiche non svolte nel rispetto delle norme di legge o delle regole fiscali rappresentano un vasto sottoinsieme della cosiddetta economia non direttamente osservata (Non-Observed Economy, NOE). Essa è formata dall'insieme di attività che non inviano "volontariamente" segnali statistici attraverso i quali sia possibile rilevarne l'entità, ma che contribuiscono alla formazione del prodotto in quanto generano valore aggiunto tramite atti di scambio svolti tra soggetti consenzienti. L'Economia Sommersa è integralmente inclusa nell'economia non osservata, e rappresenta l'insieme delle attività legali svolte contravvenendo a norme fiscali o contributive. Per rendere attendibile il confronto internazionale dei livelli e delle dinamiche del prodotto, i nuovi sistemi di Contabilità nazionale, adottati all'incirca dalla seconda metà degli anni Novanta, impongono a tutti i Paesi di contabilizzare l'economia non osservata nel PIL. Tutte le componenti sono state quindi oggetto di stima e molte di esse sono attualmente incluse nei conti economici nazionali: fa eccezione, come vedremo, la quota di economia illegale da includere nell'economia non osservata, che non ha ancora prodotto stime ritenute affidabili (Rassegna economica, 2013).

L'ISTAT (ISTAT, 2010, 2015, 2017, 2018) elabora le stime del PIL e dell'occupazione attribuibili alla parte dell'economia non osservata. Quest'ultima deriva dall'attività di produzione di beni e servizi che, pur essendo legale, sfugge all'osservazione diretta in quanto connessa al fenomeno della frode fiscale e contributiva. Tale componente è già compresa nella stima del PIL e negli aggregati economici diffusi dall'ISTAT a livello sia nazionale sia territoriale. Secondo i criteri dell'Unione europea (Regolamento 2223/96 UE), solo una misura esaustiva del PIL rende tale aggregato confrontabile fra i vari Paesi e utilizzabile come indicatore per il calcolo dei contributi che gli Stati membri versano all'Unione, per il controllo dei parametri di Maastricht e per l'attribuzione dei fondi strutturali.

Con l'introduzione del sistema SEC2010, per la compilazione dei conti nazionali dei Paesi aderenti all'Unione europea, l'ISTAT ha operato un importante rinnovamento delle fonti informative e dei metodi di stima. Uno degli sviluppi più rilevanti ha interessato le metodologie di misurazione di diverse componenti dell'economia non osservata, che hanno beneficiato anche di una serie di sviluppi nelle fonti informative sui dati d'impresa e di importanti innovazioni nei processi di stima dell'occupazione e dei redditi.

L'inclusione delle diverse componenti dell'economia non osservata nei conti nazionali non solo consente di rispettare il principio dell'esaustività nella rappresentazione dei flussi economici (stabilito nei manuali internazionali SNA e SEC e verificato dalle autorità statistiche europee), permettendo una migliore comparabilità internazionale dei dati, ma contribuisce anche a migliorare e rendere più trasparenti le stime dei principali aggregati economici, il prodotto interno lordo e il reddito nazionale lordo.

Le maggiori componenti dell'economia non osservata sono rappresentate dal **sommerso economico**, dall'**economia illegale**, dal **sommerso statistico** e dall'**economia informale**:

- Il **sommerso economico** include tutte quelle attività che sono volontariamente celate alle autorità fiscali, previdenziali e statistiche. Esso è generato da dichiarazioni mendaci riguardanti sia il fatturato

e/o i costi delle unità produttive (in modo da generare una sotto-dichiarazione del valore aggiunto), sia l'utilizzo di input di lavoro[9].

- L'**economia illegale** è definita dall'insieme delle attività produttive aventi per oggetto beni e servizi illegali, o che, pur riguardando beni e servizi legali, sono svolte senza adeguata autorizzazione o titolo.
- Il **sommerso statistico** include tutte quelle attività che sfuggono all'osservazione diretta per motivi riferibili alle inefficienze informative che caratterizzano le basi di dati (errori campionari e non campionari) o per errori di copertura negli archivi[10].
- L'**economia informale** include, infine, tutte quelle attività produttive svolte in contesti poco o per nulla organizzati, basati su rapporti di lavoro non regolati da contratti formali, ma nell'ambito di relazioni personali o familiari.

La stima del **sommerso economico** nei conti nazionali è stata profondamente rinnovata, sia per quel che riguarda la componente di sotto-dichiarazione del valore aggiunto, sia per quel che concerne la valutazione del contributo produttivo del lavoro irregolare. In particolare, la sotto-dichiarazione del valore aggiunto è connessa al deliberato occultamento di una parte del reddito da parte delle imprese, attraverso dichiarazioni volutamente errate del fatturato e/o dei costi alle autorità fiscali (con un analogo comportamento riscontrato nelle rilevazioni statistiche ufficiali). Per quanto riguarda la misura del lavoro come fattore di produzione, il sistema europeo dei conti raccomanda di stimare in modo esaustivo l'input di lavoro espresso non solo in termini di occupati, ma anche di posizioni lavorative, ore effettivamente lavorate e unità di lavoro. L'insieme delle Unità di Lavoro (ULA) è pari al numero di posizioni lavorative equivalenti a tempo pieno e include sia le posizioni lavorative regolari sia quelle riconducibili a prestazioni di lavoro svolte in forma non regolare. In occasione del passaggio al sistema SEC2010, l'accresciuta disponibilità di fonti amministrative per usi statistici ha consentito di sviluppare una metodologia di stima dell'input di lavoro fortemente basata sull'uso integrato di dati individuali da rilevazioni statistiche e amministrative. Una volta individuato l'ammontare di ore non regolari impiegate nel processo produttivo, si misura il valore aggiunto che esse generano.

Le componenti appena descritte, pur rappresentandone la parte più rilevante, non esauriscono la misurazione del fenomeno del sommerso economico. Ulteriori integrazioni derivano: (1) dalla valutazione delle mance che i lavoratori dipendenti ricevono dai clienti in alcune attività economiche (alberghi e ristoranti, parrucchieri, taxi) e che dovrebbero essere considerate parte del fatturato; (2) dai risultati della procedura di riconciliazione delle stime indipendenti dell'offerta e della domanda di beni e servizi; (3) dalla valutazione degli affitti in nero.

1.3 Metodi di misurazione dell'Economia Sommersa

I primi tentativi di misurazione e quantificazione dell'Economia Sommersa risalgono agli anni Sessanta. In particolare sono stati sviluppati tre diversi approcci (Campanelli, Comitato per l'emersione del lavoro non regolare - Presidenza del Consiglio dei Ministri; Guardia di Finanza, 2008; Schneider ed Enste, 2000; Schneider e Buehn, 2016; Zizza, 2002): il primo, **metodo diretto**, privilegia l'esame dell'attività sommersa sul campo (ad esempio con indagini di tipo campionario realizzate tramite interviste a imprenditori o testimoni privilegiati);

[9]Le principali definizioni sull'input di lavoro (SEC2010) riguardano gli occupati interni, le posizioni lavorative, le ore lavorate e le unità di lavoro. L'approccio italiano alla stima dell'input di lavoro consente di calcolare le posizioni lavorative e le corrispondenti unità di lavoro, che rappresentano la trasformazione a tempo pieno delle prestazioni lavorative offerte, per diverse categorie lavorative, regolari e non regolari, individuabili integrando e confrontando fonti statistiche diverse o utilizzando metodi indiretti di stima.

[10]L'incidenza del sommerso statistico è stata ridotta significativamente grazie alle innovazioni nelle fonti informative sui conti economici delle imprese. La stima della componente regolare dell'economia è stata ottenuta attraverso l'elaborazione di una nuova base dati annuale di tipo censuario, che contiene informazioni individuali per tutto l'universo delle imprese attive. Questo nuovo prodotto statistico (denominato Frame-SBS) nasce da una complessa procedura di integrazione di dati d'indagine e amministrativi e per le principali variabili non è affetto da errori campionari.

il secondo, **metodo indiretto**, si propone di misurare la diffusione dell'Economia Sommersa confrontando le diverse fonti statistiche e amministrative disponibili; infine il terzo, **metodo econometrico**, si basa su modelli matematici che misurano l'entità dell'Economia Sommersa mettendola in relazione ad alcune sue cause (tasso di disoccupazione, livello di sviluppo, livello di tassazione, indice di vecchiaia, ecc.).

1.3.1 Metodi diretti

Tra i metodi che si propongono di analizzare direttamente il fenomeno dell'Economia Sommersa, vediamo i più significativi.

Le indagini campionarie

Le indagini campionarie sono realizzate, generalmente, attraverso l'estrazione di un campione casuale di lavoratori o imprese, al quale è somministrato un questionario appositamente predisposto. L'uso di questo metodo ha il vantaggio di ottenere informazioni sull'Economia Sommersa in tempi relativamente brevi, ma presenta alcune debolezze: 1) l'affidabilità dei dati dipende dal grado di veridicità delle risposte da parte degli intervistati; 2) l'uso di questionari strutturati non permette di rilevare aspetti nuovi rispetto a quelli individuati inizialmente; 3) l'incompletezza degli elenchi ufficiali disponibili delle imprese implica la problematica dell'identificazione dell'universo statistico.

Il metodo dei testimoni privilegiati

Tale metodo consiste nella conduzione di interviste, attraverso la somministrazione di questionari, a persone che, per posizione professionale ed esperienza, sono in possesso di informazioni sul fenomeno oggetto d'analisi. Il metodo consente di superare il problema della reticenza, in quanto gli interlocutori, poiché non direttamente coinvolti nel fenomeno, parlano più liberamente. È molto utile nelle prime fasi di uno studio, in quanto permette di acquisire nuove informazioni del fenomeno. Di contro, la realtà raccontata dai testimoni potrebbe essere alterata dalla cultura di appartenenza, dalle specifiche ideologie e dai propri valori, rischiando, in tal modo, di cogliere solo quegli aspetti altamente correlati con il punto di vista dell'interlocutore. Vi è inoltre la difficoltà di individuare i testimoni privilegiati.

L'utilizzo di dati amministrativi provenienti dai controlli fiscali e contributivi

Si tratta di dati provenienti dall'attività di vigilanza sulle imprese effettuata dall'Agenzia delle Entrate, dalla Guardia di Finanza, dall'Ispettorato del Lavoro e dall'INPS. Tali rilevazioni non si prestano a essere utilizzate ai fini statistici per diversi motivi, primo fra tutti il fatto che i risultati delle ispezioni dipendono dalla capacità di scoprire cose che il controllato cerca di occultare. In pratica, poiché l'inadempienza a certi obblighi comporta delle sanzioni, la disponibilità del contribuente a collaborare risulta inadeguata, con la conseguenza che le informazioni fornite non possono essere ritenute sufficientemente attendibili.

Il modello di Feinstein

Il modello matematico creato da Feinstein (Feinstein, 1999) è basato sull'analisi delle caratteristiche del processo di indagine ed è in grado di stimare il numero delle infrazioni reali. Il modello si basa su due equazioni vettoriali, la cui stima consente di calcolare la probabilità che non sia scoperta una violazione commessa e la probabilità di errata segnalazione di infrazione. La prima equazione è riferita al potenziale violatore e specifica la probabilità di commettere la violazione:

$$Y_{1i}=X_{1i}\,\beta_1+\varepsilon_{1i}$$

$$L_{1i}=1 \text{ (violazione) se } Y_{1i}>0$$

$$L_{1i}=0 \text{ (non violazione) se } Y_{1i}\leq 0$$

dove:

X_{1i}= vettore delle caratteristiche del potenziale violatore; β_1= vettore dei parametri; ε_{1i}= errore di media nulla con distribuzione F.

La seconda equazione è riferita al controllore ed esamina la possibilità di scoperta dell'infrazione condizionata all'avvenuta violazione:

$$Y_{2i}=X_{2i}\,\beta_2+\varepsilon_{2i}$$

$$L_{2i}=1 \text{ (detenzione) se } Y_{2i}>0$$

$$L_{2i}=0 \text{ (non detenzione) se } Y_{2i}\leq 0$$

dove:

X_{2i}= vettore delle caratteristiche del processo di indagine; β_2 = vettore dei parametri; ε_{2i}= errore di media nulla con distribuzione G.

L_{1i} e L_{2i} non sono osservabili separatamente, mentre è osservabile il loro prodotto, $L_{1i}L_{2i}$, che rappresenta la violazione rilevata.

La stima contemporanea delle due equazioni, può essere fatta attraverso il metodo della massima verosimiglianza, e consente di calcolare la probabilità di non scoperta di una violazione commessa e quella di errata segnalazione di infrazione. Dal numero di infrazioni scoperte si eliminano quindi quelle false e si aggiungono quelle non scoperte, giungendo a una stima del numero delle infrazioni totali. La principale forza di tale modello è di porre particolare attenzione al processo di controllo incorporandolo all'interno dell'analisi, mentre molte analisi non ne tengono conto. La principale debolezza consiste, invece, nel fatto che il modello è di natura statistica e non si basa su informazioni dettagliate sulla non detenzione. Il maggior problema riguarda la difficoltà di definizione dei parametri e delle distribuzioni nel modello.

1.3.2 Metodi indiretti

Questo approccio implica l'utilizzo di strumenti di carattere macro-economico.

Metodi della differenza tra grandezze diverse

I metodi della differenza si basano sul confronto dei diversi valori di una stessa grandezza economica, rilevati da fonti indipendenti ipotizzando che una delle due grandezze rappresenti il valore totale effettivo della variabile considerata mentre l'altro il suo valore ufficiale.

- Differenza tra produzione e impiego del reddito: mette a confronto il valore della produzione (rilevato presso le imprese) e delle importazioni con la somma di consumi (rilevati presso le famiglie), investimenti, esportazioni ecc. Poiché i redditi posseduti dalle famiglie possono essere acquisiti anche con attività illecite o di tipo informale, questo metodo andrà a cogliere l'intera economia non osservata.

- Differenze tra reddito reale e reddito dichiarato: confronta i dati provenienti dalle stime di contabilità nazionale (considerate come valore reale del reddito) e quanto viene dichiarato al fisco per il pagamento dell'Irap. Una volta rese omogenee le due grandezze, la differenza esistente rappresenta l'ampiezza dell'evasione fiscale.
- Differenza tra occupati dal lato domanda e dal lato offerta: il metodo delle differenze può essere utilizzato anche per il calcolo della quantità di lavoro irregolare. L'ISTAT la calcola, infatti, attraverso il confronto fra la rilevazione trimestrale delle forze di lavoro (lato famiglie) e le rilevazioni relative alle imprese. In tal caso il lavoro che si viene a rilevare è unicamente quello irregolare (legale ma non dichiarato). Per costruire la stima del lavoro sommerso, sono armonizzate e integrate le varie fonti di informazioni, sia dal lato della domanda sia da quello dell'offerta, al fine di ottenere il numero delle posizioni lavorative. Tali stime vengono poi confrontate per ottenere il numero degli irregolari e dei non dichiarati. Le posizioni lavorative vengono, infine, trasformate in unità di lavoro equivalente a tempo pieno (ULA). A partire dalla quantità di lavoro irregolare individuato, l'ISTAT calcola il valore dell'Economia Sommersa attribuendo ai lavoratori dipendenti irregolari la stessa produttività di quelli regolari e correggendo i conti delle imprese in modo che il reddito da lavoro autonomo sia almeno uguale a quello da lavoro dipendente.
- Il tasso di partecipazione delle forze lavoro: a differenza dei metodi precedentemente mostrati, il metodo del tasso di partecipazione delle forze lavoro non parte da due misurazioni indipendenti della medesima grandezza, bensì dal confronto tra una misurazione del tasso di occupazione regolare con un valore che si considera normale. Tutto ciò sulla base dell'ipotesi che un declino del tasso di partecipazione delle forze lavoro sia un segnale di presenza di sommerso.

Modelli monetari

Fanno parte di questa famiglia una serie di metodi che vedono nella domanda di moneta lo strumento prioritario per stimare l'Economia Sommersa. Si ritiene, infatti, che se esiste un eccesso di moneta in circolazione, rispetto alle necessità dell'economia regolare, ciò sia dovuto alle necessità di economia irregolare. Riuscendo a calcolare questo eccesso, si giunge a trovare l'ampiezza dell'Economia Sommersa.

- Il metodo delle transazioni: sviluppato da Feige (Feige, 1979, 1986; Documento redatto dal Ministero del Lavoro, 2002) e applicato negli Stati Uniti, si basa sull'assunzione che ci sia un rapporto costante nel tempo tra volume totale delle transazioni (avvenute sia in contanti sia per mezzo di assegni) e reddito (PNL ufficiale). Feige appoggia la sua dimostrazione sull'equazione quantitativa della moneta MV=PQ di Fisher (con M quantità di moneta, V velocità di circolazione della moneta, P livello dei prezzi, Q indice delle transazioni, cioè delle quantità di beni prodotte e scambiate). Conoscendo e assumendo costante la velocità di circolazione della moneta e conoscendo M (moneta circolante più depositi a vista), è possibile trovare la quantità PQ, equivalente al valore totale delle transazioni (composto da transazioni finanziarie, intermedie, di beni finali e transazioni nell'economia irregolare). Per calcolare il rapporto corretto, presunto costante, tra volume delle transazioni e reddito ufficiale, Feige deve assumere come parametro di riferimento l'anno 1939, anno in cui egli suppone che non vi fosse Economia Sommersa, e in cui quindi il rapporto tra transazioni e reddito totale (in questo caso, non essendoci sommerso il reddito totale coincide con il reddito osservato) si ritenesse normale. Sulla base di questo dato, Feige rileva l'ampiezza del sommerso negli anni 1976 e 1978, nei quali nota che il rapporto tra transazioni e reddito ufficiale è cresciuto. Per stimare il volume del reddito irregolare, egli divide il volume delle transazioni rilevato in questi due anni per il rapporto transazioni/reddito ufficiale del 1939, e al risultato sottrae il reddito ufficiale per ognuno dei due anni

di riferimento, ottenendo in questo modo le due stime dell'Economia Sommersa per gli anni 1976 e 1978. Punti di debolezza: a) l'assunzione di un anno base senza Economia Sommersa; b) l'assunzione che la velocità di circolazione della moneta e degli assegni rimanga stabile nel tempo; c) l'assunzione che la moneta sia utilizzata solo a scopi transattivi; d) non c'è evidenza che il rapporto tra transazioni e reddito ufficiale sia costante nel tempo.

- Metodo della domanda di moneta: ideato da Cagan, ripreso da Gutmann e infine perfezionato da Tanzi (Tanzi, 2013), questo metodo, che stima la domanda di moneta circolante rispetto alla domanda di depositi a vista, è sicuramente uno dei metodi più conosciuti. Le assunzioni importanti su cui si basa sono le seguenti: a) le transazioni nell'Economia Sommersa sono interamente effettuate in contanti, per non lasciare tracce alle autorità; quindi, a un incremento dell'Economia Sommersa corrisponde un aumento della domanda di moneta circolante; b) l'Economia Sommersa è principalmente causata dall'elevata tassazione; c) il rapporto tra la moneta circolante e i depositi a vista è costante nel tempo. Tanzi utilizza un'equazione particolare per calcolare la domanda di moneta. L'equazione considera il rapporto tra moneta circolante e depositi a vista, e include una serie di variabili tra le quali il livello di tassazione. Per mezzo di questa equazione viene inizialmente stimata la domanda di contanti con la tassazione del momento, e poi viene stimata la domanda di contanti nello stesso anno, ma con un carico fiscale più basso o nullo (quindi in una situazione ipotetica in cui non vi è Economia Sommersa). La differenza tra queste due stime fornirà la quantità di circolante utilizzata nel sommerso. Infine, assumendo che la velocità di circolazione della moneta sia la stessa, sia nell'economia regolare che in quella sommersa, è possibile calcolare il reddito generato dall'Economia Sommersa. Vi sono però diverse obiezioni a questo metodo: a) non tutte le transazioni sono pagate in contanti; b) il carico fiscale può non essere l'unica causa dell'Economia Sommersa (altre cause possono essere l'impatto dei regolamenti, l'attitudine del contribuente verso lo stato, ecc); c) i fattori che determinano la domanda di moneta sono instabili, diversi da Paese a Paese e talvolta anche all'interno dello stesso Paese; d) è difficile sostenere che la velocità di circolazione è identica sia nell'economia ufficiale che in quella irregolare, in quanto è complicato stimarla; e) è difficile sostenere l'assenza di Economia Sommersa in un ipotetico anno di riferimento in cui la tassazione è nulla; f) il rapporto tra circolante e depositi a vista difficilmente può rimanere stabile (la scarsa richiesta di tali depositi può essere dovuta alla proliferazione di depositi di tipo alternativo).
- Metodo degli input fisici: si basa sulla rilevazione del consumo di input misurabili determinando la coerenza o la differenza con le stime ufficiali del PIL. Kaufmann e Kaliberda (Kaufmann e Kaliberta, 1996) ritengono l'elettricità il migliore indicatore di ogni attività economica. L'assunzione di base di questo metodo è che l'attività economica è strettamente correlata con il consumo di energia, e l'elasticità del loro rapporto è approssimativamente vicina al valore uno. Assumendo che questo rapporto sia relativamente costante e conoscendo il valore complessivo dell'elettricità erogata, è possibile derivare una stima del PIL complessivo. La differenza tra questo PIL complessivo e il PIL ufficiale darà la stima dell'Economia Sommersa. La differenza tra la crescita ufficiale del PIL e la crescita del consumo di elettricità (differenza nulla in una situazione ideale in cui non c'è sommerso) è quindi attribuita alla crescita del sommerso. Ma anche questo metodo, che viene ritenuto adatto per la misurazione del sommerso nei Paesi in via di sviluppo, presenta alcuni limiti: a) molte attività economiche richiedono piccole quantità di elettricità; quindi le attività irregolari di questo tipo non verranno rilevate; b) il progresso tecnico ha permesso un utilizzo più efficiente di elettricità, sia nell'economia regolare sia in quella irregolare; c) possono esserci differenze nell'elasticità del rapporto tra elettricità consumata e PIL, soprattutto in quei Paesi che stanno conoscendo

cambiamenti strutturali; d) se vengono utilizzate altre forme di energia una parte del sommerso non verrà rilevata. Una variante a questo modello fu introdotta da Lackó (Lackò, 1999), la quale assunse che una certa parte dell'Economia Sommersa vada associata al consumo domestico di elettricità, dovuto alle attività produttive "fatte in casa". L'economia non ufficiale sarà quindi tanto più grande quanto maggiore è il consumo domestico di energia. Questo metodo evita il problema dei cambiamenti strutturali e l'elasticità unitaria del rapporto elettricità/PIL, ma richiede la conoscenza del valore di PIL prodotto da una unità di elettricità nell'Economia Sommersa, che, se sconosciuto, deve essere ricavato dalla stima di un Paese diverso da quello sotto indagine. Inoltre, considerando solo l'ambiente domestico, una buona parte di sommerso viene persa.

1.3.3 Metodi econometrici

Nell'ambito dei modelli econometrici, il modello di Frey (1983) è il primo a considerare un certo numero di fattori come principali cause dell'Economia Sommersa. I fattori che egli considera sono: il livello di tassazione; il livello di regolamentazione del lavoro; la moralità nel pagamento delle tasse; la percezione del disagio creato dalle tasse; il tasso di partecipazione ufficiale al lavoro; il tasso di disoccupazione; l'orario di lavoro ufficiale. Per ognuno dei fattori considerati Frey ha stilato una classifica di alcuni Paesi e ha poi calcolato una posizione media per ciascuno di essi, ponderando ogni fattore sulla base dell'importanza che a esso veniva attribuita dalla letteratura sull'Economia Sommersa (Marino, 2015). L'anno successivo Frey e Weck-Hanmeman (Frey e Weck-Hanmeman, 1984) hanno elaborato un modello di tipo MIMIC (Multiple Indicators and Multiple Causes), che metteva in relazione le cause e gli effetti dell'Economia Sommersa con la sua entità. Esso si basava sulla teoria della variabile latente non osservata stimata attraverso un modello di tipo LISREL (Linear Interdipendent Structural RELationship), composto da due parti: il modello di misurazione e quello delle equazioni strutturali. Le variabili determinanti considerate erano il peso della tassazione[11], la moralità rispetto al pagamento delle tasse (calcolato attraverso indagini specifiche), il tasso di disoccupazione e il livello di sviluppo. Gli indicatori degli effetti dell'Economia Sommersa venivano quindi messi in relazione alla variazione del numero di lavoratori regolari. Stimati i parametri del modello attraverso il metodo della massima verosimiglianza si otteneva infine l'ampiezza dell'economia irregolare.

Nel periodo a cavallo tra la fine del 1990 e l'inizio del 2000 sono stati pubblicati nuovi lavori che hanno dato ampio spazio a riflessioni sulla materia. Tra queste ricordiamo un breve saggio (Tanzi, 1999) sulla misurazione dell'Economia Sommersa, la ricca rassegna presente sui problemi di misurazione, stima e implicazioni del settore del sommerso (Scheneider ed Enste, 2000), che mira a far emergere i fattori rilevanti che spiegano il fenomeno oggetto di studio, e un accurato approfondimento (Lucifora, 2003) in grado di spaziare dalla ricerca teorica a quella applicata, considerando allo stesso tempo le politiche adottate fino al 2003 sull'Economia Sommersa.

1.4 Breve rassegna della letteratura più recente

L'approccio modellistico ha riscosso negli ultimi anni molto successo in quanto è in grado di descrivere l'Economia Sommersa attraverso le sue cause, non limitandosi solamente all'analisi degli aspetti puramente fiscali, ma individuando anche fattori di carattere sociale ed economico che in misura diversa influenzano il fenomeno. In letteratura sono presenti una serie di lavori che seguono tale orientamento.

[11]Il peso della tassazione viene diviso in oggettivo e percepito. Il peso oggettivo viene misurato come livello della tassazione sul PIL, il peso percepito viene invece misurato come tasso di crescita del livello di tassazione e il peso delle regolamentazioni.

Iniziamo con l'analizzare (Radovanovic, 2017) i punti di vista di tre dei maggiori studiosi della materia: Roberta Zizza, direttrice delle politiche monetarie presso la Banca d'Italia, la quale in un suo rapporto (Zizza, 2002) elenca le cause del sommerso, non solo basandosi sulla teoria economica ma anche attraverso un esercizio empirico riferito al caso italiano; Friedrich Schneider, professore di economia presso l'Università Johannes Kepler di Linz in Austria e dal 2006 ricercatore presso l'Istituto tedesco di ricerche economiche, e Dominik H. Enste, professore di economia presso l'Università di Colonia in Germania.

Le ragioni che spingono a entrare nel sommerso sono molteplici e variano da nazione a nazione.

1.4.1 Cause dell'Economia Sommersa: il punto di vista di Roberta Zizza

Roberta Zizza, nel suo rapporto (Zizza, 2002), elenca le seguenti cause:

1. **L'imposizione fiscale e contributiva** come causa dell'Economia Sommera è, all'interno della letteratura, riconosciuta come la causa principale ed è rappresentata dall'evasione delle imposte dirette e indirette. Il principio sul quale si basa questa causa è che maggiore è la pressione fiscale, maggiore sarà l'incentivo per il lavoratore e il datore di lavoro a operare al di fuori delle norme. Secondo gli autori Allingham e Sandmo (Allingham e Sandmo, 1972) deve esserci coerenza tra la pressione fiscale e i suoi benefici. Se l'onere fiscale è troppo elevato rispetto a ciò che lo Stato fornisce alle imprese e alle persone, queste faranno di tutto per evadere le imposte.
2. **I fattori istituzionali** hanno un ruolo importante secondo Zizza (2002). Con il termine fattori istituzionali si intendono le diverse autorità che hanno lo scopo di monitorare le attività economiche. Una parte consistente del sommerso viene alimentata dallo scarso livello di controllo o dall'eccessiva permissività, da parte delle autorità predisposte, all'accertamento della corretta applicazione delle norme di impresa.
3. **L'eccesso di regolamentazione e burocrazia**. Un metodo per capire se la burocrazia di uno Stato sia più o meno "pesante", è vedere il numero di leggi e autorizzazioni necessarie per lo svolgimento dell'attività di un'impresa nel mercato del lavoro. Oltre al semplice numero bisogna considerare anche l'adeguatezza delle norme in vigore, per esempio nel caso in cui la procedura di licenziamento fosse particolarmente lunga e onerosa, questo disincentiverebbe il datore di lavoro ad assumere personale, in quanto se in futuro fossero necessari tagli del personale, questi genererebbero elevati costi a carico dell'impresa.
4. **La struttura industriale**. Con questo termine si intende la configurazione del tessuto economico di un determinato Stato. Nel caso in cui la struttura industriale sia composta da poche imprese di grandi dimensioni risulterà più facile per le autorità incaricate controllare che le aziende rispettino le leggi. Di conseguenza la possibilità di operare nel sommerso diminuisce. Il caso contrario, invece, cioè se il tessuto fosse composto da una vasta rete con un numero elevato di piccole-medie imprese, renderebbe più facile la "mimetizzazione" per un'azienda che decide di operare al di fuori della legge, poiché corre un rischio minore di essere sottoposta a ispezioni da parte delle autorità, favorendo così il sommerso.
5. Secondo Zizza (2002), nel caso italiano, una causa è rappresentata anche **dall'accettazione culturale**. Infatti, in Italia, come in altre nazioni, non esiste da parte della popolazione contrarietà verso chi opera in condizioni di irregolarità, anzi in alcuni casi si assiste a comportamenti di comprensione e giustificazione di tali azioni.
6. **Crescita della domanda di servizi**. La costante crescita della domanda per i servizi personalizzati come le cure a domicilio, i mercati dello svago, dell'intrattenimento, della ristorazione e del turismo,

i quali sono caratterizzati da una forte necessità di manodopera, favoriscono il ricorso al lavoro in nero.

7. **La crescente "volatilità" dell'economia** influisce sulla crescita dell'Economia Sommera. L'introduzione delle nuove tecnologie ha creato infinite opportunità di lavoro, grazie alle quali in pochi metri quadrati si è in grado di creare una vera e propria azienda. Inoltre, le nuove vie di comunicazione hanno notevolmente accorciato le distanze, rendendo determinate attività più occultabili alle autorità.

1.4.2 Cause dell'Economia Sommersa: il punto di vista di Friedrich Schneider e Dominik H. Enste

Esaminiamo adesso le cause dell'Economia Sommersa dal punto di vista di Schneider ed Enste:

1. **L'onere fiscale** riguarda l'ammontare delle imposte da pagare, dirette, indirette, compresi i contributi pensionistici e i costi della salute. All'interno della letteratura, l'importanza di questo aspetto è condivisa da diversi studiosi. Nello specifico, Schneider ed Enste hanno effettuato una analisi di regressione all'interno dei Paesi OCSE per il periodo 1995-2000 e 2001-2005. Da tale studio è emerso che all'aumentare dell'imposizione fiscale, anche il sommerso aumenta. Inoltre, riguardo alla relazione tra Economia Sommersa e imposizione fiscale, i due autori hanno constatato l'esistenza di un fenomeno definito "*circolo vizioso: maggiore è l'imposizione fiscale, maggiore sarà l'economia sommersa, tuttavia maggiore è l'Economia Sommersa, minore saranno le entrate fiscali*". Altri autori condividono quanto qui definito nei propri lavori: Thomas (1992); Johnson, Kaufmann, e Zoido-Lobatón (1998 a, b); Giles (1999 a); Tanzi (1999); Schneider (2003, 2005); Dell'Anno (2007); Dell'Anno, Gomez-Antonio e Alanon Pardo (2007); Buehn and Schneider (2012).
2. **Il livello di regolamentazione**. La regolamentazione imposta dagli Stati ha effetti sensibili sull'economia attuale e sulla crescita futura. Ogni nuova norma applicata ha un impatto diretto o indiretto sull'occupazione, sugli investimenti e sulla produttività. Per le nazioni è importante avere delle istituzioni forti e affidabili in quanto ciò significa avere fiducia nel sistema economico nazionale, all'interno del quale le imprese operano nell'ambito dell'economia ufficiale. Il pericolo che deriva da un elevato livello di regolamentazione è che tali norme influenzino le decisioni dei consumatori e le possibilità di agire dell'impresa. Gli autori sostengono che queste disposizioni abbiano effetti distorsivi sui meccanismi naturali del mercato, sull'accumulazione di capitale, sulla concorrenza e sull'innovazione. Pertanto, nonostante sia necessario un determinato livello di regolamentazione, per poter operare in armonia e proteggere la proprietà privata, un numero eccessivo di norme può generare costi aggiuntivi e creare barriere per l'economia ufficiale. Anche in questo caso, l'analisi empirica svolta dai due autori per il periodo 1995-2000 e 2001-2005 per i Paesi OCSE, ha messo in evidenza come all'aumentare della presenza di una forte regolamentazione corrisponda un aumento del sommerso. Altri lavori che condividono questa linea di pensiero sono: Schneider e Buhen, (2016); Medina e Schneider, (2017); Johnson, Kaufmann, e Shleifer, (1997); Johnson, Kaufmann e Zoido-Lobatón, (1998 b); Friedman, Johnson, Kaufmann e Zoido-Lobato, (2000); Kucera e Roncolato, (2008), Schneider, (2011).
3. **La qualità delle istituzioni** è intesa come la capacità di garantire la proprietà privata, lo sviluppo del benessere della popolazione attraverso le infrastrutture, la fornitura di beni che rispecchiano le preferenze della popolazione. Questo aspetto è legato ai due punti precedenti. Infatti, lo Stato per svolgere questi compiti avrà bisogno di mezzi finanziari che ricava tramite le imposte, e un'alta qualità delle istituzioni permette di avere maggiore controllo sulla propria economia nazionale, riducendo l'incentivo a infrangere le norme e operare nel sommerso. Lo studio empirico svolto sulle nazioni OCSE su due periodi (1995-2000 e 2001-2005) dai due autori, mostra come la presenza di

un'alta qualità di istituzioni corrisponda a una presenza inferiore di Economia Sommersa. Altri autori che condividono il pensiero di Schneider ed Enste, in merito a questa variabile determinante dell'Economia Sommersa sono: Schneider e Buhen, (2016); Medina e Schneider (2017); Johnson e al., (1998 a, b); Friedman, Johnson, Kaufmann e Zoido-Lobatón, (2000); Dreher e Schneider, (2009); Dreher, Kotsogiannis e McCorriston, (2009); Schneider, (2010); Buehn e Schneider, (2012); Teobaldelli, (2011); Teobaldelli and Schneider, (2012); Amendola e Dell'Anno, (2010); Losby e al. (2002); Schneider e Williams (2013).

4. **Moralità fiscale (Tax morality)**: può essere definita come un contratto psicologico tra i cittadini che versano le imposte e lo Stato, rappresentato dalle autorità fiscali. L'efficienza del settore pubblico ha un'influenza dominante sulla moralità fiscale, in quanto se il numero e la qualità dei servizi pubblici sono bilanciati al carico fiscale, i contribuenti sono disposti a pagare le loro imposte onestamente. Un altro aspetto importante, che influenza questa causa, è il rapporto che vige tra le autorità fiscali e i contribuenti. Se essi si sentono sottomessi alle autorità, come se esistesse un rapporto gerarchico, c'è il rischio che non adempiano al loro compito di cittadini, cioè pagare le tasse. Invece, se le due figure si trovassero sullo stesso piano, come se fossero due partner, i contribuenti tenderebbero a rispettare maggiormente l'obbligo derivante dal contratto psicologico. Ci sono ancora poche prove riguardo l'influenza che la morale fiscale esercita sull'Economia Sommersa. Si è iniziato solo recentemente ad approfondire tale campo, ed è stata comunque notata una relazione negativa tra moralità fiscale ed Economia Sommersa (cioè una diminuzione di moralità fiscale porta a un aumento del sommerso). Altri autori che condividono il pensiero di Schneider ed Enste in merito a questa variabile determinante dell'Economia Sommersa sono: Schneider e Buhen, (2016); Medina e Schneider, (2017); Feld e Frey (2007); Kirchler, (2007); Torgler e Schneider, (2009); Feld e Larsen, (2005, 2009); Feld e Schneider, (2010).

Altre determinanti dell'Economia Sommersa sono (Schneider e Buhen, 2016; Medina e Schneider, 2017):

I servizi del settore pubblico: un aumento dell'Economia Sommersa potrebbe portare a minori entrate fiscali, e una conseguente riduzione della qualità e della quantità di beni e servizi forniti pubblicamente. Pertanto, potrebbe essere necessario aumentare le aliquote fiscali per le imprese e le persone, nonostante persista un deterioramento nella qualità dei beni pubblici (come le infrastrutture pubbliche) e dell'amministrazione. La conseguenza è un incentivo ancora più forte a partecipare all'Economia Sommersa (Johnson, Kaufmann e Zoido-Lobatón, 1998).

Deterrenza: nonostante la forte attenzione alla deterrenza nelle politiche che combattono l'Economia Sommersa e le intuizioni inequivocabili della teoria economica tradizionale della non conformità fiscale, poco è noto da studi empirici sugli effetti della deterrenza. Questo perché i dati non sono disponibili su base internazionale; anche per i Paesi dell'OCSE tali dati sono difficili da raccogliere. La poca evidenza empirica disponibile dimostra che multe e punizioni non esercitano un'influenza negativa sull'Economia Sommersa e che le dimensioni dell'Economia Sommersa possono influire sulla deterrenza, ma non viceversa (Andreoni, Erard e Feinstein, 1998).

Sviluppo dell'economia ufficiale: lo sviluppo dell'economia ufficiale è un altro fattore chiave nell'Economia Sommersa. Quanto maggiore (minore) è la quota di disoccupazione (il tasso di crescita del PIL), tanto maggiore (minore) è l'incentivo a lavorare nell'Economia Sommersa (Schneider e Williams, 2013).

Lavoro autonomo: più alto è il tasso di lavoro autonomo, più attività possono essere svolte nell'Economia Sommersa (Schneider e Williams, 2013).

Disoccupazione: maggiore è il tasso di disoccupazione, più elevata è la probabilità di lavorare nel sommerso (Schneider e Williams, 2013).

Dimensione del settore agricolo: più grande è il settore agricolo, maggiori sono le possibilità di lavorare nell'Economia Sommersa (Hassan e Schneider, 2016).

L'uso dei contanti: più grande è l'Economia Sommersa, più denaro sarà usato (Hassan e Schneider, 2016; Williams e Schneider, 2016).

Quota di forza lavoro: più alta è l'Economia Sommersa, minore è il tasso di partecipazione ufficiale alla forza lavoro (Schneider e Williams, 2013).

1.4.3 Effetti dell'Economia Sommersa

Schneider ed Enste (Schneider ed Enste, 2013; Radovanovic, 2017) analizzano anche gli effetti dell'Economia Sommersa.

Il primo aspetto analizzato è quello che riguarda l'**utilizzo delle risorse** e l'**influenza sull'economia ufficiale**.

Spreco di risorse: uno degli aspetti principali dell'Economia Sommersa è lo **spreco di risorse** che essa porta con sé. Si pensi ai costi aggiuntivi che lo Stato si trova a sostenere in termini di verifica del rispetto delle norme e, qualora non fossero rispettate, l'onere delle procedure di sanzione.

Economia Ufficiale: tra l'Economia Sommersa e **l'Economia Ufficiale** si crea una concorrenza sleale. Coloro che operano nel settore regolare devono sostenere costi quali l'onere fiscale, rispettare determinate norme a sostegno della forza lavoro, dei salari minimi e degli orari di lavoro imposti. Invece le aziende che operano nel sommerso non devono sottostare a questo tipo di normative, rendendo così il costo del capitale umano decisamente inferiore. Il fatto di decidere se entrare nel sommerso o no è a discrezione delle singole aziende; tuttavia, nel caso in cui la possibilità di essere scoperti sia bassa, le imprese saranno più stimolate a operare fuori dai confini legali. I consumatori, basando le loro decisioni sul prezzo, sceglieranno i beni e i servizi offerti dalle aziende che fanno uso di lavoratori illeciti in quanto saranno in grado di offrire un prezzo più concorrenziale rallentando l'Economia Ufficiale. Spesso le persone che lavorano in nero hanno lacune nell'istruzione o nella formazione, di conseguenza quando le autorità competenti fanno emergere queste irregolarità, tali lavoratori faticano a ricollocarsi, entrando in una disoccupazione di medio-lungo periodo, ostacolando la crescita dell'Economia Ufficiale. Se si prende in considerazione il pensiero di Loayza (Loayza, 1997), che ha studiato il sommerso per gli Stati dell'America Latina, si può concludere che l'Economia Sommersa rallenta la crescita dell'economia ufficiale, infatti, nello studio è stata trovata una correlazione negativa tra dimensione del sommerso e crescita dell'Economia Ufficiale.

Spinta all'innovazione dell'Economia Sommersa: oltre agli aspetti negativi analizzati in precedenza, il fenomeno del sommerso porta con sé un forte potenziale innovativo. Infatti, nel sommerso si ha maggiore possibilità di trovare degli individui che siano disposti a prendersi dei rischi al fine di trovare dei possibili impieghi di mercato per i loro prodotti innovativi. La spinta innovativa del sommerso deriva dal fatto che questo può essere svolto raggirando i regolamenti e l'imposizione fiscale, diminuendo così i costi a carico di questi "imprenditori".

Effetti sugli indicatori macroeconomici: gli studiosi estrapolano dagli indicatori macroeconomici informazioni come la variazione dei prezzi, la modifica del tasso di disoccupazione oppure il tasso di crescita del PIL, e poi forniscono tali informazioni alle autorità competenti o alle persone a capo degli enti istituzionali interessati che hanno il compito di prendere determinate decisioni. Queste decisioni dovrebbero essere prese grazie all'ausilio di un quadro della situazione preciso e affidabile, ma l'Economia Sommersa altera gli indicatori. La distorsione dei dati ufficiali può portare a valutazioni errate del reale quadro economico e di conseguenza a scelte sbagliate. I possibili impatti sugli indicatori dell'Economia Sommersa sono:

A. Dimensione errata del PIL;
B. Tasso di crescita dell'economia reale inesatto, dovuto al fatto che l'Economia Sommersa cresce a una velocità diversa rispetto a quella ufficiale;
C. I prezzi nell'Economia Sommersa, essendo esenti da imposte, tendono a crescere più lentamente, di conseguenza il tasso di inflazione ufficiale (o rilevato) risulta più alto di quello reale;
D. Il tasso di disoccupazione è influenzato dai lavoratori in nero che sono ancora registrati come in cerca di impiego. In questo modo alle autorità risulta che la forza lavoro senza un'occupazione sia maggiore rispetto a quella reale.

Effetti fiscali: il sommerso ha implicazioni sostanziose a livello fiscale, in quanto gli evasori sono considerati free rider, cioè individui che utilizzano i beni e i servizi pubblici messi a disposizioni dallo Stato alla popolazione senza però contribuire al finanziamento. Questo fattore, oltre ad avere caratteristiche di ingiustizia sociale, ha effetti negativi anche sulle casse dello Stato, che si trova a dovere offrire un servizio a un bacino di individui maggiore rispetto a quello che ha realmente contribuito all'investimento. Per tale ragione lo Stato sarà obbligato ad aumentare le aliquote fiscali per recuperare i mezzi necessari al mantenimento di tali beni e servizi.

Effetti sul Sistema Sociale: il concetto legato ai sistemi sociali statali è simile a quello dell'onere fiscale, infatti, coloro che operano nel sommerso non contribuiscono al finanziamento delle pensioni e di tutti gli altri contributi che si basano sul valore della solidarietà. Inoltre, esiste la possibilità che questi individui beneficino di tali prestazioni senza però aver contribuito. Il sommerso è un'alternativa attraente nel caso in cui un individuo riceva delle prestazioni come la disoccupazione o l'invalidità. Queste indennità sono legate all'assenza di un'attività lavorativa, l'incentivo deriva dal fatto che una persona può decidere di ricevere contemporaneamente un indennizzo di disoccupazione e lavorare in nero, abusando così delle prestazioni e arricchendosi in maniera illecita.

BIBLIOGRAFIA CAPITOLO 1

Allingham, G. e Sandmo, A. (1972). Income tax evasion: a theoretical analysis.
Andreoni, J.Erard, B. e Feinstein, J. (1998). Tax compliance. Journal of Economic Literature, Volume 36, Issue 2, 818-860.
Campanelli, L. I metodi di analisi statistica per la ricerca sull'economia sommersa (Comitato per l'emersione del lavoro non regolare - Presidenza del Consiglio dei Ministri).
Dell'Anno, R. (2003). Stimare l'economia sommersa con un approccio ad equazioni strutturali. Un'applicazione all'economia italiana (1962-2000). SIEP.
Dell'Arno, R. e Schneider, F. (2003). The Shadow Economy of Italy and other OECD Countries: What do we know? Journal of Public Finance and Public Choice.
Documento redatto dal Ministero del Lavoro (2002). Analisi delle metodologie adottate per la rilevazione del sommerso.
Feige, E. (1979). How big is the irregular economy? Research Gate
Feige, E. (1986). Sweden underground economy.
Feige, E. (1989). The Underground Economy. Cambridge University Press, Cambridge.
Feinster, J. (1999). Approaches for estimating noncompliance: example from Federal taxation in the United States. The Economic Journal 109.
Frey, B. - Weck-Hanmeman, H. (1984), "The hidden economy as an unobserved variable". European Economic Review n. 26/1.
Gruppo di lavoro MEF (gennaio/giugno 2011). Economia non osservata e flussi finanziari.
Guardia di Finanza. (2008). Economia Sommersa: Profili di analisi comparata tra i principali paesi dell'Unione Europea. A cura dei Frequentatori del 35° Corso Superiore di Polizia Tributaria.
Hassan, M. and Schneider, F. (2016). Size and Development of the Shadow Economies of 157 Countries Worldwide: Updated and New Measures from 1999 to 2013.
ISTAT (2010). La misura dell'economia sommersa secondo le statistiche ufficiali. Anni 2000-2008.
ISTAT (2015). L'economia non osservata nei conti nazionali. Anni 2011-2013
ISTAT (2017). L'economia non osservata nei conti nazionali. Anni 2012-2015
ISTAT (2018). L'economia non osservata nei conti nazionali. Anni 2013-2016
Johnson, S., Kaufmann, D., e Zoido-Lobaton, P. (1998). Regulatory Discretion and the Unofficial Economy. American Economic Review, 88 (2).
Kaufmann, D. Kaliberta, A. (1996). Integrating the Unofficial Economy into the Dynamics of Post-Socialist Economies. Policy Research Working Paper 1691.
Lackò, M. (1999). Hidden Economy - An Unknown Quantity? Comparative Analysis of Hidden Economies in Transition Countries in 1989-1995.Research Project of the Jubiläumsfonds Nr. 6882
Loayza, A. (1997). The economics of the informal sectors. A simple model and some empirical evidence from Latin America. Policy Research Working Paper 1727. The World Bank Policy Research Department Macroeconomics and Growth Division.
Lucifora, C. (2003). Economia sommersa e lavoro nero. Il Mulino.
Marino, A. (2015) Indicatori, sostenibilità, crescita: l'economia non direttamente osservabile. La Sinistra Rivista.
Medina, L. e Schneider, F. (2017). Shadow Economies Around the World: What Did We Learn Over the Last 20 Years? WP/18/17 – IMF Working Paper
Morlacchi, A. (2014). Manuale di scienza delle finanze, XXI edizione.
Radovanovic, I. (2017). Economia Sommersa. Un'analisi delle economie dell'Unione Europea e della Confederazione Elvetica alla scoperta di ciò che non vediamo. Tesi di Bachelor in Economia Aziendale Scuola Universitaria Professionale della Svizzera Italiana Dipartimento economia aziendale, sanità e sociale.
Rassegna Economica (2013). Ampiezza e dinamica dell'economia sommersa e illegale. Rivista internazionale di economia e territorio.
Regolamento (CE) N. 2223/96 del Consiglio Europeo del 25 giugno 1996 relativo al Sistema europeo dei conti nazionali e regionali nella Comunità

Schneider, F. (2004). The Size of the Shadow Economies of 145 Countries all over the World : First Results over the Period 1999 to 2003. Papers, No. 1431, Institute for the Study of Labor (IZA), Bonn.
Schneider, F. (2005). Shadow Economies around the World: What do we really know? IAW Diskussionspapiere, No. 16, Institut für Angewandte Wirtschaftsforschung (IAW), Tübingen
Schneider, F. (2013). Size and Development of the Shadow Economy of 31 European and 5 other OECD Countries from 2003 to 2013: A Further Decline.
Schneider, F. (2013). Size and Development of the Shadow Economy of 31 European and 5 other OECD Countries from 2003 to 2012: Some New Facts. Research Gate
Schneider, F. (2013). The Shadow Economy in Europe. ATKearny.
Schneider, F. e Williams, C. (2013). The Shadow Economy. The Institute of Economic Affairs.
Schneider, F. e Buehn, C. (2016). Estimating the Size of the Shadow Economy: Methods, Problems and Open Questions. Discussion Paper No. 9820.The Institute for the Study of Labor (IZA) in Bonn.
Smith, P. (1997). The Underground Economy: Global Evidence of its Size and Impact. The Fraser Institute Vancouver, British Columbia, Canada
Tanzi, V. (2013). L'economia sotterranea degli Stati Uniti: stime e implicazioni. Moneta e Credito
Viel L. (a.a. 2015/2016). L'economia sommersa. Il caso italiano. Tesi di dottorato. Corso di laurea in economia e management. Università degli studi di Padova
Zizza, R. (2002). Metodologie di stima dell'economia sommersa: un'applicazione al caso italiano. Banca d'Italia. Temi di discussione del Servizio Studi.

2 – I modelli con dati panel

2.1 I dati panel

Il termine dati panel o dati longitudinali (Stock e Watson, 2009), si riferisce a dati relativi a N entità diverse osservate in T periodi temporali diversi. Pertanto, essi combinano le informazioni delle caratteristiche di N entità, nello stesso istante temporale, con quelle rilevate per le stesse entità in T diversi periodi di tempo. Quindi possono essere visti sia come:

- dati cross section o trasversali: per un dato istante sono osservate le caratteristiche di più individui;
- dati time series o serie temporali: per un dato collettivo di individui sono rilevate le diverse caratteristiche in diversi istanti.

Nel descrivere i dati sezionali si utilizza un pedice per indicare l'entità; per esempio Y_i si riferisce alla variabile Y per la i-esima entità. Per descrivere una serie temporale si utilizza un pedice per indicare il tempo: per esempio Y_t si riferisce alla variabile Y per la t-esima osservazione temporale. Pertanto, nel descrivere i dati panel, sarà necessario utilizzare entrambe le notazioni per tenere conto sia dell'entità sia del tempo. Per fare questo si usano due pedici invece di uno: il primo i, si riferisce all'entità e il secondo, t, si riferisce al tempo dell'osservazione. Perciò Y_{it} indica la variabile Y osservata per la i-esima delle N unità nel t-esimo dei T tempi:

$$Y_{it},\ i=1,\ldots.,N \ \text{ e } \ t=1,\ldots\ldots,T$$

Se i dati contengono osservazioni sulle variabili X e Y, allora essi si indicano con:

$$(X_{it},Y_{it}),\ i=1,\ldots.,N \ \text{ e } \ t=1,\ldots\ldots,T$$

Un dataset panel può essere, ad esempio, così rappresentato:

Unità	**Tempo**	$\mathbf{Y_{it}}$	$\mathbf{X_{1it}}$	$\mathbf{X_{2it}}$	$\mathbf{X_{3it}}$
1	2000	$Y_{1,00}$	$X_{1,1,00}$	$X_{2,1,00}$	$X_{3,1,00}$
1	2001	$Y_{1,01}$	$X_{1,1,01}$	$X_{2,1,01}$	$X_{3,1,01}$
1	2002	$Y_{1,02}$	$X_{1,1,\ 02}$	$X_{2,1,\ 02}$	$X_{3,1,\ 02}$
2	2000	$Y_{2,00}$	$X_{1,2,00}$	$X_{2,2,00}$	$X_{3,2,00}$
2	2001	$Y_{2,01}$	$X_{1,2,01}$	$X_{2,2,01}$	$X_{3,2,01}$
2	2002	$Y_{2,02}$	$X_{1,2,02}$	$X_{2,2,02}$	$X_{3,2,02}$
…..	…..	…..	…..	…..	…..
N	2000	$Y_{N,00}$	$X_{1,N,00}$	$X_{2,N,00}$	$X_{3,N,00}$
N	2001	$Y_{N,01}$	$X_{1,N,01}$	$X_{2,N,01}$	$X_{3,N,01}$
N	2002	$Y_{N,02}$	$X_{1,N,02}$	$X_{2,N,02}$	$X_{3,N,02}$

dove sono state ipotizzate N unità di osservazione, un periodo di osservazione di 3 anni (2000, 2001, 2002) e K=3 variabili esplicative.

Un dataset panel può essere:

- *bilanciato*, ossia contenere tutte le sue osservazioni, cioè, le variabili sono osservate per ciascuna entità e ciascun periodo temporale;
- *non bilanciato*, ossia vi sono dei dati mancanti per almeno un periodo e per almeno una entità.

Un esempio ben noto di dataset panel proviene dagli Stati Uniti (Baltagi, 2005) e riguarda il National Longitudinal Surveys (NLS), che consiste in un insieme di sondaggi sponsorizzati dall'Ufficio di presidenza di Statistica del lavoro[12].

Il dataset NLS è formato da una serie di sondaggi progettati che raccolgono informazioni sulle attività del mercato del lavoro e altri eventi significativi della vita di diversi gruppi di uomini e donne, quali istruzione, formazione, salute, vita familiare:

- NLSY97 è una indagine sui giovani uomini e sulle giovani donne nati negli anni 1980-84. I rispondenti avevano un'età compresa tra i 12 e i 17 anni quando sono stati intervistati per la prima volta nel 1997. La coorte NLSY97 comprende due campioni: un campione trasversale e un ulteriore campione di rispondenti neri/ispanici/latini. La coorte NLSY97 è stata selezionata in due fasi. Nella prima fase è stato estratto un elenco di unità abitative per i due campioni attraverso un campionamento multistadio con stratificazione. Ciò ha garantito una rappresentazione accurata della popolazione definita secondo la caratteristica di razza, reddito, regione. Nella seconda fase sono state identificate in ciascuna famiglia tutte le persone idonee all'indagine NLSY97 nate tra il 1980 e il 1984.
- NLSY79 è una indagine su uomini e donne intervistati nel 1979. Comprende tre sottocampioni: i) un campione trasversale di 6.111 intervistati, progettato per rappresentare persone che vivevano negli Stati Uniti nel 1979, nati tra il 1 ° gennaio 1957 e il 31 dicembre 1964 (età 14-21 al 31 dicembre 1978); ii) un campione supplementare di 5.295 persone, economicamente svantaggiate che vivevano negli Stati Uniti nel 1979, nate tra il 1° gennaio 1957 e il 31 dicembre 1964; iii) un campione di 1.280 intervistati, progettato per rappresentare la popolazione appartenente alle forze armate statunitensi al 30 settembre 1978, nati tra il 1° gennaio 1957 e il 31 dicembre 1961 (età 17-21 anni al 31 dicembre 1978).
- NLSY79 di ragazzi e bambini è un dataset panel che include due sottocampioni: il campione NLSY79 bambini e il campione NLSY79 ragazzi. Il primo campione comprende i bambini nati da madri appartenenti all'indagine NLSY79. A partire dal 1986, i bambini delle madri NLSY79 sono stati seguiti ogni due anni per valutare il loro sviluppo cognitivo, fisico e socio-emotivo. A partire dal 1994, i bambini che hanno raggiunto l'età di 15 anni entro la fine dell'anno di indagine, non sono più stati valutati in quanto sono entrati a far parte del campione di ragazzi NLSY79.
- NLSW di donne adulte e giovani. L'indagine sulle donne giovani include donne che sono state intervistate per la prima volta nel 1968 quando avevano un'età compresa tra i 14 ei 24 anni. Il sondaggio delle donne più mature comprende donne che sono state intervistate per la prima volta nel 1967 quando avevano un'età compresa tra i 30 ei 44 anni.
- NLS di uomini adulti e giovani è una indagine attualmente interrotta. Era composta da due campioni: NLS degli uomini giovani, che è stato interrotto nel 1981, e includeva uomini di età compresa tra i 14 e i 24 anni nel 1966; NLS degli uomini più anziani, che è stato interrotto nel 1990, e includeva uomini di età compresa tra 45 e 59 anni nel 1966.

Altro esempio di dati panel è il Panel Europeo sulle Famiglie (Gallo, Mastrovita e Siciliani, 2004 e ISTAT, 2008).

Il Panel Europeo sulle Famiglie (European Community Household Panel – ECHP) è un'indagine campionaria che, dal 1994, è stata effettuata con cadenza annuale in tutti i Paesi dell'Unione europea fino al 2001. Dal 2004 è stata sostituita dal progetto EU-SILC[13](Statistics on Income and Living Conditions). Il progetto ha come

[12] http://www.bls.gov/nls/home.htm

[13] Il Regolamento del Parlamento europeo, Statistics on Income and Living Conditions, n°1177/2003.

obiettivo principale la produzione sistematica di statistiche comunitarie su reddito, povertà ed esclusione sociale, sia a livello trasversale che longitudinale. Il Regolamento[14] precisa le responsabilità dei Paesi membri e di Eurostat nel progetto, e definisce un insieme di regole comuni al fine di migliorare la qualità, la comparabilità e la tempestività dei dati, oltre a promuovere una migliore integrazione delle nuove statistiche nei sistemi statistici nazionali. L'indagine è realizzata dagli Istituti Nazionali di Statistica o da Istituti di ricerca nazionali con il coordinamento dell'EUROSTAT. Per assicurare la comparabilità dei dati tra i Paesi membri, il Regolamento definisce alcune regole comuni riguardo la popolazione target, la definizione delle variabili, le dimensioni del campione, le regole di inseguimento delle famiglie e dei loro componenti[15], lasciando ai singoli Paesi alcuni margini di flessibilità rispetto all'impiego di differenti fonti di dati (indagine campionaria/archivi), al periodo di riferimento del reddito (fisso/mobile), alla modalità di raccolta delle informazioni sui redditi lordi (indagine/archivi/microsimulazione) e alla struttura dei questionari nazionali.

L'ISTAT conduce l'indagine campionaria con le seguenti modalità:

1) Popolazione di riferimento. La popolazione di riferimento è costituita da tutti i componenti delle famiglie residenti in Italia, anche se temporaneamente all'estero. Sono escluse le famiglie residenti in Italia che vivono abitualmente all'estero e i membri permanenti delle convivenze istituzionali (ospizi, brefotrofi, istituti religiosi, caserme, eccetera).

2) Unità di rilevazione. L'unità di rilevazione è la famiglia di fatto definita famiglia campione. Questa va intesa come un insieme di persone legate da vincoli di matrimonio, parentela, affinità, adozione, tutela o da vincoli affettivi, coabitanti e aventi dimora abituale nello stesso comune (anche se non residenti secondo l'anagrafe nello stesso domicilio).

3) Individuo campione, famiglia campione e famiglia longitudinale. Tutti gli individui appartenenti alle famiglie campione debbono essere intervistati a patto che abbiano compiuto 15 anni[16] nell'anno di riferimento del reddito. Ogni individuo che appartiene alla famiglia campione intervistato nella prima fase diviene individuo campione e va intervistato anche nelle fasi successive a meno che, nel frattempo, sia deceduto o si sia trasferito all'estero. La famiglia longitudinale si riferisce alla famiglia al tempo t=1, cioè alla prima fase dell'indagine. Dalla seconda fase in poi, le famiglie campione sono quelle composte da almeno un individuo campione e tutti i componenti che a qualsiasi titolo si aggiungono alla famiglia campione sono considerati individui coabitanti.

4) Periodicità e riferimento temporale. L'indagine viene svolta annualmente, in un periodo successivo alle dichiarazioni dei redditi in modo da dare la possibilità alle famiglie e agli individui di poter utilizzare le informazioni derivanti dalle proprie dichiarazioni fiscali. Le notizie acquisite fanno riferimento a due periodi distinti: alcune alla data di indagine (anno t) e altre, principalmente quelle sul reddito, all'anno precedente (t-1).

I riferimenti temporali delle notizie raccolte sono:

- periodo dell'intervista (anno t). A questo periodo vanno associate le informazioni familiari e individuali che caratterizzano la condizione di vita attuali (come ad esempio, le caratteristiche

[14] Il Regolamento è stato pubblicato il 3 luglio 2003 (Official Journal n. 165).

[15] Le regole di inseguimento impongono di continuare a seguire gli individui campione nei loro spostamenti residenziali sul territorio nazionale e, data la scala sovra nazionale dell'indagine, anche nell'ambito dei Paesi dell'Unione europea. Per questo è necessario mantenere contatti con le famiglie per avere informazioni su eventuali cambiamenti di residenza di tutta la famiglia o di singoli componenti usciti per formare nuove famiglie.

[16] In realtà Eurostat chiede di intervistare tutti gli individui che nell'anno di riferimento del reddito abbiano compiuto 16 anni. L'Italia ha scelto di intervistare anche i quindicenni per uniformare le definizioni a quelle di altre indagini sulle famiglie.

dell'abitazione, il possesso dei beni durevoli, le condizioni di salute degli individui, l'istruzione, l'attuale condizione lavorativa).

- ultimi dodici mesi. A questi periodi vanno ricondotte, ad esempio, le principali spese per l'abitazione.
- anno t-1. A questo periodo si riferiscono tutte le informazioni familiari e individuali che caratterizzano la situazione economica della famiglia e degli individui (come, ad esempio, i mutui, i prestiti e i redditi).

5) Il disegno di rilevazione. Il disegno di rilevazione tiene conto dell'insieme di requisiti indicati da EUROSTAT:

- l'indagine deve presentare una componente trasversale e una componente longitudinale, pur assumendo maggiore importanza la componente trasversale, in quanto la priorità dell'indagine è fornire stime comparabili, tempestive e di alta qualità a livello trasversale;
- le due componenti devono avere cadenza annuale, il periodo di riferimento dei quesiti riferiti al reddito deve essere l'anno solare precedente il momento dell'intervista, mentre per gli altri quesiti tale periodo di riferimento può essere variabile;
- la componente longitudinale deve rilevare informazioni riferite a un arco temporale di durata di almeno quattro anni;
- devono essere prodotti archivi di dati aggiornati annualmente, con le seguenti caratteristiche: i) gli archivi di dati trasversali, che si riferiscono sia alle famiglie che agli individui che le compongono, devono contenere per ciascuna unità sia le variabili di tipo economico che quelle di tipo sociale; ii) gli archivi di dati longitudinali, riferiti agli individui, devono riportare, per ciascun individuo rilevato, i valori delle variabili di interesse osservati in almeno quattro occasioni di indagine (quella corrente e le tre precedenti) in cui detto individuo è stato intervistato;
- la componente trasversale ha la finalità di fornire stime delle famiglie e degli individui che le compongono con riferimento a parametri di livello con cadenza annuale e a variazioni nette, con principale interesse per le variazioni a un anno di distanza;
- la componente longitudinale ha lo scopo di produrre stime degli individui con riferimento a flussi e durate medie per quanto riguarda la dinamica dei processi legati all'esclusione sociale e alla povertà;
- la numerosità campionaria (in termini di unità finali) di ciascuna delle due componenti deve rispettare certi livelli minimi già stabiliti.

È stato definito un disegno di rilevazione di tipo panel, in base al quale le informazioni relative alle variabili di interesse sono raccolte sulle medesime unità campione in tempi differenti. Lo schema utilizzato è quello di un panel ripetuto costituito da una serie di panel, ciascuno dei quali è di durata fissa e che si sovrappone agli altri. Pertanto, due o più panel coprono parte dello stesso periodo temporale. Questo schema è equivalente al campionamento ruotato, in quanto in entrambi i casi ciascun panel ha durata limitata nel tempo e due o più panel vengono seguiti nello stesso periodo di tempo. All'interno di ogni singolo panel il disegno di campionamento usa uno schema standard a due stadi comuni-famiglie con stratificazione dei comuni che vengono suddivisi in comuni Auto Rappresentativi (Ar), ossia i comuni con maggior dimensione demografica, e Non Auto Rappresentativi (Nar), costituito dai rimanenti comuni. I comuni Nar sono stratificati in base alla dimensione demografica.

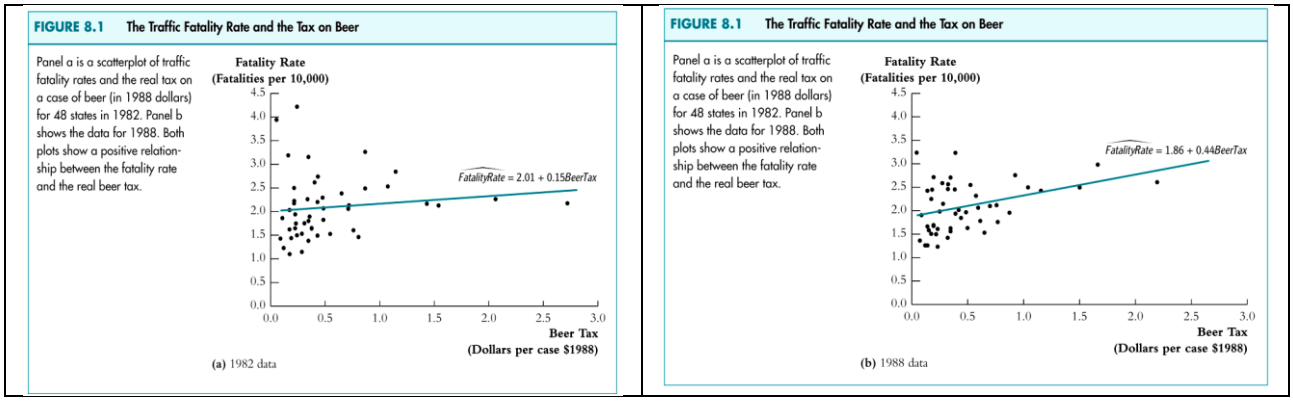
Esistono una serie di vantaggi e di svantaggi che riguardano l'utilizzo di questa tipologia di dati.

Un primo vantaggio è quello di superare il problema delle variabili omesse. Infatti, sotto certe condizioni l'uso dei dati panel permette di ottenere stimatori consistenti in presenza di variabili omesse (Wooldridge, 2010).

Siano y e x = (x_1,…,x_k) le variabili casuali osservabili e *c* una variabile casuale non osservabile. La popolazione di interesse è costituita dal vettore (y,x_1,…,x_k,*c*). Siamo interessati agli effetti che le variabili esplicative hanno sulla variabile y. Se *c* non è correlata con le x_j risulta essere un altro fattore non osservabile che influenza la y. Viceversa se *c* è correlato con le x_j non è possibile inserirlo all'interno dell'errore perché porterebbe a stime inconsistenti. Un modo per superare questo problema è quello di introdurre una struttura dei dati che sia ripetuta nel tempo, ossia i dati panel. Secondo tale struttura, e supponendo che *c* sia costante nel tempo, *c* risulta essere un effetto inosservato che cattura la caratteristica dell'individuo che è data e non cambia nel tempo.

A titolo esemplicativo di seguito si riporta uno studio tratto da Stock e Watson, 2009, che cerca di valutare se gli interventi pubblici volti a scoraggiare la guida in stato di ebbrezza siano efficaci nel ridurre effettivamente i decessi causati da incidenti stradali. Come misura degli incidenti stradali l'autore utilizza il tasso di mortalità espresso dal numero dei morti sulle strade in un anno per 10.000 abitanti. Le imposte sugli alcolici sono invece valutati attraverso l'imposta su una cassa di birra.

Figura1: Tasso di mortalità sulle strade e l'imposta sulla birra



Fonte: Stock, J.H. e Watson, M.W. (2009)

La figura 1 mostra la relazione esistente tra il tasso di mortalità sulle strade e l'imposta reale su una cassa di birra, nel 1982 e nel 1988. In entrambi i casi si nota una relazione positiva tra il tasso di mortalità e la tassa reale sulla birra.

Si dovrebbe quindi concludere che un inasprimento dell'imposta sulla birra porterebbe ad un maggior numero di incidenti stradali mortali, mentre obiettivo di un policy maker è quello di diminuirne il numero. Ciò avviene perché le analisi della regressione utilizzate potrebbero avere una sostanziale distorsione da variabili omesse. Molti fattori, infatti, influenzano il tasso di mortalità, inclusa la qualità delle automobili guidate nello stato, la condizione delle autostrade, il fatto che il traffico si concentri in zone urbane o rurali, la densità delle automobili sulla strada. Ognuno di questi fattori potrebbe essere correlato con le imposte sugli alcolici; se ciò accadesse si avrebbe distorsione da variabili omesse. Un modo per superare questo problema è quello di utilizzare una struttura di dati panel.

Pertanto, i dati panel vengono usati, sotto alcune condizioni, per ottenere stimatori consistenti in presenza di variabili omesse, in quanto sono in grado di controllare l'eterogeneità individuale (Wooldridge, 2010)**.**

Si riportano di seguito ulteriori significativi vantaggi associati ai dati panel (Hsiao, 2003; Klevmarken, 1989; Baltagi, 2005; Benfratello, 2013, 2015):

Permettono di trovare gli effetti che non è possibile investigare con semplici dati trasversali o serie temporali.

Con le serie temporali è possibile semplicemente dedurre quanto varia ogni anno un certo fenomeno che si sta studiando, mentre con i dati trasversali è possibile fotografare il fenomeno solamente in un determinato istante di tempo. Ad esempio:

1) supponiamo di avere una serie temporale sul consumo e vogliamo rilevare se il consumo aggregato aumenta del 2% ogni anno. Da una serie temporale in realtà possiamo solo dedurre che il consumo medio è aumentato del 2%. Con i dati panel, invece, possiamo controllare il comportamento a livello individuale. Possiamo pertanto vedere se metà della popolazione ha aumentato il consumo del 4% e metà della popolazione non ha subito cambiamenti.

2) supponiamo di avere uno studio sezionale riguardante donne con un tasso di partecipazione medio annuo alle forze di lavoro pari al 50%. Ciò potrebbe essere dovuto ai seguenti motivi: i) ogni donna può avere il 50% di possibilità di essere nel mondo del lavoro, in un dato anno; ii) il 50% delle donne lavora tutto il tempo e il 50% non lavora affatto. Solo i dati panel possono discriminare tra questi due casi.

I dati panel sono più informativi, sono uno dei metodi che permettono di rappresentare in modo semplice dati ad elevata variabilità e di ridurre la multicollinearità indotta da fattori non esplicitamente considerati, hanno più gradi di libertà e più efficienza. Gli studi su serie temporali hanno una forte presenza di multicollinearità. Per esempio, nel caso della domanda di sigarette, esiste un'elevata collinearità tra prezzo e reddito nelle serie temporali aggregate per gli Stati Uniti. Questo è meno probabile con una struttura panel nella quale viene inserita la dimensione trasversale, la quale aggiunge molta variabilità, poiché presenta più dati informativi sul prezzo e sul reddito. Con ulteriori dati più informativi si possono produrre stime dei parametri più affidabili.

I dati panel vengono preferiti nello studio delle dinamiche di alcuni fenomeni economici. Distribuzioni trasversali di fenomeni economici quali la disoccupazione, il lavoro, il turnover, la migrazione e il reddito, sembrano spesso stabili, mentre nascondono una moltitudine di cambiamenti. Tali fenomeni riescono a fare emergere le proprie dinamiche con l'uso dei dati panel. Infatti, con riferimento allo studio delle dinamiche del fenomeno della disoccupazione, i dati trasversali permettono di osservare la proporzione di popolazione disoccupata in un determinato momento. I dati panel, invece, fornendo oltre ai dati trasversali anche quelli longitudinali, consentono di osservare come cambia l'andamento della proporzione di popolazione disoccupata al variare del tempo. Pertanto, solo attraverso l'utilizzo di una struttura panel si può osservare quale percentuale dei disoccupati, riferita a un determinato arco temporale, può rimanere tale in un altro periodo.

Le limitazioni dei dati panel includono (Baltagi, 2005):

Problemi di progettazione e raccolta dei dati. Come in tutte le survey, includono problemi di copertura (resoconto incompleto della popolazione di interesse), mancata risposta (a causa della mancanza di collaborazione dell'intervistato o dell'errore dell'intervistatore), richiamo (il rispondente non ricorda correttamente), frequenza dell'intervista.

Distorsioni dovuti a errori di misura. Gli errori di misura sono errori costituiti dalla differenza fra il vero valore della caratteristica da misurare su una data unità statistica e il valore effettivamente osservato nell'indagine. Tali differenze possono essere introdotte dal rispondente stesso (per dimenticanza, imprecisione o dolo)

oppure dallo svolgimento delle fasi di elaborazione successive alla raccolta del dato. Esempi di questo secondo caso sono tutti gli errori introdotti dalle operazioni di registrazione su supporto informatico o di codifica dei quesiti aperti.

La "non risposta panel" (panel non response). Comprende tre differenti tipologie di non risposta specifiche delle indagini panel: i) **la non risposta iniziale**, ovvero la presenza di unità che rifiutano o non sono in grado di cooperare alla prima fase, accettando però di rispondere successivamente; ii) **l'attrition**, o caduta definitiva dal panel, che si manifesta con la presenza di unità che, dopo aver risposto a una o più fasi, lasciano il panel senza più farvi ritorno; iii) **la non risposta saltuaria**, data dalla presenza di unità con un comportamento irregolare di risposta alle varie fasi. In ciascuna fase del panel il campione si assottiglia per molteplici ragioni quali il mancato contatto, la mancata risposta, l'insuccesso nel seguire il campione e così via. Questa diminuzione campionaria è compensata dall'inclusione di famiglie split o da famiglie che, non avendo partecipato al massimo a una fase, ritornano successivamente a far parte del campione.

Il logoramento. Nonostante la mancata risposta si verifichi anche negli studi trasversali, nel caso dei dati panel è più grave perché le fasi successive del panel sono ancora soggette a mancata risposta. Gli intervistati, infatti, potrebbero morire o trasferirsi. Se questo processo è endogeno (cioè sistematicamente correlato con la variabile oggetto di studio), conduce a risultati inferenziali in generale non corretti.

2.2 I modelli con dati panel

Consideriamo il modello lineare (Baltagi, 1984, 2005; Stock e Watson, 2009; Wooldridge, 2009, 2010):

$$y_{it} = x'_{it}\,\beta + \alpha_i + \varepsilon_{it} \qquad t = 1,\ldots,T \;\; e \;\; i = 1,\ldots,N \qquad (1)$$

dove y_{it} è l'osservazione della variabile dipendente per l'unità i al tempo t, $\mathbf{x}_{it}$ è un vettore di variabili indipendenti (1xk), $\boldsymbol{\beta}$ è un vettore di coefficienti ($\beta_1,\ldots,\beta_k$) (kx1), α_i è un fattore latente (non osservato) il cui ruolo è quello di catturare gli effetti non osservabili costanti nel tempo. Infatti, tale componente può tener conto dell'eterogeneità degli individui, ossia di caratteristiche peculiari di ciascun individuo che non si possono osservare direttamente e che permangono nel tempo. Infine, ε_{it} sono i disturbi o gli errori *idiosincratici*, detti tali perché variano sia rispetto alle entità i che al tempo t.

2.2.1 I modelli statici

I modelli di regressione con dati panel statici permettono di studiare il comportamento dell'individuo in un ambiente che muta nel tempo.

Modello Pooled OLS (POLS)

Il più semplice modello di stima per dati panel è il modello pooled OLS, che sfrutta lo stimatore dei minimi quadrati ordinari. Nella maggior parte dei casi è improbabile che esso sia adeguato, ma offre una *guideline* per un confronto con modelli più complessi.

Consideriamo il modello seguente:

$$y_{it} = x'_{it}\,\beta + v_{it} \qquad i = 1,\ldots,N \; e \; t = 1,\ldots,T$$

dove $v_{it} = \alpha_i + \varepsilon_{it}$ è il termine di errore composito tale che $E(x'_{it}, v_{it}) = 0$ $t = 1,2,\ldots T$.

Il modello POLS ignora la correlazione seriale nell'errore composito dovuto alla presenza di α_i. Questo modello è adeguato nel caso in cui si ritiene di aver incluso tutti i possibili regressori rilevanti e gli effetti di eterogeneità individuale osservabile per la determinazione di y_{it}. In tale situazione si pensa che non ci sia nessun effetto inosservato e dunque in sostanza α_i possa essere eliminato.

Se però così non fosse, e quindi si trascurasse qualche effetto α_i correlato con x'_{it}, allora la stima POLS risulterebbe distorta e inconsistente. La distorsione dovuta a questo problema viene chiamata anche *heterogeneity bias*, ma è semplicemente una distorsione causata dal fatto che non è stata inclusa alcuna variabile *time-constant* (Wooldridge, 2009 e 2010).

La stima tramite POLS necessita della validità di cinque assunzioni:

- Linearità: la variabile dipendente è descritta da una funzione lineare di variabili indipendenti e del termine di disturbo.
- Esogeneità: il valore atteso dei disturbi è nullo o gli errori non sono correlati con alcun regressore.
- Omoschedasticità e non autocorrelazione: i disturbi hanno la stessa varianza e non sono correlati gli uni agli altri.
- Le osservazioni delle variabili indipendenti non sono stocastiche ma fisse in campioni ripetuti senza errori di misurazione.
- Rango pieno: non c'è multicollinearità, ovvero non c'è alcuna relazione lineare perfetta tra le variabili indipendenti.

L'alternativa al modello POLS è il modello ad effetti fissi (fixed effect o within group - FE) e il modello ad effetti casuali (random effect - RE).

Nei lavori metodologici inizialmente la discussione sull'utilizzo dei due stimatori FE e RE si è incentrata sul fatto che α_i fosse vista come una variabile casuale o come parametro da stimare. Pertanto, nell'approccio tradizionale ai modelli per dati panel, α_i viene chiamato un "effetto casuale" quando viene trattato come una variabile casuale e un "effetto fisso'' quando viene trattato come un parametro da stimare. Nel moderno linguaggio econometrico la distinzione tra i due metodi si basa sul fatto che venga ammessa o meno la correlazione fra gli effetti individuali α_i e le variabili osservate per le varie unità x_{it}: $E(x_{it}, \alpha_i) = 0$ oppure $E(x_{it}, \alpha_i) \neq 0$.

Modello Fixed Effect (FE)

Il modello ad effetti fissi si concentra sull'eliminazione dell'intercetta α_i, che è costante nel tempo e descrive valori non osservabili. Tali valori potrebbero essere correlati con le variabili esplicative x_{it}, restituendo una stima distorta. L'eliminazione del termine α_i si basa sul procedimento di *data-demeaning* (Wooldridge, 2009 e 2010).

Vediamo di seguito la formalizzazione del modello FE:

1. Linearità $y_{it} = x'_{it}\,\beta + \alpha_i + \varepsilon_{it}$ $t = 1,\ldots,T$ e $i = 1,\ldots,N$
2. Stretta esogeneità $E(\varepsilon_{it} | X_i, \alpha_i) = 0$ $t = 1,\ldots,T$
3. Rango $\text{rank}[E(X'_i M^0 X_i)] = \text{rank}\,[\sum_{t=1}^{N} E(x_{it} - \bar{x}_i)(x_{it} - \bar{x}_i)'] = K$

dove M^0 è la matrice che centra le variabili rispetto alla media

$$M^0 = \begin{bmatrix} 1-1/T & \cdots & -1/T \\ \vdots & \ddots & \vdots \\ -1/T & \cdots & 1-1/T \end{bmatrix}$$

4. Omoschedasticità $E(\varepsilon_i\varepsilon'_i|X_i, \alpha_i) = \sigma_\varepsilon^2 I_T$
5. Normalità asintotica $\varepsilon_i \sim N(0, \sigma_\varepsilon^2 I_T)$

Lo stimatore FE è uno stimatore OLS applicato ai dati centrati rispetto alla media individuale. Infatti, i dati originali $y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it}$ possono essere trasformati applicando una trasformazione entro i gruppi (within group). Ciò viene fatto attraverso i seguenti passaggi:

$$\overline{y}_i = \overline{x}_{it}\beta + \alpha_i + \varepsilon_{it}$$

$$\text{dove } \overline{y}_i = \frac{1}{T}\sum_{t=1}^{T} y_{it} \qquad \overline{x}_I = \frac{1}{T}\sum_{t=1}^{T} x_{it} \qquad \overline{\varepsilon}_i = \frac{1}{T}\sum_{t=1}^{T} \varepsilon_{it}$$

$$y_{it} - \overline{y}_i = (x_{it} - \overline{x}_i)'\beta + (\alpha_i - \alpha_i) + (\varepsilon_{it} - \overline{\varepsilon}_I)$$

$$y_{it} - \overline{y}_i = (x_{it} - \overline{x}_i)'\beta + (\varepsilon_{it} - \overline{\varepsilon}_I)$$

Pertanto, i parametri β possono essere stimati attraverso l'equazione trasformata che elimina gli effetti individuali. Lo stimatore ad effetti fissi o tra gruppi (within group) non è altro che lo stimatore OLS applicato alle variabili trasformate:

$$\hat{\beta}_{FE} = \left(\sum_{i=1}^{N}\left(\sum_{t=1}^{T}((x_{it} - \overline{x}_i)(x_{it} - \overline{x}_i)')\right)^{-1}\left(\sum_{i=1}^{N}\left(\sum_{t=1}^{T}((x_{it} - \overline{x}_i)(y_{it} - \overline{y}_i)\right)\right)\right)$$

Modello Random Effect (RE)

Nell'ambito del modello ad effetti casuali, diversamente dal modello a effetti fissi, gli α_i non sono trattati come parametri fissi, ma come realizzazioni di una variabile aleatoria (da qui la denominazione effetti casuali), non correlati ai regressori. In tal modo, questi effetti si possono trattare nel modello come se fossero parte del termine di errore. Si effettua, quindi, una trasformazione dei dati che produce un dataset con errori non autocorrelati. I dati trasformati soddisfano le assunzioni del teorema di Gauss-Markov, e quindi le stime finali sono efficienti.

Vediamo di seguito la formalizzazione del modello RE

1) Linearità $y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it} = x'_{it}\beta + \nu_{it}$ $t = 1,\ldots,T$ e $i = 1,\ldots,N$
   dove $\nu_{it} = \alpha_i + \varepsilon_{it}$ è il termine di errore composito
2) Stretta esogeneità $E(\varepsilon_{it}| X_i,\alpha_i) = 0$ $t = 1,\ldots,T$
   $E(\alpha_i | X_i) = 0$
3) Rango $\text{rank}[E(X'_i \Omega^{-1} X_i)] = K$
   dove Ω è una matrice TxT

$$\Omega = \begin{bmatrix} \sigma_\alpha^2 + \sigma_\varepsilon^2 & \cdots & \sigma_\alpha^2 \\ \vdots & \ddots & \vdots \\ \sigma_\alpha^2 & \cdots & \sigma_\alpha^2 + \sigma_\varepsilon^2 \end{bmatrix}$$

4) Omoschedasticità $E(\varepsilon_i\varepsilon'_i|X_i, \alpha_i) = \sigma_\varepsilon^2 I_T$
   $E(\alpha^2_i | X_i) = \sigma_\alpha^2$

Il teorema di Gauss-Markov, secondo il quale lo stimatore OLS è lo stimatore lineare corretto più efficiente, dipende dall'assunzione che il termine d'errore sia indipendente e identicamente distribuito (iid). Nel

contesto dei dati panel, tale assunzione implica una serie di ipotesi che sono difficilmente verificate, quindi lo stimatore OLS non sarebbe lo stimatore migliore in tal caso. Il problema della correlazione seriale può essere però risolto usando lo stimatore GLS, tenendo in considerazione la struttura della covarianza del termine d'errore. Pertanto, si applica un approccio GLS (Generalised Least Square):

$$\hat{\beta}_{RE} = (\sum_{i=1}^{N}\sum_{t=1}^{T}((x_{it} - \overline{x}_i)(x_{it} - \overline{x}_i)')+\psi T \sum_{i=1}^{N}((\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})'))^{-1}(\sum_{i=1}^{N}(\sum_{t=1}^{T}((x_{it} - \overline{x}_i)(y_{it} - \overline{y}_i))) + \psi T \sum_{i=1}^{N}((\overline{x}_i - \overline{x})(\overline{y}_i - \overline{y}))$$

dove $\psi = \frac{\sigma_\varepsilon^2}{\sigma_\varepsilon^2+\sigma_\alpha^2 T}$, dove $\overline{y} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} y_{it}$ $\overline{x} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} x_{it}$

Nel caso in cui $\psi = 0$ gli stimatori FE e RE sono equivalenti, mentre nel caso in cui $\psi = 1$ allora $\sigma_\alpha^2 = 0$,ciò significa che la matrice Ω è diagonale e quindi viene applicato lo stimatore OLS ai dati iniziali ritornando al modello POLS.

Più in generale lo stimatore RE può essere visto come una media pesata di due stimatori:

- Il primo è uno stimatore FE;
- Il secondo è uno **stimatore between (BE)**, che è uno stimatore OLS applicato alla media dei dati per individuo: $\overline{y}i = \mu + \overline{x}_i'\beta + (\alpha_i + \overline{\varepsilon}_i)$ i= 1,…., N. Pertanto $\hat{\beta}_{BE} = (\sum_{i=1}^{N}((\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})'))^{-1}(\sum_{i=1}^{N}((\overline{x}_i - \overline{x})(\overline{y}_i - \overline{y}))$

Quindi $\hat{\beta}_{RE} = \Delta\hat{\beta}_{BE} + (1- \Delta)\hat{\beta}_{FE}$, dove Δ è una matrice KxK che pesa ogni stimatore sulla base dell'inverso della sua varianza.

Sia lo stimatore RE che lo stimatore Pooled OLS, che si ottiene quando $\psi = 1$, sono una combinazione degli effetti fissi e between. Mentre lo stimatore RE combina in modo ottimale le informazioni provenienti dalla variazione entro i gruppi e dalla variazione tra i gruppi, lo stimatore Pooled OLS non li combina in modo ottimale in quanto considera le informazioni tutte insieme (Pooled).

Modello First Difference (FD)

Un altro stimatore utile nel contesto dei modelli panel è lo stimatore **First Difference (FD)**. È uno stimatore OLS di un modello trasformato, nel quale la trasformazione in differenze prime viene così applicata:

$$y_{it} = x'_{it}\beta + (\alpha_i + \varepsilon_{it})$$

$$y_{it-1} = x'_{it-1}\beta + (\alpha_i + \varepsilon_{it-1})$$

$$y_{it} - y_{it-1} = (x_{it} - x_{it-1})'\beta + (\alpha_i + \varepsilon_{it}) - (\alpha_i + \varepsilon_{it-1})$$

$$\Delta y_{it} = \Delta x'_{it}\beta + \Delta\varepsilon_{it}$$

<u>Ricapitolando:</u> esistono cinque differenti stimatori, ognuno dei quali è una applicazione OLS ad una differente trasformazione dei dati originali. Ricordando la proprietà di consistenza dello stimatore OLS è possibile definire le condizioni di consistenza per gli stimatori introdotti, e precisare gli aspetti computazionali.

1) <u>Lo stimatore BE</u> è uno stimatore OLS che si applica alla media dei dati individuali

$$\overline{y}i = \mu + \overline{x}_i'\beta + (\alpha_i + \overline{\varepsilon}_i)\ i= 1,\ldots., N$$

La condizione di consistenza è che le $\overline{x}_i$ devono essere incorrelate con $(\alpha_i + \overline{\varepsilon}_i)$. Questo implica che x_{is} deve essere incorrelato con α_i e con ε_{it} per ogni s, t=1,…, T, dove con s vengono indicati tutti i periodi diversi dal

periodo t (s≠t). Dal punto di vista computazione, tale stimatore fornisce solo la variabilità between (*between variability*). Infatti la variabilità tra stesso individuo viene eliminata, mentre viene mantenuta la variabilità tra diversi individui. L'informazione dello stesso individuo nel tempo si perde, pertanto lo stimatore $\hat{\beta}_{BE}$ sarà pari a Kx1, dove K è il numero dei parametri.

2) Lo stimatore FE è uno stimatore OLS applicato alla deviazione dalla media del gruppo

$$y_{it} - \overline{y}_i = (x_{it} - \overline{x}_i)' \beta + (\varepsilon_{it} - \overline{\varepsilon}_I)$$

Tale stimatore utilizza solo la variabilità within (*within variability*). La condizione di consistenza è che $x_{it} - \overline{x}_I$ deve essere incorrelato con $(\varepsilon_{it} - \overline{\varepsilon}_I)$. Ciò implica che x_{is} deve essere incorrelato con ε_{it} per ogni s, t=1,…, T. Non è necessaria alcuna correlazione tra α_i e x_{is}. Questa trasformazione ha il vantaggio di ridurre il carico computazionale rispetto allo stimatore OLS. Infatti con dati trasformati within la dimensione dello stimatore è K×1.

3) Lo stimatore RE è uno stimatore OLS applicato alla trasformazione dei dati attraverso una trasformazione GLS

$$y_{it} - \Theta\overline{y}_i = (1 - \Theta)\mu + (x_{it} - \Theta\overline{x}_I)'\beta + (v_{it} - \Theta\overline{v}_i) \text{ dove } \Theta = 1 - \sqrt{\psi}$$

Questo stimatore fornisce sia la variabilità within che la variabilità between. La condizione di consistenza è che $(x_{it} - \Theta\overline{x}_I)$ deve essere incorrelato con $(v_{it} - \Theta\,\overline{v}_i)$. Come nel caso dello stimatore BE, ciò implica che x_{is} sia incorrelato con α_i e con ε_{it} per ogni s, t=1,…, T. Dal punto di vista computazionale, gli stimatori POLS e FE sono definiti modi "all-or-nothing" di utilizzare l'informazione fra individui: POLS tratta indistintamente tutte ("all") le fonti di variabilità, mentre FE non dà nessun peso ("nothing") alla variabilità between. Lo stimatore RE, invece rappresenta il caso intermedio fra i due approcci estremi nel considerare gli effetti individuali. Ricordando che $\hat{\beta}_{RE} = \Delta\hat{\beta}_{BE} + (1- \Delta)\,\hat{\beta}_{FE}$, dove Δ è una matrice KxK, che pesa ogni stimatore sulla base dell'inverso della sua varianza, lo stimatore $\hat{\beta}_{RE}$, considerando che $\hat{\beta}_{BE}$ e $\hat{\beta}_{FE}$ hanno un carico computazionale pari a kx1, sarà pari a kx1.

4) Lo stimatore POLS è uno stimatore OLS che viene applicato ai dati originali

$$y_{it} = x'_{it}\,\beta + \alpha_i + \varepsilon_{it}$$

La condizione di consistenza è che x_{it} deve essere incorrelato con $(\alpha_i + \varepsilon_{it})$. Ciò implica che x_{is} sia incorrelato con α_i e con ε_{it} per s=t. La condizione di consistenza per questo stimatore è più debole rispetto allo stimatore RE e BE. Come già anticipato, dal punto di vista computazionale lo stimatore POLS non tiene conto del fatto che, T osservazioni temporali per N diversi individui, non siano la stessa cosa di NT diversi individui. Ciò perché, dando lo stesso peso alle due fonti di variabilità, within e between, ignora la struttura panel dei dati. Pertanto, tale stimatore fornisce sia la variabilità within che la variabilità between, ma non in modo ottimale come fa lo stimatore RE.

5) Lo stimatore FD è uno stimatore OLS applicato alla trasformazione in differenze prime

$$\Delta y_{it} = \Delta\, x'_{it}\,\beta + \Delta\varepsilon_{it}$$

Tale stimatore richiede che $\Delta\, x_{it}$ sia incorrelato solo con $\Delta\varepsilon_{it}$, pertanto x_{it} deve essere incorrelato con ε_{it-1}, ε_{it}, ε_{it+1}. Dal punto di vista computazionale, il calcolo della differenza prima fa perdere N osservazioni delle NT

complessive; in altri termini, si perde l'osservazione temporale iniziale di ciascuna cross-section, che rimane con T-1, invece che T osservazioni disponibili.

Tabella 1: Riepilogo modelli panel

Metodo	Ipotesi	Modello	Stimatore
BE	Esogeneità	$\overline{y}i = \mu + \overline{x}_i'\beta + (\alpha_i + \overline{\varepsilon}_I)$	$\hat{\beta}_{BE} = (\sum_{i=1}^{N}((\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})'))^{-1}(\sum_{i=1}^{N}((\overline{x}_i - \overline{x})(\overline{y}_i - \overline{y}))$
FE	Linearità, Stretta esogeneità, Rango pieno, Omoschedasticità, Normalità asintotica	$y_{it} - \overline{y}_i = (x_{it} - \overline{x}_i)'\beta + (\varepsilon_{it} - \overline{\varepsilon}_I)$	$\hat{\beta}_{FE} = (\sum_{i=1}^{N}(\sum_{t=1}^{T}((x_{it} - \overline{x}_i)(x_{it} - \overline{x}_i)'))^{-1}$ $(\sum_{i=1}^{N}(\sum_{t=1}^{T}((x_{it} - \overline{x}_i)(y_{it} - \overline{y}_i)))$
RE	Linearità, Stretta esogeneità, Rango pieno, Omoschedasticità	$y_{it} - \Theta\overline{y}_i = (1 - \Theta)\mu + (x_{it} - \Theta\overline{x}_i)'\beta + (v_{it} - \Theta\,\overline{v}_i)$	$\hat{\beta}_{RE} = (\sum_{i=1}^{N}\sum_{t=1}^{T}((x_{it} - \overline{x}_i)(x_{it} - \overline{x}_i)') + \psi T \sum_{i=1}^{N}((\overline{x}_i - \overline{x})(\overline{x}_i - \overline{x})'))^{-1}(\sum_{i=1}^{N}(\sum_{t=1}^{T}((x_{it} - \overline{x}_i)(y_{it} - \overline{y}_i))) + \psi T \sum_{i=1}^{N}((\overline{x}_i - \overline{x})(\overline{y}_i - \overline{y}))$
POLS	Linearità, Esogeneità, Rango pieno, Omoschedasticità	$y_{it} = x'_{it}\beta + \alpha_i + \varepsilon_{it}$	$\hat{\beta}_{POLS} = (\sum_i X'_i X_i)^{-1}\sum_i X'_i Y_i$
FD	Esogeneità, Rango pieno, Omoschedasticità	$\Delta y_{it} = \Delta x'_{it}\beta + \Delta\varepsilon_{it}$	$\hat{\beta}_{FD} = (\sum_{i=1}^{N}\sum_{t=2}^{T}(\Delta x'_{it}\Delta x_{it})^{-1})$ $(\sum_{i=1}^{N}\sum_{t=2}^{T}(\Delta y'_{it}\Delta y_{it}))$

2.2.2 La scelta dello stimatore

In prima approssimazione, la scelta dello stimatore può essere effettuata in relazione alla natura del dataset. Se il panel comprende osservazioni di un insieme limitato e fisso di unità d'interesse, è corretto presupporre che il modello a effetti fissi sia il più efficiente. Se il panel include osservazioni su un alto numero di individui selezionati casualmente, si presuppone che sia più adatto il modello a effetti casuali.

La scelta definitiva del modello più adatto per una specifica analisi, si effettua con lo studio di peculiari test statistici.

FE vs RE - Il test di Hausman

In presenza di correlazione tra α_i e x_{it} le stime RE non sono consistenti, mentre quelle FE continuano ad esserlo. Dunque, una differenza statisticamente significativa fra le stime FE e le stime RE viene interpretata come evidenza contro gli stimatori RE. Se invece le ipotesi di non correlazione tra le esplicative e gli effetti individuali è valida, RE produce stime consistenti e risulta più efficiente di FE. L'ipotesi nulla di non correlazione fra le esplicative e gli effetti individuali e l'ipotesi alternativa, può essere verificata con il test di Hausman.

L'idea alla base del test di Hausman è il confronto tra due stimatori, uno consistente sotto l'ipotesi nulla e sotto l'ipotesi alternativa, l'altro consistente ed efficiente sotto l'ipotesi nulla ma inconsistente sotto l'ipotesi alternativa. In questo contesto il test è utile per scegliere tra lo stimatore RE e FE, dove l'ipotesi nulla vuole

verificare che non vi sia correlazione tra gli effetti fissi e le x_{it}. Infatti, in presenza di correlazione non nulla tra α_i e x_{it} le stime RE sono non consistenti, mentre quelle FE continuano ad esserlo. Dunque, una differenza statisticamente significativa fra le stime FE e quelle RE viene interpretata come evidenza contro gli stimatori RE. Se invece l'ipotesi di non correlazione fra le esplicative e gli effetti individuali è valida ($E(\alpha_i|x_{1t}, x_{2t},...., x_{iT}) = E(\alpha_i) = 0$), RE produce stime consistenti ed è più efficiente di FE.

Viene verificata (sotto l'ipotesi di esogeneità in senso stretto) l'ipotesi

$$H_0: E(\alpha_i|x_{1t}, x_{2t},...., x_{iT}) = E(\alpha_i) = 0$$

$$H_1: E(\alpha_i|x_{1t}, x_{2t},...., x_{iT}) \neq E(\alpha_i)$$

Hausmann fornisce un test basato sulla statistica

$$W = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'[var(\hat{\beta}_{FE} - \hat{\beta}_{RE})]^{-1}(\hat{\beta}_{FE} - \hat{\beta}_{RE})$$

Che si distribuisce come un χ^2 con K gradi di libertà.

Se H_0 non viene rigettata dal test	$\hat{\beta}_{FE}$ è consistente ma inefficiente $\hat{\beta}_{RE}$ è consistente ed efficiente
Se H_0 viene rigettata dal test	$\hat{\beta}_{FE}$ rimane consistente $\hat{\beta}_{RE}$ è non consistente

Il risultato cruciale di Hausman è mostrare che la covarianza fra uno stimatore efficiente e la sua differenza con uno inefficiente è zero: $cov(\hat{\beta}_{RE}, (\hat{\beta}_{FE} - \hat{\beta}_{RE})) = cov(\hat{\beta}_{RE}, \hat{\beta}_{FE}) - var(\hat{\beta}_{RE}) = 0$, da cui sotto l'ipotesi di stretta esogeneità, rango pieno e omoschedasticità dei residui si ottiene

$var(\hat{\beta}_{FE} - \hat{\beta}_{RE}) = var(\hat{\beta}_{FE}) + var(\hat{\beta}_{RE}) - 2cov(\hat{\beta}_{FE}, \hat{\beta}_{RE}) = var(\hat{\beta}_{FE}) - var(\hat{\beta}_{RE})$ che è definita positiva.

La statistica di Hausman è $W = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'[var(\hat{\beta}_{FE} - \hat{\beta}_{RE})]^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE})$ la quale non è esente da criticità:

- viene condotto sotto l'ipotesi di stretta esogeneità $E(\varepsilon_{it}|x_{i1},...x_{iT}, \alpha_i)=0$. Dato che $\nu_{it} = \alpha_i + \varepsilon_{it}$ e il test viene condotto su $\hat{\nu}_{it}$, Ho può essere rigettata sia da $E(\alpha_i| x_{i1},...x_{iT}) \neq 0$ che da $E(\varepsilon_{it} | x_{i1},...x_{iT}) \neq 0$
- nel caso in cui $\hat{\beta}_{RE}$ non sia BLUE allora $var(\hat{\beta}_{FE} - \hat{\beta}_{RE}) \neq var(\hat{\beta}_{FE}) - var(\hat{\beta}_{RE})$, pertanto poiché nel calcolo di W viene indicato il primo termine dell'uguaglianza, essa è distorta e tende ad over rigettare l'ipotesi nulla a favore di FE.

Tali criticità vengono superate con il test di Wooldridge - overidentifying restrictions.

FE vs RE - Il test di Wooldridge - overidentifying restrictions

Un altro test di confronto per la scelta tra un modello a effetti fissi e uno a effetti casuali è il test di Wooldridge. Lo stimatore a effetti fissi utilizza le condizioni di ortogonalità, secondo le quali i regressori non sono correlati con l'errore idiosincratico e_{it}, ovvero $E(X_{it}e_{it}) = 0$, mentre lo stimatore degli effetti casuali utilizza le condizioni di ortogonalità aggiuntive, in base alle quali i regressori non sono correlati con l'errore specifico u_i, ovvero $E(X_{it}u_i)=0$. Queste condizioni di ortogonalità aggiuntive sono definite restrizioni di sovra identificazione (overidentifying restrictions). Il test utilizza l'approccio di regressione descritto da Wooldridge (2002), nel quale l'equazione degli effetti casuali viene nuovamente stimata utilizzando delle variabili aggiuntive calcolate come scostamento dei regressori originali che variano nel tempo dalle medie individuali.

La statistica test è un test di Wald attraverso il quale viene verificata l'ipotesi nulla che tutte le variabili create precedentemente siano congiuntamente uguali a zero. Se tale ipotesi viene rigettata significa che vengono rilevati gli effetti fissi, pertanto il modello a effetti casuali non è adeguato.

Siano quindi $\tilde{x}_{it}$ e $\tilde{y}_{it}$ le variabili calcolate come scostamento dei regressori originali che variano nel tempo dalle medie individuali:

$$\tilde{x}_{it} = x_{it} - \bar{x}_i \text{ e } \tilde{y}_{it} = y_{it} - \bar{y}_i$$

e w_{ij} un sottoinsieme di elementi time-varying di x_{it}. Definita la variabile $\ddot{w}_{it} = w_{it}-\bar{w}_i$, allora l'equazione degli effetti casuali estesa, utilizzando le variabili calcolate come scostamento dei regressori originali che variano nel tempo dalle medie individuali, è data da:

$$\tilde{y}_{it} = \tilde{x}_{it}\beta + \ddot{w}_{it}\xi + \varepsilon_{it} \qquad t=1,\dots,T;\ i=1,\dots,N$$

dove ξ è un vettore Mx1.

Il test di Hausman può essere implementato testando l'ipotesi nulla H_0: ξ = 0, attraverso una analisi pooled OLS. L'approccio più semplice è quello di calcolare una statistica test $F = [(SSR_r-SSR_{ur})/SSR_{ur}][(NT-K-M)/M]$, dove SSR_r e SSR_{ur} sono la somma dei quadrati dei residui ristretti e non ristretti. Se non viene rispettata l'ipotesi di omoschedasticità, è necessario utilizzare una forma robusta della statistica di Hausman. L'approccio più semplice è verificare l'ipotesi H_0: ξ = 0 attraverso una statistica Wald o sostituendo $\bar{w}_i$ a $\ddot{w}_{it}$. Quest'ultima intuizione è dovuta a Mundlak (1978), il quale si è posto l'obiettivo di verificare una assunzione diversa dell'omoschedasticità, utilizzando $E(c_i|x_i) = E(c_i|w_i)= y_0+\bar{w}_i y$. L'equivalenza dei due approcci deriva dal fatto che i regressori ($\tilde{x}_{it}$, $\ddot{w}_{it}$) sono una trasformazione non singolare dei regressori ($\tilde{x}_{it}$, $\bar{w}_i$), e quindi la somma dei quadrati dei residui nella regressione non ristretta sono gli stessi, così come quella ristretta.

La chiave dell'approccio di Mundlak (1978) è determinare se α_i e x_{it} sono correlati. Si consideri la media di α_i condizionata alla parte invariante nel tempo dei regressori, allora:

$$\alpha_i E(\alpha_i \mid x_i) = \bar{x}_i \theta + \nu_i \bar{x}_i \theta$$

In questa espressione, $\bar{x}_i$ è la media di x_{it}, e ν_i è una variabile non osservabile invariante nel tempo che non è correlata ai regressori.

Come nella regressione vogliamo testare che, se θ = 0, allora α_i e le covariate non sono correlate. Il modello è dato da:

$$y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it}$$

$$y_{it} = x_{it}\beta + \bar{x}_i \theta + \nu_i + \varepsilon_{it}$$

$$E(y_{it} \mid x_{it}) = x_{it}\beta + \bar{x}_i \theta$$

La seconda uguaglianza sostituisce α_i con $\bar{x}_i \theta + \nu_i$. La terza uguaglianza si basa sul fatto che i regressori e le variabili non osservabili sono indipendenti in media. Il test è dato da Ho: θ = 0.

**Effetti non correlati con le variabili esplicative: POLS vs RE*

Se il modello non contiene effetti latenti, gli stimatori pooled OLS sono efficienti. L'assenza del fattore non osservato è statisticamente equivalente all'ipotesi nulla H_0: $\sigma^2_\alpha = 0$, che può essere verificata con un test di

correlazione seriale o con il test di Breush e Pagan. Se non esiste eterogeneità individuale POLS è preferibile a RE.

**Effetti correlati con le variabili esplicative: FE vs FD*

Per t=2 le stime e i test statistici FD e le stime FE sono identiche. Per t=3 entrambi gli stimatori sono non distorti e consistenti.

Quando N è grande e T è piccolo la decisione deve essere basata sull'efficienza:

- se gli errori ε_{it} sono non serialmente correlati, FE è più efficiente;
- se gli errori ε_{it} seguono una random walk, FD è più efficiente;
- quando c'è qualche correlazione, ma non random walk, non si possono fare paragoni in termini di efficienza.

Quando N è piccolo e T è grande:

- FD è preferibile in presenza di elevata autocorrelazione positiva;
- l'inferenza con FE è molto sensibile a violazioni dell'ipotesi di normalità, omoschedasticità e correlazione seriale nulla negli errori idiosincratici;
- FE è meno sensibile all'ipotesi di esogeneità in senso stretto, quindi FE è preferibile quando i processi sono debolmente dipendenti rispetto al tempo.

2.2.3 I Modelli dinamici

Uno sviluppo della letteratura sui modelli di tipo panel è quello relativo ai panel dinamici, caratterizzati dalla presenza della variabile dipendente ritardata all'interno della matrice dei regressori. Il problema principale di questo tipo di modelli è dato dal fatto che il termine di errore non è incorrelato con la variabile ritardata, questo genera stime OLS e GLS inconsistenti. La soluzione a tale inconveniente è quella di considerare un modello in termini di differenze prime e ricorrere allo stimatore a variabili strumentali (Baltagi, 2005; Arellano e Bond, 1991). Gli stimatori più utilizzati e conosciuti sono lo stimatore di Anderson-Hsiao e Arellano-Bond. Di tali modelli si parlerà in modo più esteso ed esaustivo nel Capitolo 5.

2.2.4 Ulteriori metodologie (Baltagi, B.H., 2005)

<u>Panel dati non bilanciati</u>

Fino ad ora sono stati affrontati modelli con panel dati bilanciati. Può capitare però che i panel dati siano non bilanciati. Ad esempio, nella raccolta di dati sulle compagnie aeree statunitensi nel tempo, un ricercatore potrebbe scoprire che alcune aziende hanno abbandonato il mercato, mentre ci sono nuove compagnie che potrebbero far parte del campione. Allo stesso modo, durante l'utilizzo dei panel sulle famiglie, si potrebbe scoprire che alcune famiglie si sono trasferite e non possono più essere incluse nel panel. Inoltre, se si raccolgono dati su un insieme di paesi nel tempo, un ricercatore può capire che alcuni paesi hanno serie storiche più lunghe rispetto ad altri paesi. Questi scenari determinano panel sbilanciati o incompleti, a causa della mancanza casuale di osservazioni. Per la stima di tali modelli sono state sviluppate valide metodologie:

- Gli stimatori ANOVA sono i migliori stimatori con errori quadratici non distorti di componenti della varianza (Searle, 1971). Sono disponibili metodi ANOVA sbilanciati, ma perdono alcune proprietà. Pertanto vengono generalmente adattati i metodi ANOVA di panel bilanciati al caso sbilanciato;

- Un metodo alternativo per stimare le componenti della varianza è la stima di massima verosimiglianza (ML). Gli stimatori di massima verosimiglianza sono funzioni di statistiche sufficienti e sono coerenti e asintoticamente efficiente (Harville, 1977). L'approccio ML è stato però criticato in quanto non tiene conto della perdita di gradi di libertà dovuta ai coefficienti di regressione nella stima delle componenti della varianza;

- Sotto l'ipotesi di normalità dei residui, è possibile utilizzare indifferentemente le procedure MINQUE (Minimum Norm Quadratic Unbiased Estimators) e MIVQUE (Minimum Variance Quadratic Unbiased Estimators) (Rao, 1971) per la stima delle componenti della varianza;

- Baltagi e Chang (1994), hanno affrontato il problema sviluppando il metodo Monte Carlo;

- Nel caso di modello fixed effect sbilanciato, Wansbeek e Kapteyn (1989) hanno mostrato che la trasformazione è più complicata ma comunque gestibile.

Panel data con variabili a scelta discreta

In molti studi economici, la variabile dipendente è discreta, indicando ad esempio che la famiglia ha acquistato un'auto o che una persona è disoccupata o che è entrata a far parte di un sindacato o è inadempiente su un prestito o gli è stato negato il credito. Questa variabile dipendente è solitamente rappresentata da una variabile di scelta binaria e che viene indicata da $y_{it} = 1$ se l'evento accade e $y_{it} = 0$ se non accade, dove con i viene indicato l'individuo e con t il tempo. Se p_{it} è la probabilità che l'individuo partecipi alla forza lavoro al tempo t, allora $E(y_{it}) = 1 \cdot p_{it} + 0 \cdot (1 - p_{it}) = p_{it}$, e considerando le variabili esplicative x_{it}, $p_{it} = Pr[y_{it} = 1] = E(y_{it} / x_{it}) = F(x'_{it}\beta)$.

Per il modello di probabilità lineare, $F(x'_{it}\beta) = x'_{it}\beta$ si applicano i metodi usuali dei dati panel, tranne il caso in cui non venga garantito che $\hat{y}_{it}$ si trovi nell'intervallo unitario. La soluzione standard è stata quella di utilizzare le funzioni di distribuzione cumulative logistica o normale che vincolano $F(x'_{it}\beta)$ ad essere tra zero e uno. Queste funzioni di probabilità sono conosciute in letteratura come logit e probit, e corrispondono rispettivamente alla distribuzione logistica e normale. Ad esempio, quando il lavoratore partecipa alla forza lavoro e il suo salario supera la sua riserva inosservata di salario. Questa soglia può essere descritta come:

$$y_{it} = 1 \text{ if } y^*_{it} > 0$$

$$y_{it} = 0 \text{ if } y^*_{it} \leq 0$$

dove

$$y^*_{it} = x_{it}\beta + u_{it}$$

e

$$Pr[y_{it} = 1] = Pr[y^*_{it} > 0] = Pr[u_{it} > -x'_{it}\beta] = F(x'_{it}\beta)$$

Nello specifico, l'ultima uguaglianza vale fino a che la funzione di densità che descrive F è simmetrica intorno allo zero. Tutto ciò è vero per le funzioni logistiche e di densità normale. Generalmente il modello logit viene

utilizzato per la soluzione di panel ad effetti fissi (Chamberlain, 1980 e 1984), (Winkelmann e Winkelmann, 1998), mentre la specifica probit è più utilizzata per il modello ad effetti casuali (Sickles e Taubman, 1986).

Panel non stazionari

Con il crescente utilizzo di dati cross-country nel tempo, il focus dell'econometria dei dati panel si è spostato verso lo studio dei panel asintotici macro con N grande (numero di paesi) e T grande (lunghezza delle serie temporali), piuttosto che sui soliti micro panel asintotici con N grande N e T piccola. Il fatto che T possa aumentare all'infinito nei dati macro panel, ha generato due filoni di studi. Fanno parte del primo filone: Pesaran e Smith (1995), Im, Pesaran e Shin (2003), Lee, Pesaran e Smith (1997), Pesaran, Shin e Smith (1999) e Pesaran e Zhao (1999). Questa letteratura considera, in modo critico, il fatto che T sia grande per stimare separatamente la regressione di ogni paese, e sconsiglia l'uso di stimatori come Fixed Effect, per stimare il modello di dati panel dinamico. Tale letteratura sostiene che questi modelli sono soggetti a un ampio potenziale di bias, quando i parametri sono eterogenei tra i paesi, e i regressori sono serialmente correlati. Il secondo filone della letteratura ha applicato le procedure delle serie temporali panel, preoccupandosi della non stazionarietà, delle regressioni spurie e della cointegrazione. Phillips e Moon (2000) sostengono che la serie temporale di variabili, quali la crescita del PIL pro capite, hanno una forte non stazionarietà. Questo è comprensibile nel caso di panel micro, mentre nel caso di macro panel, la non stazionarietà merita maggiore attenzione.

Panel rotanti

Biorn (1981) considera il caso dei panel rotanti. Per mantenere lo stesso numero di famiglie in un sondaggio, la frazione di famiglie che nel secondo periodo del sondaggio non può più partecipare al panel, viene sostituito da un numero uguale di nuove famiglie appena intervistate. Nello studio di Biorn e Jansen (1983), basato su dati delle indagini sul bilancio familiare, viene proposto di ruotare metà del campione per ogni periodo. In altre parole, metà delle famiglie intervistate escono dal campione ogni periodo e vengono sostituite da nuove famiglie. La metodologia applicata è il GLS.

Pseudo Panel

Per alcuni paesi, i dati del panel potrebbero non esistere. In questi casi il ricercatore può trovare delle indagini annuali sulle famiglie basate su un ampio campione casuale della popolazione. Esempi di alcuni di questi sono le indagini trasversali sulla spesa dei consumatori quali: i) l'indagine sulla spesa familiare nel Regno Unito, che esamina circa 7000 famiglie all'anno; ii) una serie di indagini sulle famiglie da paesi meno sviluppati come le indagini sulle famiglie della Banca mondiale. Gli pseudo-panel hanno il vantaggio di non soffrire del problema di attrito, che affligge invece i dati panel, e possono essere inoltre disponibili per periodi di tempo più lunghi rispetto a questi ultimi. Deaton (1985), suggerisce di tenere traccia delle coorti e di stimare le relazioni economiche basate sulle coorti invece che su osservazioni individuali. Una coorte, ad esempio, potrebbe essere l'insieme di tutti i maschi nati tra il 1945 e il 1950. Infatti, questa coorte di nascita è ben definita e può essere facilmente identificata dai dati.

Panel spaziali

Esiste una vasta letteratura che utilizza le statistiche spaziali e che tratta la correlazione spaziale. I modelli di dipendenza spaziale sono popolari nella scienza regionale e nell'economia urbana. Più specificamente, questi modelli si occupano dell'interazione spaziale (autocorrelazione spaziale) e struttura spaziale (eterogeneità

spaziale) (Anselin (1988, 2001)). Con la crescente disponibilità di dati panel di livello micro e macro, i modelli di dati panel spaziali stanno diventando sempre più numerosi e attraenti nella ricerca economica empirica.

Data la natura delle unità spaziali dei modelli panel, è possibile ipotizzare l'esistenza di una dipendenza spaziale tra esse, che consente di verificare se la presenza di effetti geografici forniscono una spiegazione migliore del modello di regressione. Un filone di letteratura (Fingleton, McCombie, 1998; Fingleton, 2001; 2003; Anselin, 1998) ha mostrato come la presenza di significativi *spillovers* regionali determini delle stime OLS distorte o rappresenti una delle principali cause della non significatività dei relativi test statistici.

Per misurare la correlazione spaziale esistente tra le varie unità che compongono il dataset, è necessario costruire una matrice dei pesi spaziali W, di dimensione *N x N*, tale per cui:

$$W = \begin{bmatrix} 0 & \cdots & w_{1,j} \\ \vdots & \ddots & \vdots \\ w_{i,1} & \cdots & 0 \end{bmatrix}$$

Ogni elemento della matrice w_{ij} rappresenta l'interazione esistente tra l'unità *i* in riga e l'unità *j* in colonna. Gli elementi lungo la diagonale principale sono, per convenzione, pari a 0. Nel caso più semplice, anche per avere una certa normalizzazione dei pesi in modo che varino tra 0 ed 1, il peso sarà uguale a 1 se esiste dipendenza spaziale, 0 viceversa. L'assegnazione del peso dell'interazione nella matrice di contiguità può seguire diversi criteri. Un modo è quello di considerare l'inverso della distanza tra i centroidi delle due unità spaziali.

Grazie alla matrice spaziale, quindi, è possibile assegnare un peso quantitativo agli spillovers spaziali, che, tuttavia, possono dipendere sia dalla posizione che tale unità spaziale assume nello spazio considerato, sia da come essa interagisce con le altre unità spaziali confinanti. Per capire se effettivamente esista una autocorrelazione spaziale tra le unità statistiche, si può calcolare l'indice I di Moran (Moran, 1950):

$$I= (N/\Sigma\Sigma W_{ij}N_j)(\Sigma\Sigma(x_i-\overline{x})(x_j-\overline{x})/\Sigma(x_i-\overline{x})^2)$$

dove N è il numero delle unità geografiche spaziali, W è la matrice di contiguità dei pesi spaziali, x_i e x_j sono le variabili rispettivamente dello spazio i e dello spazio j, $\overline{x}$ è la media campionaria. L'indice ha un range compreso tra -1 ed 1: tanto più si avvicina ad 1 (-1), tanto positiva (negativa) è la correlazione spaziale.

La matrice dei pesi spaziali può dar luogo a due macro categorie di modelli con effetti spaziali:

1. *lag spaziali*, nei quali è la variabile dipendente a subire gli effetti geografici;
2. *errori spaziali*, in cui la matrice interviene sul termine d'errore.

Inoltre, queste due categorie possono anche combinarsi tra loro, dando origine a diversi modelli spaziali di ordine superiore.

In letteratura, esistono molteplici modelli che rientrano nella branchia dell'econometria spaziale. Di seguito vengono elencati i modelli maggiormente utilizzati:

1) modello SDM (Spatial Durbin Model), che tiene conto sia degli effetti esogeni che endogeni:

$$y_{it} = \alpha_i + \beta \mathbf{X}_{it} + \rho W_{ij}\, y_{it} + \delta W_{ij}\, \mathbf{X}_{it} + \varepsilon_{it}$$

2) modello SAR (Spatial Autoregressive Model), che tiene conto solo degli effetti endogeni:

$$y_{it} = \alpha_i + \beta \mathbf{X}_{it} + \rho W_{ij}\, y_{it} + \varepsilon_{it}$$

3) modello SAC (Spatial Autoregressive with Spatially Autocorrelated Error Model), che è un'estensione del modello SAR considerando anche autocorrelazione spaziale nel termine d'errore:

$$y_{it} = \alpha_i + \beta \mathbf{X}_{it} + \rho W_{ij}\, y_{it} + \varepsilon_{it} + \upsilon_{it}$$

$$\upsilon_{it} = \lambda W\, \upsilon_i + u_{it}$$

4) modello SEM (Spatial Error Model), che tiene conto degli effetti spaziali sul termine d'errore:

$$y_{it} = \alpha_i + \beta X_{it} + \upsilon_{it}$$

$$\upsilon_{it} = \lambda W \upsilon_{it} + \varepsilon_{it}$$

Per i primi tre modelli, la stessa unità spaziale può influenzare sia la variabile dipendente che le altre variabili esplicative spazialmente correlate. Per cui, è possibile distinguere tra gli effetti marginali diretti e gli effetti marginali indiretti, i quali, uniti, danno naturalmente vita agli effetti marginali totali. Gli effetti marginali diretti indicano l'effetto che una variazione di un'esplicativa dell'unità spaziale i subisce confrontato con la media della stessa esplicativa rispetto a tutte le unità spaziali. Gli effetti marginali indiretti, invece, misurano l'effetto che ha una variazione dell'esplicativa di tutte le unità spaziali sulla stessa variabile dell'unità spaziale i, sempre in media rispetto a tutte le unità spaziali. L'effetto marginale totale indica, invece, l'effetto di una variazione dell'esplicativa di tutte le unità spaziali, che colpisce l'unità spaziale i, sempre in media rispetto a tutte le unità spaziali.

Questi effetti possono essere suddivisi anche tra breve e lungo periodo: i primi si hanno qualora siano presenti tra le esplicative i ritardi della dipendente o di una o più esplicative, mentre gli effetti di lungo periodo si hanno a prescindere dall'inclusione dei ritardi tra le variabili indipendenti.

Panel eterogenei

Con l'aumento della dimensione temporale di dati panel, alcuni ricercatori (Robertson e Symons, 1992; Pesaran e Smith 1995), hanno messo in dubbio la comparabilità dei dati tra unità eterogenee.

BIBLIOGRAFIA CAPITOLO 2

Anselin, L., 1988, *Spatial Econometrics: Methods and Models* (Kluwer Academic Publishers, Dordrecht).

Anselin L. (1998), Explanatory spatial data analysis in a geocomputational environment, in Longley P. Brooks A., McDonnell S.M., Macmillan B. (eds), *Geocomputation: a primer*, Wiley, Chichester.

Anselin, L., 2001, Spatial econometrics, Chapter 14 in B. Baltagi, ed., *A Companion to Theoretical Econometrics* (Blackwell Publishers, Massachusetts), 310–330.

Arellano, M. e Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. Review of Economic Studies.

Arellano, M. e Bover, O. (1995). Another look at instrumental variable estimation of error components model. Journal of Econometrics 68 (1995) 29-51

Ashenfelter, O. e Krueger, K. (1994). Estimates of the economic returns to schooling from a new sample of twins. The American economic Review.

Baltagi, B.H. (1984). Panel data methods. Department of Economics Texas A&M University

Baltagi, B.H. (2005). Econometric Analysis of Panel data. John Wiley and Son. Third Edition

Baltagi, B.H. e Y.J. Chang, 1994, Incomplete panels: A comparative study of alternative estimators for the unbalanced one-way error component regression model, *Journal of Econometrics* **62**, 67–89.

Belloc, M. (2011). Modelli Panel. Lezioni di dottorato. Econometria III. Università La Sapienza Roma.

Benfratello, L. (2013). System of Equations and Linear (Static and Dynamic) panel data models. Master in econometria applicata 2012-2013 SSEF

Benfratello, L. (2015). Linear (Static and Dynamic) panel data models. Scuola Nazione dell'Amministrazione (SNA)

Biorn, E., 1981, Estimating economic relations from incomplete cross-section/time-series data, *Journal of Econometrics* **16**, 221–236.

*Biorn, E. e E.S. Jansen, 1983, Individual effects in a system of demand functions, Scandinavian Journal of Economics **85**, 461–483.*

Blundell, R. Bond, S. (1998). GMM estimation with persistent panel data: an application to production function. Eight International Conference on Panel data – Goteborg University – June 11-12, 1998

Chamberlain, G., 1980, Analysis of covariance with qualitative data, *Review of Economic Studies* **47**, 225–238.

Chamberlain, G., 1984, Panel data, Chapter 22 in Z. Griliches and M. Intrilligator, eds., *Handbook of Econometrics* (North-Holland, Amsterdam), 1247–1318.

Deaton, A., 1985, Panel data from time series of cross-sections, *Journal of Econometrics* **30**, 109–126.

Fingleton B., McCombie J. (1998) Increasing returns and economic growth: some evidence for manufacturing from the European Union regions, *Oxford Economic Papers*, 50, 89-105.

Fingleton, B. (2001) Theoretical Economic Geography and SpatialEconometrics: Dynamic Perspectives, *Journal of EconomicGeography*, 1, 2, 201-225.

Fingleton B. (2003) Externalities, Economic Geography and Spatial Econometrics. Conceptual and Modeling Development, *International Regional Science Review*, 26, 197-207.

Frisch R. and Waugh F. V. (1933) Partial Time Regressions as Compared with Individual Trends, Econometrica, 1,387-401.

Gallo, F. Mastrovita, S. e Siciliani, I. (2004). Il processo di produzione dell'Indagine ECHP. ISTAT.

Gallo, F. Mastrovita, S. e Siciliani, I. (2004). Un'analisi dell'attrition e delle sue determinanti nell'indagine panel europeo sulle famiglie. ISTAT.

Harville, D.A., 1977, Maximum likelihood approaches to variance component estimation and to related problems, *Journal of the American Statistical Association* **72**, 320–340.

Hsiao, C. (2003). Analysis of panel data. Cambridge University Press.

Klevmarken, A. (1989). Modelling labor supply in a dynamic economy.

Im, K.S., M.H. Pesaran and Y. Shin, 2003, Testing for unit roots in heterogeneous panels, *Journal of Econometrics* **115**, 53–74.

ISTAT (2008). L'indagine europea sui redditi e le condizioni di vita delle famiglie (Eu-Silc). Metodi e Norme n. 37

Lee, K., M.H. Pesaran and R. Smith, 1997, Growth and convergence in a multi-country empirical stochastic Solow model, *Journal of Applied Econometrics* **12**, 357–392.

Lovell M. C. (1963) Seasonal Adjustment of Economic Time Series, Journal of the American Statistical Association, 58, 993-1010.

Moran, P. A. P. (1950), *Notes on continuos stochastic phenomena*, Biometrika, vol. 37, pp. 17-33

Mundlak, Y. (1961) Empirical Production Function Free of Management Bias, *American Journal of Agricultural Economics*, 43, issue 1, p. 44-56, https://EconPapers.repec.org/RePEc:oup:ajagec:v:43:y:1961:i:1:p:44-56.

Mundlak, Y. (1978) On the pooling of time series and cross section data, *Econometrica*, 46, No. 1.

Roodman, D. (2009). How to do xtabond2: An introduction to difference and system GMM in STATA.

Pesaran, M.H. and R. Smith, 1995, Estimating long-run relationships from dynamic heterogenous panels, *Journal of Econometrics* **68**, 79–113.

Pesaran, M.H., Y. Shin and R. Smith, 1999, Pooled mean group estimation of dynamic heterogeneous panels, *Journal of the American Statistical Association* **94**, 621–634.

Pesaran, M.H. and Z. Zhao, 1999, Bias reduction in estimating long-run relationships from dynamic heterogeneous panels, Chapter 12 in C. Hsiao, K. Lahiri, L.F. Lee and M.H. Persaran, eds., *Analysis of Panels and Limited Dependent Variable Models* (Cambridge University Press, Cambridge), 297–322.

Rao, C.R., 1970, Estimation of heteroscedastic variances in linear models, *Journal of the American Statistical Association* **65**, 161–172.

Phillips, P.C.B. e H. Moon, 2000, Nonstationary panel data analysis: An overview of some recent developments, *Econometric Reviews* **19**, 263–286.

Searle, S.R., 1971, *Linear Models* (John Wiley, New York).

*Sickles, R.C. e P. Taubman, 1986, A multivariate error components analysis of the health and retirement study of the elderly, Econometrica **54**, 1339–1356.*

Stock, J.H. e Watson, M.W. (2009). Introduction to Econometric. Pearson.

Wansbeek, T.J. e A. Kapteyn, 1989, Estimation of the error components model with incomplete panels, *Journal of Econometrics* **41**, 341–361.

*Winkelmann, L. e R.Winkelmann, 1998, Why are the unemployed so unhappy? Evidence from panel data, Economica **65**, 1–15.*

Wooldridge, J. M. (2002). Econometric Analysis of cross section and panel data.

Wooldridge, J. M. (2009). Introductory Econometrics: a modern approach.

Wooldridge, J. M. (2010). Econometric Analysis of Cross Section and Panel data. The MIT Press.

3 – Il caso italiano – Analisi per regioni

3.1 I dati sull'Economia Sommersa

In accordo con l'ipotesi che il lavoro irregolare è "il principale fattore produttivo su cui si basa il funzionamento dell'Economia Sommersa" (Lucifora, 2003), viene considerata come variabile dipendente il tasso di irregolarità del lavoro pubblicato dall'ISTAT, calcolato come la quota percentuale delle unità di lavoro irregolari sul totale delle unità di lavoro (irr) (ISTAT, 2015, 2017, 2018).

Il concetto di occupazione regolare e non regolare è strettamente connesso a quello di attività produttive osservabili e non osservabili, comprese nei confini della produzione del sistema di contabilità nazionale. In particolare, le prestazioni non regolari sono definite come quelle attività lavorative svolte senza il rispetto della normativa vigente in materia fiscale-contributiva, quindi non osservabili direttamente presso le imprese, le istituzioni e le fonti amministrative. Rientrano in questa categoria le prestazioni lavorative continuative, svolte non rispettando la normativa vigente; occasionali, svolte da persone non attive in quanto studenti, casalinghe o pensionati; svolte dagli stranieri non residenti e non regolari; plurime, cioè attività ulteriori rispetto alla principale non dichiarata alle istituzioni fiscali. Per misurare il lavoro non regolare si fa riferimento al concetto di unità di lavoro (ULA), che rappresenta una misura di quanto il fattore lavoro contribuisca alla produzione del Paese in un determinato periodo. Le unità di lavoro sono calcolate attraverso la trasformazione in unità a tempo pieno delle posizioni lavorative ricoperte da ciascuna persona occupata nel periodo di riferimento. Il tasso di irregolarità del lavoro è quindi costruito come rapporto percentuale tra unità di lavoro non regolare[17]e unità di lavoro totali.

Su indicazione dell'OCSE, l'ISTAT calcola il fattore lavoro (input lavoro[18]), non solo in termini di unità di lavoro (ULA[19]), ma anche in termini di ore lavorate e occupati. Per ognuno di essi viene calcolata la percentuale dovuta all'attività irregolare. I dati al livello regionale relativi al tasso di irregolarità del lavoro, costruito come rapporto percentuale tra unità di lavoro non regolare e unità di lavoro totali (irr) sono attualmente fermi al 2012. Sono invece stati rilevati dal database on-line dell'ISTAT (I.Stat) dati al livello regionale più aggiornati relativi al fattore lavoro in termini di occupati irregolari (irr1). Per quest'ultima sono stati considerati dati al livello regionale fino al 2015.

Quest'ultima variabile, quale proxy dell'Economia Sommersa, non viene usualmente utilizzata negli studi econometrici. Tale variabile appare pertanto meritevole di approfondimento quale elemento di novità nell'ambito della letteratura dedicata all'argomento, anche in considerazione di un facile reperimento di dati aggiornati.

Si vogliono esplorare le due diverse definizioni del tasso di irregolarità del lavoro usando le variabili sopra definite, i cui dati sono esposti in Tabella 1. I dati sono disponibili per tutte le 20 regioni italiane per l'arco

[17]Si riferiscono alle unità di lavoro relative a prestazioni lavorative svolte senza il rispetto della normativa vigente in materia di lavoro, fiscale e contributiva, quindi non osservabili direttamente presso le imprese, le istituzioni e le fonti amministrative.

[18]L'input lavoro è la quantità di lavoro utilizzata (in modo regolare o irregolare) dal sistema produttivo. Questa misurazione consente di integrare metodi indiretti di stima con fonti statistiche ed amministrative, cercando in questo modo di minimizzare i problemi derivanti dalla presenza di attività produttive non osservabili, e di cogliere indirettamente queste attività. Il primo concetto, che si tiene in considerazione, per misurare l'input di lavoro è il numero degli occupati, cioè quelle persone, dipendenti o indipendenti, che esercitano un'attività in unità produttive residenti, indipendentemente dalla loro nazionalità e dalla durata della prestazione (tempo pieno o tempo parziale).

[19]Secondo la definizione dell'ISTAT, le unità di lavoro misurano in modo omogeneo il volume di lavoro prestato da tutti coloro i quali, a prescindere dalla propria residenza, concorrono alle attività di produzione realizzate sul territorio economico di un Paese. Le unità di lavoro rappresentano tutte le posizioni lavorative (principali o secondarie) ricoperte dagli occupati, trasformate in unità equivalenti a tempo pieno. Come stabilito dal regolamento dei conti nazionali (SEC2010), le unità di lavoro sono calcolate come rapporto tra il totale delle ore effettivamente lavorate e il numero medio di ore lavorate a tempo pieno.

temporale 2001-2015 ad esclusione del tasso di irregolarità del lavoro calcolato come percentuale di unità di lavoro irregolari sul totale delle unità di lavoro (**irr**) e del asso di partecipazione all'istruzione secondaria superiore (**istr**), variabili per le quali sono disponibili i dati regionali solo per il periodo 2001-2012.

La scelta delle variabili esplicative da includere nell'analisi è stata suggerita da un attento studio della letteratura economica in materia di Economia Sommersa:

- **Struttura socio-demografica.** Si è scelto di inserire la densità di popolazione (**dens**), al fine di considerare l'eterogeneità della distribuzione della popolazione sul territorio italiano (Morvillo, 2016), il tasso di partecipazione all'istruzione secondaria superiore e di terzo livello (**istr e istr1**) che ha un ruolo di contrasto rispetto al fenomeno in esame (Cappariello e Zizza, 2009);
- **Struttura economica regionale**. Fanno parte di questo gruppo le entrate tributarie (**tax**) (Amendola e Dell'Anno, 2008; Dell'Anno, 2003), variabile che si pone l'obiettivo di sintetizzare il contesto economico-istituzionale comune a tutte le 20 regioni; il tasso di industrializzazione (**indu**), il tasso di imprenditorialità femminile e giovanile (**imprendfem** e **impredgiov**) e il tasso di natalità delle imprese (**natalitàimprend**), poiché nelle regioni in cui la dotazione di industrie e una presenza di imprese medio grandi è particolarmente carente, ci si attende una maggior diffusione di Economia Sommersa (Zizza, 2002; Daniele e Marani, 2008);
- **Variabili di controllo**. Tra le variabili di controllo vengono considerate le determinanti del sommerso (Amendola e Dell'Anno, 2008; Lisi, 2009). Rappresentano questo gruppo il PIL pro capite regionale (**pilpc**), in grado di fornire la dimensione della crescita economica locale. Nella letteratura empirica si evidenzia una notevole eterogeneità rispetto alle relazioni trovate tra questa variabile e gli indicatori di Economia Sommersa, mostrando così una situazione di ambiguità rispetto al segno della relazione (Dell'Anno, 2003; Busato e Chiarini, 2004). Infatti, le attività irregolari sono da considerarsi, per alcuni studi, anticicliche. Più precisamente il settore sommerso appare con caratteristiche anticicliche (Chiarini, 2004), indicando che questo può esercitare un qualche ruolo di "copertura" per famiglie e imprese durante le fasi negative del ciclo. D'altra parte, tali attività contribuiscono al reddito e alla produzione nazionale e quindi ci si potrebbe aspettare un segno positivo della relazione. Altre variabili di particolare interesse sono la partecipazione femminile al mercato del lavoro (**fem**) e la disoccupazione giovanile (**disocgiov**). Queste variabili sono state incluse seguendo Lucifora (2003), nel quale si afferma "I paesi con maggiori livelli di partecipazione femminile al mercato del lavoro presentano dimensioni minori di Economia Sommersa. Ci si attende un effetto negativo sulle dimensioni del sommerso. Tuttavia, controllando per la disoccupazione, una maggior partecipazione può aumentare alcuni tipi di lavoro irregolare.". Infine, è stata considerata l'intensità della regolamentazione, indicatore espresso in questo studio dal rapporto tra il numero dei dipendenti pubblici regionali e il numero di lavoratori nella stessa regione (**dippub**), utile a fornire una fotografia del contesto istituzionale italiano. La teoria economica fa generalmente riferimento a indicatori di regolamentazione costruiti tenendo conto dello stock di tutte le leggi in vigore, dello Stato e degli enti locali, relative all'accesso al lavoro, alla sicurezza sociale, alle ore lavorative, alle condizioni di lavoro, all'esercizio dell'attività d'impresa. Nei sistemi economici più regolamentati si verifica, in genere, una maggiore diffusione delle attività sommerse, poiché l'intensità della regolamentazione comporta un aumento dei costi necessari a svolgere un'attività economica. L'Economia Sommersa, per le imprese meno efficienti, rappresenta una modalità di riduzione di tali costi. Essendo difficile reperire indicatori di questo tipo per la realtà regionale italiana, in questo contesto si è deciso di utilizzare quale indicatore di regolamentazione, il rapporto tra i dipendenti ascrivibili al pubblico impiego e le forze di lavoro in età 15-64 anni. Tale indicatore, così come

costruito, è stato utilizzato da Frey e Weck-Hanneman nel 1984 (Frey e Weck-Hanmeman, 1984; Zizza, 2002).

- **Diffusione della criminalità**. Vi sono poi una serie di indicatori che cercano di cogliere l'incidenza del fenomeno della criminalità, in generale sul territorio italiano, utilizzando la percentuale di famiglie che avvertono, con un livello medio-alto, il disagio del rischio di criminalità nella zona in cui vivono rispetto al totale delle famiglie (**crimperc**). In più sono state considerate alcune variabili, che forniscono informazioni in merito alla reale attività criminosa diffusa nel Paese (**furti**, **rapine** e **omicidi**), che influiscono positivamente sull'Economia Sommersa (Daniele e Marani, 2008; Marini e Turato, 2002). In particolare, il tasso di furti e il tasso delle rapine denunciate vengono calcolate su 1.000 abitanti, mentre il tasso di omicidi volontari commessi, su 100.000 abitanti.

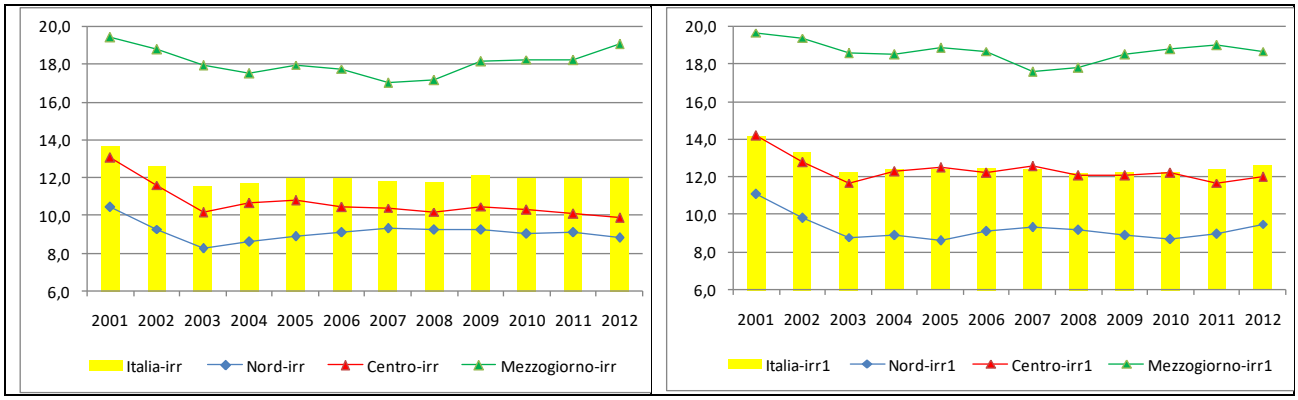
Tabella1: Elenco dettagliato delle variabili utilizzate e delle fonti utilizzate

Sigla		Descrizione	Fonte	Segno atteso
Irr	*Variabile oggetto di studio*	Tasso di irregolarità del lavoro calcolato come percentuale di unità di lavoro irregolari sul totale delle unità di lavoro	ISTAT, (2001-2012)	
irr1		Tasso di irregolarità del lavoro calcolato come percentuale di occupati irregolari sul totale degli occupati	ISTAT, (2001-2015)	
Dens	*Struttura demografica delle regioni*	Densità abitativa calcolata come rapporto tra numero di residenti e superficie territoriale in kmq	ISTAT, (2001-2015)	**-**
Istr		Tasso di partecipazione all'istruzione secondaria superiore	ISTAT, (2001-2012)	**-**
istr1		Tasso di partecipazione all'istruzione di terzo livello	EUROSTAT, (2001-2015)	**-**
Indu	*Struttura economica delle regioni*	Quota di occupati nell'industria in senso stretto (attività estrattiva; attività manifatturiere; fornitura di energia elettrica, gas, vapore e aria condizionata; fornitura di acqua; reti fognarie, attività di trattamento dei rifiuti e risanamento) sul totale degli occupati	ISTAT, (2001-2015)	**-**
Tax		Quota delle entrate tributarie sul prodotto interno lordo (variabile individual invariant)	MEF-DF, (2001-2015)	**+**
imprenfem		Totali imprese con titolare femminile sul numero totale di imprese	ISTAT, (2001-2015)	**-**
imprengiov		Totali imprese con titolare minore di 30 anni sul numero totale di imprese	ISTAT, (2001-2015)	**-**
natalitàimp		Rapporto tra imprese nate all'anno t e le imprese attive dello stesso anno	ISTAT, (2001-2015)	**-**
fem	*Variabili di controllo*	Tasso di occupazione femminile pari al rapporto tra il numero di donne occupate in età 15-64 anni sulla popolazione	ISTAT, (2001-2015)	**-**
pilpc		Prodotto Interno Lordo ai prezzi di mercato (prezzi correnti) per abitante	ISTAT, (2001-2015)	**+/-**
dippub		Rapporto tra i dipendenti ascrivibili al pubblico impiego e le forze di lavoro in età 15-64 anni	MEF, (2001-2015); ISTAT, (2001-2015)	**-**
disocgiov		Percentuale Persone in cerca di occupazione in età 15-24 anni su forze di lavoro stessa classe di età	ISTAT, (2001-2015)	**+**
crimperc	*Variabili sulla criminalità*	Percezione delle famiglie del rischio di criminalità nella zona in cui vivono	ISTAT, (2001-2015)	**+**
furti		Furti denunciati per 1.000 abitanti	ISTAT, (2001-2015)	**+**
rapine		Rapine denunciate per 1.000 abitanti	ISTAT, (2001-2015)	**+**
Omicidi		Omicidi volontari consumati per 100.000 abitanti	ISTAT, (2001-2015)	**+**

3.2 Analisi descrittiva delle variabili oggetto di studio

In questo paragrafo si riporta un'analisi esplorativa delle due proxy (irr e irr1) del fenomeno Economia Sommersa.

Figura 1: Tasso di irregolarità del lavoro nelle due accezioni scelte (irr e irr1) in Italia e per ripartizione geografica, 2001-2012

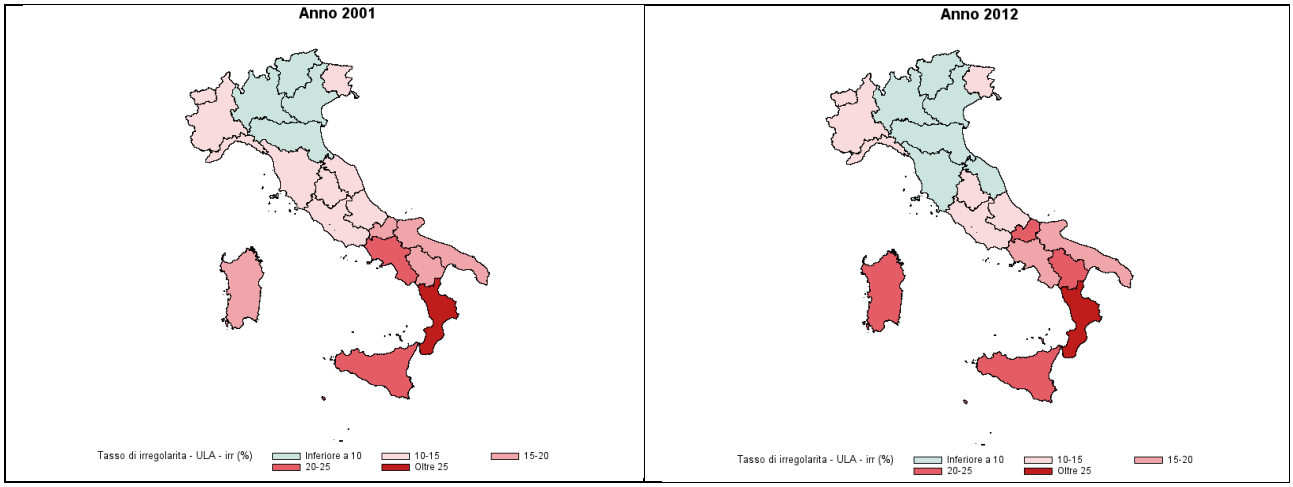


Fonte: Elaborazione su dati ISTAT

Le variabili irr e irr1 mostrano un andamento simile, sia per ripartizione geografica che per anno, sebbene a livelli percentuali lievemente differenti.

Le mappe confermano, per entrambe le accezioni, valori più elevati per le regioni meridionali rispetto a quelle settentrionali[20].

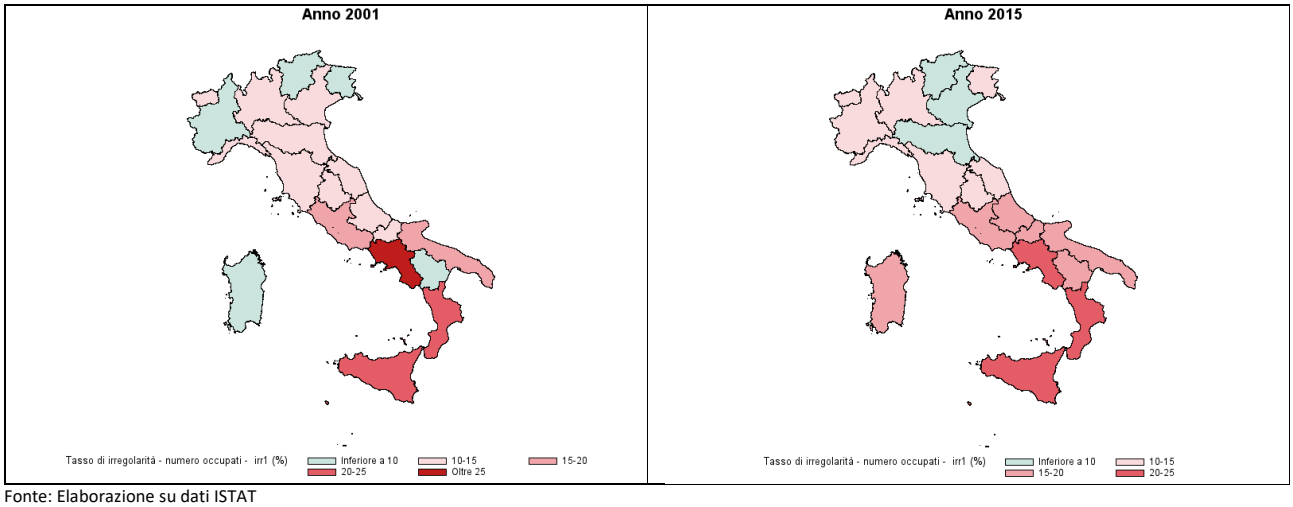
Figura 2: Tasso di irregolarità del lavoro nell'accezione irr per regione, 2001 e 2012



Fonte: Elaborazione su dati ISTAT

[20]Figura A1 e A2 in appendice

Figura 3: Tasso di irregolarità del lavoro nell'accezione irr1 per regione, 2001 e 2015



Fonte: Elaborazione su dati ISTAT

3.3 I modelli panel statici per lo studio dell'Economia Sommersa

Di seguito vengono esposte le risultanze dello studio svolto sui dati a disposizione. Il database è costituito da un panel bilanciato relativo alle 20 regioni italiane, composto da 2 variabili dipendenti, irr e irr1, che rappresentano le due accezioni del fenomeno oggetto di studio, dato dal tasso di irregolarità del lavoro. La prima accezione è disponibile solo per l'arco temporale 2001-2012, mentre la seconda per il periodo 2001-2015. Le variabili esplicative a disposizione sono 16 (Tabella 1), tutte fruibili per l'arco temporale 2001-2015, ad esclusione di istr, tasso di partecipazione all'istruzione secondaria superiore, che è disponibile solo per il periodo 2001-2012.

Si procede all'analisi attraverso i seguenti passi[21]:

A. analisi descrittiva e grafica della variabile dipendente;
B. definizione di un modello guideline;
C. analisi della correlazione e dell'indice VIF per la verifica di eventuali multicollinearità;
D. confronto dei diversi modelli panel analizzati e scelta del modello più idoneo;
E. affinamento del modello scelto;
F. test per la diagnosi del modello (omoschedasticità, correlazione e normalità).

A. Analisi descrittiva e grafica delle singole variabili

Dall'esame delle statistiche descrittive[22], emerge che il tasso di irregolarità del lavoro, nelle due accezioni e per l'arco temporale 2001-2012, varia molto di più tra le regioni che non nel tempo. La deviazione standard è dovuta, infatti, più alla variabilità tra regioni che alla variabilità nel tempo. Ciò fa presupporre la presenza di una certa eterogeneità tra le regioni, che giustifica l'analisi con stimatori panel rispetto alla semplice modellizzazione pooled OLS. Questa affermazione verrà successivamente sottoposta a test di verifica. Tale caratteristica è presente anche per le altre variabili esplicative ad eccezione di istr1 e tax[23]. In questi due casi c'è maggiore variabilità nel corso degli anni.

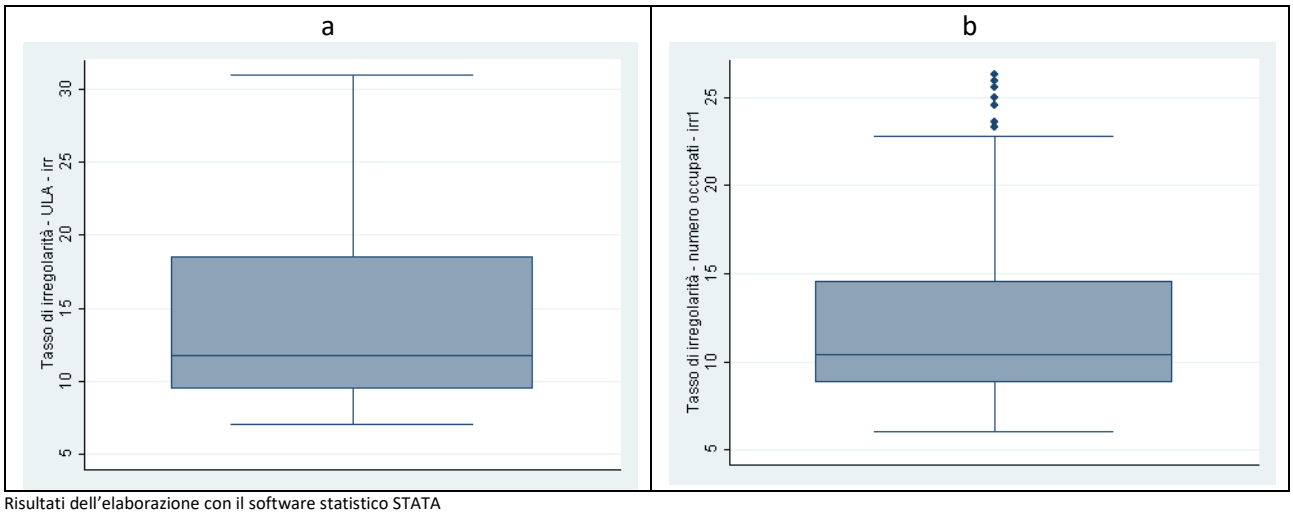
[21]Tutte le elaborazioni sono state sviluppate con il software statistico di elaborazione dati STATA.

[22]Si veda la Tabella A3 e la Tabella A4 in appendice.

[23]Per la variabile tax questo è dovuto al fatto che è individual invariant per costruzione.

L'analisi grafica della variabile risposta, effettuata attraverso i box plot (Figura 4), nelle due accezioni irr e irr1 mostra che entrambe le distribuzioni sono asimmetriche positive con valore degli indici di asimmetria[24] pari, rispettivamente a 0,88 e 1,14. La distribuzione relativa a irr1 presenta alcuni valori anomali e in entrambi i casi si evince come ci sia una maggiore frequenza dei valori medio-bassi e un conseguente spostamento verso il basso della scatola. Tali caratteristiche vengono esplorate attraverso una analisi grafica per anno, per regione e sopra l'ottantesimo percentile, per entrambe le accezioni[25], che conferma una prevalente asimmetria positiva, la presenza di alcuni valori anomali, e la maggiore variabilità del fenomeno per regione.

Figura 4: Rappresentazioni grafiche (box plot) della variabile dipendente (tasso di irregolarità del lavoro) nelle due accezioni irr(a) e irr1(b)



Risultati dell'elaborazione con il software statistico STATA

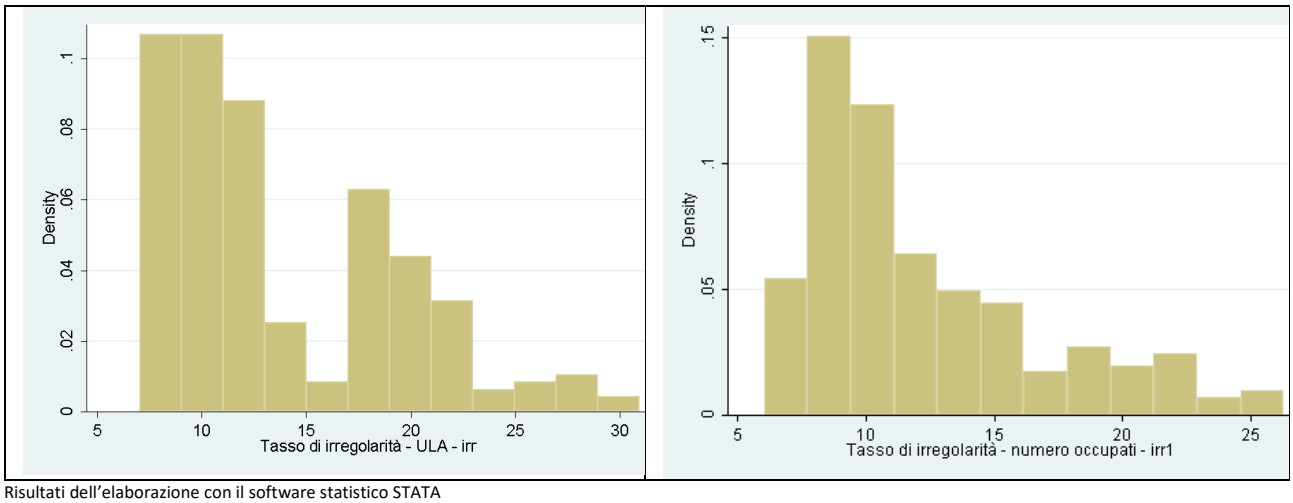
Le variabili dipendenti oggetto di studio non si distribuiscono normalmente (Figura 5). La variabile irr mostra una tendenza alla bimodalità, mentre la variabile irr1 ha una distribuzione asimmetrica. Tali particolarità vengono investigate attraverso una analisi grafica con istogrammi per entrambe gli indicatori, per livello regionale e per anno[26]. La tendenza alla bimodalità del tasso di irregolarità calcolato in termini di unità di lavoro (irr) viene confermato solo per alcuni anni (2006, 2009, 2010, 2011, 2012). Tutti i grafici avvalorano l'asimmetria della distribuzione del tasso di irregolarità, calcolato attraverso il numero di occupati (irr1). Si conferma quanto già detto in precedenza, ossia entrambi gli indicatori variano poco nel tempo sia al livello regionale che nazionale, pertanto la variabilità dipende più da una variabilità regionale che temporale.

[24] L'indice di asimmetria utilizzato è $m_3.m_2^{-3/2}$.

[25] Si veda la Figura A3, A4 e A5 in appendice.

[26] Si veda la Figura A6, A7, A8 e A9 in appendice.

Figura 5: Istogrammi del tasso di irregolarità del lavoro nelle due accezioni scelte (irr e irr1)



Risultati dell'elaborazione con il software statistico STATA

B. Definizione del modello guideline

Si procede inizialmente alla stima del modello pooled OLS, inserendo tutte le variabili a disposizione. Come anticipato nel Capitolo 2, il modello pooled OLS, nella maggior parte dei casi, è improbabile che sia adeguato, ma offre una guideline per un confronto con modelli più complessi.

Vengono studiati tre modelli: due per il periodo 2001-2012 con variabile risposta irr e irr1 e tutte le variabili esplicative a disposizione per lo stesso arco temporale, e uno per il periodo 2001-2015 con variabile risposta irr1 e tutte le variabili esplicative a disposizione per lo stesso arco temporale.

Indicando con:

Y_{kit} la variabile dipendente, dove con K=1 viene definito l'indicatore irr e con k=2 l'indicatore irr1 nella regione i e nell'anno t;

D_t le dummy temporali;

α_i l'effetto individuale;

ε_{it} l'errore residuo;

X_{hit} le p covariate (h=1,…,p) nella regione i e nell'anno t, così definite:

Variabile	Formalizzazione della covariata
Densità di popolazione	X_{1it}
Istruzione di secondo livello	X_{2it}
Istruzione di terzo livello	X_{3it}
Industrializzazione	X_{4it}
Tassazione	X_{5it}
Imprenditorialità femminile	X_{6it}
Imprenditorialità giovanile	X_{7it}
Natalità delle imprese	X_{8it}
Occupazione femminile	X_{9it}
Pil pro capite	X_{10it}
Dipendenti pubblici	X_{11it}

Disoccupazione giovanile	X_{12it}
Percezione della criminalità	X_{13it}
Furti	X_{14it}
Rapine	X_{15it}
Omicidi	X_{16it}

La formalizzazione dei modelli è data dalle seguenti espressioni:

Modello 1: $Y_{1it}=\Sigma_h\beta_h X_{hit}+D_t+\alpha_i+\varepsilon_{it}$ i=1,…..,20; t=1,….,12; h=1,…,16

Modello2: $Y_{2it}=\Sigma_h\beta_h X_{hit}+D_t+\alpha_i+\varepsilon_{it}$ i=1,…..,20; t=1,….,12; h=1,…,16

Modello 3: $Y_{2it}=\Sigma_h\beta_h X_{hit}+D_t+\alpha_i+\varepsilon_{it}$ i=1,…..,20; t=1,….,15; h=1,…,15

C. Analisi della multicollinearità

L'analisi della multicollinearità viene eseguita sia attraverso lo studio della matrice di correlazione che attraverso il fattore di inflazione della varianza (VIF - Variance Inflation Factor) per ciascuna variabile esplicativa. La formula è definita come $VIF_j= 1/1-R^2_j$ dove con j si indica la variabile esplicativa e R^2_j è l'R^2 della previsione della variabile x_j usando le altre variabili esplicative. Il VIF fornisce l'aumento proporzionale nella varianza di $\hat{\beta}$, rispetto a quello che sarebbe stata se le variabili esplicative fossero completamente incorrelate. Se il VIF è grande significa che ci potrebbe essere un problema di collinearità. Un buon riferimento per il livello soglia che segnala un problema di multicollinearità è VIF< max $(10,1/1-R^2_{modello})$. Sotto questa condizione le variabili esplicative sono maggiormente legate alla variabile risposta e non alle altre esplicative, o non sono particolarmente legate tra loro. In queste circostanze le stime dei coefficienti non saranno particolarmetne instabili, quindi la collinearità non sembra essere un problema. Questo indicatore suggerisce di continuare l'analisi escludendo il tasso di occupazione femminile (fem pari al rapporto tra il numero di donne occupate in età 15-64 anni sulla popolazione) per il Modello 1 e anche il Prodotto Interno Lordo ai prezzi di mercato, prezzi correnti, per abitante (pilpc)[27] per i Modelli 2 e 3.

D.Confronto dei diversi modelli panel analizzati e scelta del modello più idoneo

Si procede alla stima dei cinque modelli panel (POLS, FE, RE, BE e FD) discussi nel Capitolo precedente e alla scelta del modello tramite gli strumenti di seguito descritti.

La stima POLS non tiene conto della presenza degli effetti individuali, pertanto nel caso tali effetti risultassero correlati con i regressori le stime potrebbero essere inconsistenti. Infatti, gli effetti individuali, se esistono, sono nei residui. La scelta tra la stima FE e POLS si effettua attraverso l'F-test[28] per gli effetti fissi. Tale test verifica l'ipotesi nulla che gli effetti fissi siano assenti. Pertanto, nel caso in cui l'ipotesi nulla sia rifiutata, così come nel nostro caso[29], si afferma che l'effetto fisso contribuisce in modo significativo alla qualità di adattamento del modello, in questo caso il modello FE è migliore del modello POLS.

La scelta tra il modello RE e il modello POLS si realizza attraverso il test di Breusch e Pagan, che verifica l'ipotesi nulla di assenza del fattore individuale, ossia che la varianza degli effetti individuali sia costante. Se

[27]Si veda Tabella A5 in appendice.

[28]La statistica F è il quadrato della statistica t.

[29]Si veda Tabella A16 in appendice

l'ipotesi nulla è rifiutata, come nel nostro caso[30], si può concludere che gli effetti individuali sono presenti e vanno considerati. Pertanto, il modello POLS non appare opportuno.

I test esaminati fino ad ora hanno portato ad affermare che i modelli FE e RE sono in grado di tener conto degli effetti individuali non osservati, mentre il modello POLS no. Questa conclusione era stata già anticipata dall'analisi descrittiva delle variabili[31], quando era stato osservato che la deviazione standard della maggior parte delle variabili era dovuta più alla variabilità between regioni che alla variabilità within regione, facendo presupporre la presenza di una certa eterogeneità tra le regioni che avrebbe giustificato l'analisi con modelli panel rispetto alla semplice modellizzazione POLS.

Infine, per la scelta tra il modello FE e il modello RE utilizziamo il test di Hausman. L'ipotesi nulla vuole verificare che non vi sia correlazione tra α_i e x_{it}. Infatti, in presenza di correlazione non nulla tra α_i e x_{it} le stime RE non sono consistenti, mentre quelle FE continuano ad esserlo. Dunque, una differenza statisticamente significativa fra le stime FE e quelle RE viene interpretata come evidenza contro gli stimatori RE. Se invece l'ipotesi di non correlazione fra le esplicative e gli effetti individuali è valida ($E(\alpha_i|x_{1t}, x_{2t},\ldots., x_{iT}) = E(\alpha_i) = 0$), RE produce stime consistenti ed è più efficiente di FE. In tutti i casi in esame si rifiuta l'ipotesi nulla, pertanto il test indica di scegliere il modello FE. Ma la matrice delle differenze delle varianze dei coefficienti è definita non positiva, pertanto non è invertibile.

Ricordiamo, infatti, che il risultato cruciale di Hausman è mostrare che la covarianza fra uno stimatore efficiente e la sua differenza con uno inefficiente è zero: $\text{cov}(\hat{\beta}_{RE}, (\hat{\beta}_{FE} - \hat{\beta}_{RE}) = \text{cov}(\hat{\beta}_{RE}, \hat{\beta}_{FE}) - \text{var}(\hat{\beta}_{RE}) = 0$, da cui sotto l'ipotesi di stretta esogeneità, rango pieno e omoschedasticità dei residui si ottiene

$$\text{var}(\hat{\beta}_{FE} - \hat{\beta}_{RE}) = \text{var}(\hat{\beta}_{FE}) + \text{var}(\hat{\beta}_{RE}) - 2\text{cov}(\hat{\beta}_{FE}, \hat{\beta}_{RE}) = \text{var}(\hat{\beta}_{FE}) - \text{var}(\hat{\beta}_{RE})$$

che è definita positiva. La statistica di Hausman è data dalla seguente espressione

$$W = (\hat{\beta}_{FE} - \hat{\beta}_{RE})'[\text{var}(\hat{\beta}_{FE} - \hat{\beta}_{RE})]^{-1} (\hat{\beta}_{FE} - \hat{\beta}_{RE}),$$

pertanto se la matrice $\text{var}(\hat{\beta}_{FE} - \hat{\beta}_{RE})$ non è definita positiva, non è invertibile e quindi non è possibile trarre nessuna conclusione definitiva.

Per tale motivo consideriamo il test di Wooldridge, che è un test più robusto in quanto tiene conto della correlazione seriale nel tempo e dell'eteroschedasticità generale. Il test di Wooldridge indica che le variabili calcolate come scostamento delle variabili che variano nel tempo dalle medie individuali non sono congiuntamente uguali a zero (si rifiuta H0), e questo ci porta definitivamente a rigettare il modello a effetti casuali.

Conseguenza del risultato appena trovato, è che il modello BE non sia consistente. Infatti, lo stimatore RE è una media pesata dello stimatore BE e dello stimatore FE. Ma il test di Hausman e di Wooldridge hanno ritenuto lo stimatore FE consistente[32], pertanto la non consistenza del modello RE influisce anche sui risultati del modello BE.

Il modello FE è quindi risultato il modello più idoneo allo studio del fenomeno in esame. Questa scelta è coerente con la letteratura esistente (Baltagi, 2005; Lisi, 2009; Lisi, 2010), in base alla quale il modello a effetti

[30]Si veda Tabella A16 in appendice

[31]Tabella A3 e A4 in appendice.

[32]Si veda Tabella A17 in appendice.

fissi è un'adeguata specificazione se l'analisi è focalizzata su un insieme specificato di N unità e l'inferenza è indirizzata al comportamento nel tempo delle medesime unità. Viceversa, il modello a effetti casuali è la specificazione più adeguata se le N unità osservate sono casualmente selezionate dalla popolazione e si vuole inferire dal campione all'intera popolazione.

Sono state eseguite una serie di elaborazioni riferite al periodo 2001-2012, con variabile dipendente irr e irr1, e al periodo 2001-2015, con variabile dipendente irr1. Si è proceduto al raffinamento del modello FE inserendo anche le variabili dummy temporali, per catturare gli effetti fissi time series, e utilizzando i metodi backward elimination e forward selection, per selezionare l'insieme di predittori che abbia la migliore relazione con la variabile dipendente. Per i modelli così ottenuti, è stata eseguita l'analisi dei residui che sono stati sottoposti a test di verifica per la normalità dei residui, per la correlazione e per l'omoschedasticità. Poiché quest'ultima proprietà non è mai stata verificata, sono state utilizzate varie versioni robuste della stima della matrice di varianza e covarianza: *robust*, che stima in maniera consistente la matrice di varianza e covarianza perché per ogni individuo considera l'autocorrelazione delle osservazioni relative ad ogni individuo, e stimatori basati su ricampionamento (*bootstrap* e *jackknife*). Inoltre, l'analisi dei residui ha fatto emergere una correlazione dei residui di primo ordine, pertanto tutti i modelli sono stati sottoposti a una trasformazione dei dati che rimuove la componente AR(1)[33].

D1: Robust, Bootstrap e Jackknife

Robust è la procedura del software statistico STATA che implementa alcuni stimatori di varianza robusti, in particolare gli stimatori di Huber, di White o gli stimatori sandwich, che sono appunto stimatori usati per contrastare gli effetti dell'eteroschedasticità (Friedman, 2006; Soliani, 2005; Cameron e Trivedi, 2010).

Consideriamo la regressione dei minimi quadrati ordinaria. Lo stimatore per i coefficienti è:

$$\hat{\beta}=(X'X)^{-1}X'y$$

dove y è un vettore nx1 che rappresenta la variabile dipendente e X è una matrice di covariate nxk. La varianza di $\hat{\beta}$ sarà pari a $V(\hat{\beta})= (X'X)^{-1}V(X'y)(X'X)^{-1}$. Lo stimatore robusto della varianza per le stime del coefficiente di regressione lineare nel caso di osservazioni indipendenti sarà dato da:

$$\hat{V}(\hat{\beta})= (X'X)^{-1}\hat{V}(X'y)(X'X)^{-1} \text{ dove } \hat{V}(X'y)=\Sigma\hat{e}^2_j x_j' x_j$$

chiamato **stimatore sandwich**.

Lo stimatore Huber Sandwich (Friedman, 2006) può essere utilizzato per stimare la varianza nel caso di stimatori di massima verosimiglianza, quando il modello non è corretto, ossia quando Θ_0 non è il vero valore di Θ.

Sia i l'indice delle osservazioni con valori pari a y_i e $\Theta \epsilon R^p$ un parametro vettore px1. Sia inoltre $y \sim f_i(y/\Theta)$ una densità positiva tale che $f_i(0/\Theta)>0, f_i(1/\Theta)>0$ e $f_i(0/\Theta)+f_i(1/\Theta)=1$. La funzione di verosimiglianza dei valori osservati Y_i è data da $\Pi f_i(y_i/\Theta)$ e la funzione di log verosimiglianza è $L(\Theta)=\Sigma \log f_i(y_i/\Theta)$. Le derivate prima e seconda di L rispetto a Θ sono $L'(\Theta)=\Sigma g_i(Y_i/\Theta)$ e $L''(\Theta)=\Sigma h_i(Y_i/\Theta)$. Supponiamo che il modello sia corretto e che Θ_0 sia il vero valore di Θ. Allora la funzione di log verosimiglianza può essere scritta come una serie di Taylor intorno a Θ_0. Ignorando i termini di ordine più alto, la funzione di verosimiglianza sarà essenzialmente

[33] Viene utilizzata una trasformazione di Cochrane-Orcutt

quadratica, e il massimo potrà essere trovato risolvendo l'equazione $L'(\Theta)=0$. Pertanto l'equazione sarà $L'(\Theta_0) + (\Theta-\Theta_0)^T L''(\Theta_0)=0$ quindi $\hat{\Theta}-\Theta_0=[-L''(\Theta_0)]^{-1}L'(\Theta_0)^T$. Allora

$$cov_{\Theta 0}\hat{\Theta}= [-L''(\Theta_0)]^{-1}cov_{\Theta 0}L'(\Theta_0)[-L''(\Theta_0)]^{-1} \quad (1)$$

L'idea sandwich è quella di stimare $L''(\Theta_0)$ direttamente dai dati del campione come $L''(\hat{\Theta})$ e allo stesso modo $cov_{\Theta 0}L'(\Theta_0)$ come $\Sigma g_i(Y_i/\hat{\Theta})^T g_i(Y_i/\hat{\Theta})$. Pertanto, la (1) può essere così stimata:

$$\hat{V} =(-A)^{-1}B(-A)^{-1}$$

dove $A=L''(\hat{\Theta})$ e $B=\Sigma g_i(Y_i/\hat{\Theta})^T g_i(Y_i/\hat{\Theta})$.

$\hat{V}$ viene chiamato lo stimatore Huber sandwich, e la radice quadrata degli elementi della diagonale di $\hat{V}$ sono gli errori standard robusti o gli errori standard Huber-White.

Le procedure di riuso del campione, ed in particolare le metodologie che vanno sotto il nome di **Bootstrap e Jackknife** (Soliani, 2005), hanno assunto nei problemi di inferenza un ruolo sempre più rilevante come vie alternative a quella analitica classica. Una caratteristica specifica su cui poggiano queste tecniche di ricampionamento è la simulazione con metodi Monte Carlo di una procedura statistica, utilizzando il minor numero possibile di assunzioni a priori. In molte situazioni è difficile o impossibile ottenere la distribuzione di probabilità di una statistica, altre volte l'approssimazione asintotica non è soddisfacente per piccoli campioni. In questi casi si può tentare di stimare la distribuzione di una statistica con metodi di simulazione Monte Carlo basati sul ricampionamento da $X=(x_1 ,...,x_n)$. Le procedure Bootstrap e Jackknife sono di questo tipo.

La procedura Jackknife (chiamata anche Tukey's jackknife) serve per ridurre le distorsioni sistematiche, che dipendono dai dati campionari, nella stima delle statistiche di una popolazione, fornendone l'errore standard. Permette quindi di calcolare l'intervallo di confidenza per la statistica in esame. Il termine jackknife in inglese indica il coltello a serramanico. L'idea di base del metodo jackknife, come proposta da Tukey nel 1958 sviluppando l'idea proposta da Quenouille nel 1949, serve anche per costruire intervalli di confidenza intorno alla media.

La metodologia può essere schematizzata nei seguenti passaggi.

- Calcolare la statistica S_t desiderata utilizzando le N osservazioni del campione raccolto.
- Dividere il campione in sottogruppi; se il campione è di grandi dimensioni, i sottogruppi sono formati da k unità; se il campione è di piccole dimensioni, come spesso succede, i sottogruppi possono essere formati da una sola unità.
- Calcolare il valore della statistica desiderata senza un sottogruppo, ignorando ogni volta un sottogruppo diverso S_{t-i}; si ottengono N/k differenti stime della statistica.
- Calcolare i cosiddetti pseudo valori θ_i (chiamati in questo modo perché cercano mediamente di stimare il parametro θ riproducendo le variabili originarie) per ogni stima di S_{t-i}, mediante la differenza $\theta_i = N \cdot S_t – (N – 1) \cdot S_{t-i}$
- La stima con la procedura jackknife della statistica in oggetto $\hat{S}_t$, è semplicemente la media aritmetica θ di questi valori θ_i
- Con il valore della t di Student alla probabilità α prescelta e per gdl (gradi di libertà) N-1, si stimano i limiti di confidenza $\hat{S}_t \pm t_{(\alpha/2,\, N-1)} \cdot es_{(St)}$ entro i quali, alla probabilità α prefissata, si troverà il parametro della popolazione.

Questa procedura può essere applicata a varie analisi statistiche, delle quali vengono ricordate quelle che ricorrono con frequenza maggiore nella ricerca applicata alle discipline biologiche e ambientali.

L'idea alla base della forma più semplice della procedura jackknife, implementata nel software statistico STATA, è quella di calcolare ripetutamente la statistica in questione, ogni volta omettendo solo una delle osservazioni del set di dati.

La procedura Bootstrap è stata proposta da Bradley Efron nel 1979, come evoluzione del metodo Jackknife. In pochi anni, tale procedura ha avuto una evoluzione rapida a tal punto da renderla la tecnica di ricampionamento più nota e diffusa. Il nome Bootstrap, letteralmente stringhe o lacci da scarpe, deriva dall'espressione inglese "to pull oneself up by one's bootstrap"[34], che significa "tirarsi su attaccandosi ai lacci delle proprie scarpe". Evidenzia, in modo scherzoso, il fatto paradossale che l'unico campione disponibile serve per generarne molti altri e per costruire la distribuzione teorica di riferimento. Per meglio comprendere i concetti in modo operativo, si possono definire i passaggi fondamentali richiesti dalla metodologia.

- A partire dal campione osservato (x_1, x_2, …, x_k), si costruisce una popolazione fittizia, ripetendo n volte ognuno dei k dati; oppure si estrae un dato alla volta, reinserendolo immediatamente, in modo che la probabilità di estrazione di ogni valore sia sempre costante ed uguale per tutti.
- Si estrae un campione casuale, chiamato campione bootstrap (*bootstrap sample*), estraendo k dati; nella striscia possono essere presenti una o più repliche dello stesso dato e quindi mancare un numero corrispondente di valori presenti nel campione originale. Ad esempio, con k = 10, è possibile avere la striscia o campione bootstrap x_1, x_3, x_3, x_4, x_5, x_6, x_6, x_6, x_7, x_{10} in cui sono assenti i valori x_2, x_8 e x_9, mentre x_3 è ripetuto due volte e x_6 tre volte.
- Per ciascuno di tali campioni bootstrap si calcola lo stimatore θ ottenendo una replica bootstrap (*bootstrap replication*).
- Con n estrazioni o n campioni di repliche casuali (*bootstrap samples*), si ottiene la successione di stime o repliche bootstrap (*bootstrap replications*), che sono la realizzazione della variabile casuale "stimatore bootstrap T".
- La funzione di ripartizione empirica degli n valori θ ottenuti fornisce una stima accurata delle caratteristiche della variabile casuale T.
- L'approssimazione è tanto più precisa quanto più n è elevato.
- Infine, dalla serie dei valori θ, oltre alla media è possibile ottenere:
    - la stima della distorsione,
    - la stima dell'errore standard,
    - l'intervallo di confidenza.

Cameron e Trivedi (2010) discutono molti argomenti riguardo il bootstrap e dimostrano come eseguirli nel software STATA. La logica alla base del bootstrap è che tutte le misure di precisione provengono dalla distribuzione del campione di una statistica. Quando la statistica viene stimata su un campione di dimensione N da una popolazione, la distribuzione campionaria indica le frequenze relative dei valori della statistica. La distribuzione campionaria, a sua volta, è determinata dalla distribuzione della popolazione e dalla formula utilizzata per stimare la statistica. L'accuratezza con cui la distribuzione bootstrap stima la distribuzione campionaria dipende dal numero di osservazioni nel campione originale e dal numero di repliche nel

[34]L'espressione è tratta dal romanzo del diciottesimo secolo "Adventures of Baron Munchausen" di Rudolph Erich Raspe.

bootstrap. In generale, le repliche dell'ordine di 1.000 producono stime molto buone, ma sono necessarie solo 50-200 repliche per le stime degli errori standard.

D2: Trasformazione di Cochrane-Orcutt (1949)

Consideriamo il semplice modello di regressione lineare con errori autocorrelati del primo ordine. Supponiamo di trasformare la variabile risposta in modo che:

$$Y^*_t = Y_t - \rho Y_{t-1}$$

dove ρ è un parametro sconosciuto.

Sostituendo Y_t e Y_{t-1} si ottiene:

$$Y^*_t = \beta^*_0 + \beta^*_t X^*_t + a_t$$

dove

$$\beta^*_0 = \beta_0(1-\rho);\ \beta^*_t = \beta_t;\ X^*_t = X_t - \rho X_{t-1};\ a_t = \varepsilon_t - \rho \varepsilon_{t-1}$$

I termini di errore, nel modello riparametrizzato, sono variabili casuali indipendenti. Pertanto, trasformando il regressore e le variabili risposta, viene prodotto un modello che soddisfa le usuali ipotesi di regressione e possono quindi essere utilizzati nei minimi quadrati ordinari. Il modello riparametrizzato non può essere utilizzato direttamente, perché le nuove variabili di regressione e risposta X^*_t e Y^*_t sono funzioni del parametro sconosciuto ρ. Tuttavia, il processo autoregressivo del primo ordine ($e_t = \rho e_{t-1} + a_t$) può essere visto come una regressione attraverso l'origine. Quindi ρ può essere stimato ottenendo i residui e_t da una normale regressione dei minimi quadrati di Y_t su X_t, e quindi regredendo e_t su e_{t-1}. La stima dei minimi quadrati di ρ è:

$$\hat{\rho} = \sum_{t=2}^{T}(e_t e_{t-1}) / \sum_{t=2}^{T}(e^2_{t-1})$$

Usando questa stima di ρ, si ottengono il regressore e le variabili risposta trasformate. Il metodo dei minimi quadrati si applica ai dati trasformati. Viene utilizzato il test di Durbin-Watson ai residui del modello riparametrizzato. Se questa procedura indica che i residui non sono correlati, non è necessaria alcuna analisi aggiuntiva. Viceversa, se è ancora indicata l'autocorrelazione positiva, è necessaria un'altra iterazione. Nella seconda iterazione, si stima ρ con i nuovi residui ottenuti utilizzando i coefficienti di regressione del modello riparametrizzato con il regressore originale e le variabili di risposta. Se necessario, tale procedura iterativa può essere continuata fino a quando i termini di errore nel modello riparametrizzato non sono correlati.

La trasformazione di Cochrane-Orcutt, nella sua semplicità, è risolutiva sia dei problemi di stazionarietà o correlazione, che di autocorrelazione seriale. Il problema verrà definitivamente sorpassato con i modelli panel dinamici (Capitolo 5).

E e F. Affinamento del modello scelto e test per la diagnosi del modello (omoschedasticità, correlazione e normalità)

Vengono di seguito esplicitati i migliori tre modelli trovati:

- MODELLO CON VARIABILE DIPENDENTE IRR – Periodo 2001-2012
- MODELLO CON VARIABILE DIPENDENTE IRR1 – Periodo 2001-2012

- MODELLO CON VARIABILE DIPENDENTE IRR1 – Periodo 2001-2015

MODELLO CON VARIABILE DIPENDENTE IRR - Periodo 2001-2012

Per la variabile dipendente irr sono stati individuati due modelli, che sono stati valutati attraverso gli indici di bontà di adattamento del modello AIC e BIC[35] e la capacità predittiva tramite l'indicatore RMSE[36]:

Tabella 2: Valutazione dei modelli Fixed Effect (FE) con variabile dipendente irr per il periodo 2001-2012 tramite gli indicatori AIC, BIC e RMSE

Nome Modello	Modello	AIC	BIC	RMSE
FE1 (irr)	$Y_{1it} = \beta_1 x_{1it} + \beta_5 x_{5it} - \beta_6 x_{6it} + \beta_7 x_{7it} + \beta_{10} x_{10it} + \beta_{11} x_{11it} + \beta_{12} x_{12it} + \beta_{13} x_{13it} + \beta_{14} x_{14it} + \beta_{15} x_{15it} + \beta_{16} x_{16it} + D_t + \alpha_i + \varepsilon_{it}$	**621.69**	**698.26**	**0.88**
FE2 (irr)	$Y_{1it} = \beta_1 x_{1it} + \beta_4 x_{4it} + \beta_6 x_{6it} + \beta_8 x_{8it} + \beta_{11} x_{11it} + \beta_{13} x_{13it} + D_t + \alpha_i + \varepsilon_{it}$	**650.84**	**713.50**	**0.95**

Il modello FE1(irr) ha valori degli indicatori AIC, BIC e RMSE più bassi rispetto al modello FE2(irr), pertanto il modello FE1(irr) è il modello migliore. Il modello **FE1(irr)** è stato ottenuto con il metodo backward elimination.

MODELLO CON VARIABILE DIPENDENTE IRR1 - Periodo 2001-2012

Per la variabile irr1, nel periodo 2001-2012, sono stati individuati quattro modelli, che sono stati valutati con gli indici di bontà di adattamento del modello AIC e BIC e la capacità predittiva tramite l'indicatore RMSE:

[35] Akaike Information Criterion (AIC) e Bayesian Information Criterion (BIC). Entrambi gli indici sono ricavati partendo dal logaritmo della varianza residua pesata per il numero delle osservazioni, e penalizzando i modelli con parametri aggiuntivi. AIC e BIC differiscono nella funzione di penalizzazione, dato che il BIC impone una penalità maggiore ai modelli con più parametri, e sono molto utili per la comparazione fra modelli.

[36] RMSE (Root Mean Square Error) è l'indicatore che esprime la radice dell'errore quadratico medio.

Tabella 3: Valutazione dei modelli Fixed Effect (FE) con variabile dipendente irr1 e per il periodo 2001-2012 tramite gli indicatori AIC, BIC e RMSE

Nome Modello	Modello	AIC	BIC	RMSE
FE1 (irr1)	$Y_{2it} = \beta_1 x_{1it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_5 x_{5it} + \beta_8 x_{8it} + \beta_{11} x_{11it} + \beta_{12} x_{12it} + \beta_{15} x_{15it} + \beta_{16} x_{16it} + \alpha_i + \varepsilon_{it}$	571.57	606.38	0.81
FE2(irr1)	$Y_{2it} = \beta_1 x_{1it} + \beta_3 x_{3it} + \beta_5 x_{5it} + \beta_8 x_{8it} + \beta_{11} x_{11it} + \beta_{12} x_{12it} + \beta_{13} x_{13it} + \alpha_i + \varepsilon_{it}$	582.78	610.62	0.84
FE3 (irr1)	$Y_{2it} = \beta_1 x_{1it} + \beta_3 x_{3it} + \beta_5 x_{5it} + \beta_{11} x_{11it} + \beta_{12} x_{12it} + \beta_{13} x_{13it} + \beta_{15} x_{15it} + \beta_{16} x_{16it} + D_t + \alpha_i + \varepsilon_{it}$	557.42	623.55	0.78
FE4 (irr1)	$Y_{2it} = \beta_1 x_{1it} + \beta_3 x_{3it} + \beta_5 x_{5it} + \beta_{12} x_{12it} + \beta_{13} x_{13it} + D_t + \alpha_i + \varepsilon_{it}$	574.15	629.89	0.81

Il modello FE3(irr1) ha valori degli indicatori AIC e RMSE più bassi rispetto agli altri tre modelli. Il valore dell'indicatore BIC per il modello FE3(irr1) varia di poco rispetto agli altri modelli, pertanto il modello migliore è il modello FE3(irr1). Il modello **FE3(irr1)** è stato ottenuto con il metodo backward elimination.

I residui dei modelli appena esaminati, si distribuiscono normalmente sia secondo la rappresentazione grafica che secondo i test utilizzati[37]. Essi risultano essere eteroschedastici, pertanto sono state utilizzate le procedure robust, jackknife e bootstrap[38]. Inoltre, l'analisi dei residui, ha fatto emergere la presenza di autocorrelazione di primo ordine[39], che è stata opportunamente risolta attraverso la trasformazione dei dati di Cochrane-Orcutt (metodo esposto nel paragrafo D2) [40] . Il problema congiunto eteroschedasticità/correlazione dei residui è stato superato applicando alla procedura robust, l'opzione cluster in quanto il raggruppamento produce uno stimatore coerente quando vi è una correlazione seriale dei residui.

[37]Figura A10 e Tabella A6; Figura A11 e Tabella A7 in appendice.

[38]Tabella A11 e A12 in appendice.

[39] I risultati sono esposti nell'appendice del Capitolo 5.

[40] Tabella A13 e A14 in appendice.

Risultati dei modelli FE1(irr) e FE3(irr1)

Tabella 4: Coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE1(irr)

Variabile	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Densità di popolazione	-0.1726881	0.0206675	-8.36	0.000	-0.213435 -0.1319327
Tassazione	1.828424	0.347616	5.26	0.000	1.14294 2.513908
Imprenditorialità femminile	-0.3848288	0.132116	-2.91	0.004	-0.6453558 -0.1243017
Imprenditorialità giovanile	-0.9502937	0.222758	-4.27	0.000	-1.389563 -0.5110246
Pil pro capite	0.3527608	0.1797479	1.96	0.051	-0.0016943 0.7072159
Dipendenti pubblici	-0.1486721	0.0666317	-2.23	0.027	-0.280067 -0.0172773
Disoccupazione giovanile	0.0379435	0.0171077	2.10	0.037	0.0022359 0.0736512
Percezione della criminalità	0.0424319	0.0212581	2.00	0.047	-0.0005118 0.084352
Furti	0.0823281	0.0340697	2.42	0.017	0.0151442 0.149512
Rapine	0.9702864	0.5613383	1.73	0.085	-0.1366484 2.077221
Omicidi	-0.3410333	0.1692493	-2.01	0.045	-0.6747856 0.0072811
Dummy2002	1.229681	0.3752887	3.28	0.001	0.4896279 1.969734
Dummy2003	0.2303791	0.3280873	0.70	0.483	-0.4165949 0.8773531
Dummy2004	-0.8020854	0.2575837	-3.11	0.002	-1.310029 -0.2941415
Dummy2005	Variabile omessa				
Dummy2006	-2.688425	0.5955098	-4.51	0.000	-3.862744 -1.514105
Dummy2007	-3.798799	0.8004707	-4.75	0.000	-5.377293 -2.220306
Dummy2008	-3.482437	0.8408884	-4.14	0.000	-5.140632 -1.824241
Dummy2009	-3.236864	0.8545415	-3.79	0.000	-4.921983 -1.551745
Dummy2010	-1.64368	0.6557825	-2.51	0.013	-2.936854 -0.350505
Dummy2011	-1.72572	0.6772911	-2.55	0.012	-3.061309 -0.3901313
Dummy2012	-3.305763	0.9244877	-3.58	0.000	-5.128812 -1.482713
Cons	6.568574	13.00361	0.51	0.614	-19.07399 32.21114

Risultati dell'elaborazione con il software statistico STATA

Tabella 5: Coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE3(irr1)

Variabile	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Densità di popolazione	-0.1979079	0.0176204	-11.23	0.000	-0.2326514 -0.1631645
Istruzione di terzo livello	0.3120811	0.0737892	4.23	0.000	0.1665853 0.4575769
Tassazione	1.003738	0.2128593	4.72	0.000	0.5840264 1.423449
Dipendent ipubblici	-0.0885674	0.0531436	-1.67	0.097	-0.1933547 0.0162199
Disoccupazione giovanile	0.0618856	0.0157932	3.92	0.000	0.030745 0.0930262
Percezione della criminalità	0.0330629	0.0184903	1.79	0.075	-0.0033957 0.0695216
Rapine	1.926509	0.4807452	4.01	0.000	0.9785864 2.874431
Omicidi	-0.3245098	0.1510128	-2.15	0.033	-0.6222734 -0.0267462
Dummy2002	0.4732817	0.2711679	1.75	0.082	-0.0614009 1.007964
Dummy2003	-0.1363701	0.2639559	-0.52	0.606	-0.6568324 0.3840921
Dummy2004	-0.6063633	0.2226309	-2.72	0.007	-1.045342 -0.1673848
Dummy2005	Variabileomessa				
Dummy2006	-1.079123	0.3246359	-3.32	0.001	-1.719232 -0.4390131
Dummy2007	-1.497433	0.3935671	-3.80	0.000	-2.273459 -0.7214061
Dummy2008	-1.471014	0.4258831	-3.45	0.001	-2.310761 -0.6312677
Dummy2009	-1.381837	0.4807013	-2.87	0.004	-2.329673 -0.4340011
Dummy2010	-0.4185706	0.3779764	-1.11	0.269	-1.163856 0.3267148
Dummy2011	-0.3224366	0.3761798	-0.86	0.392	-1.06418 0.4193063
Dummy2012	-1.590962	0.5827389	-2.73	0.007	-2.739993 -0.4419305
cons	16.77847	7.132778	2.35	0.020	2.714221 30.84272

Risultati dell'elaborazione con il software statistico STATA

La Tabella 4 e la Tabella 5 mostrano i risultati principali dei modelli ritenuti migliori, F1(irr) e F3(irr1), per le due accezioni in esame, irr e irr1, nello stesso arco temporale 2001-2012[41].

I dati relativi al coefficiente di determinazione e all'esistenza della correlazione tra i residui e le variabili esplicative[42] confermano che la scelta del modello fixed effect è quella ottimale. Il test F garantisce che in entrambi i modelli i coefficienti siano differenti da zero[43].

Le variabili dummy riferite all'anno vengono modellizzate secondo una parametrizzazione corner point: una dummy viene considerata nulla e costituisce il termine di riferimento per valutare le restanti dummy. In entrambi i modelli il 2005 viene posto pari a zero, identificando il 2005 come l'intercetta o l'anno base. Nella tabella vengono riportati non i coefficienti, ma la variazione rispetto all'anno base 2005. Dal modello FE1(irr) (Tabella 4) emerge che l'anno 2003 non è significativamente diverso dall'anno base, mentre per il modello FE3(irr1) (Tabella 5), oltre l'anno 2003 anche gli anni 2010 e 2011 non sono significativamente diversi dal 2005.

Il numero dei regressori è lievemente superiore nel modello FE1(irr). Concorrono alla definizione della variabile risposta nelle due accezioni utilizzate, irr e irr1, sette esplicative uguali: la densità di popolazione, la quota delle entrate tributarie, l'indice di regolamentazione, la disoccupazione giovanile, la percezione della

[41] In appendice la Tabella A10 sintetizza i risultati principali dei tre modelli esaminati.

[42] Modello FE1(irr): R-square (within =0.5280 between = 0.0018 overall = 0.0011);corr(u_i, Xb) = -0.9550. Modello FE3(irr1): R-square (within = 0.5651 between = 0.1089 overall = 0.0982); corr(u_i, Xb) = -0.9798.

[43] Modello FE1(irr): F(21,199) = 10.60; Prob> F = 0.0000. Modello FE3(irr1):F(18,202) = 14.58; Prob> F = 0.0000.

criminalità, le rapine denunciate e gli omicidi volontari consumati. Rimangono invece escluse dalla definizione dell'indicatore irr1, presenti invece per irr, il tasso di imprenditorialità femminile e giovanile, che definiscono la struttura economica delle regioni ma la cui presenza viene garantita dalla quota delle entrate tributarie; il prodotto interno lordo pro capite, quale variabile di controllo, il cui ruolo viene assicurato dalla quota dei dipendenti pubblici e dalla disoccupazione giovanile; i furti, variabile appartenente al gruppo sulla criminalità, il quale risulta ampiamente rappresentato dalle restanti variabili esplicative prese in considerazione. Infine, l'unica variabile non presente nel modello FE1(irr) è la variabile che esprime il livello di istruzione.

I coefficienti dei regressori sono tutti significativi e mostrano un segno atteso coerente con la teoria economica, sebbene perdano lievemente di significatività sotto le procedure per la correzione dell'eteroschedasticità e della correlazione.

MODELLO CON VARIABILE DIPENDENTE IRR1 - Periodo 2001-2015

Avendo a disposizione per la variabile irr1 dati aggiornati fino all'anno 2015, si è ritenuto utile allo studio definire un modello che considerasse la variabile dipendente irr1 nel periodo 2001-2015. Sono stati quindi individuati due modelli che hanno ottenuto i seguenti valori per gli indici di bontà di adattamento del modello AIC e BIC e la capacità predittiva tramite l'indicatore RMSE:

Tabella 6: Valutazione dei modelli Fixed Effect (FE) con variabile dipendente irr1 e per il periodo 2001-2015 tramite gli indicatori AIC, BIC e RMSE

Modello	Con tutte le variabili	AIC	BIC	RMSE
FE5(irr1)	$Y_{2it} = \beta_1 X_{1it} + \beta_3 X_{3it} + \beta_5 X_{5it} + \beta_6 X_{6it} + \beta_7 X_{7it} + \beta_8 X_{8it} + \beta_{11} X_{11it} + \beta_{12} X_{12it} + \beta_{14} X_{14it} + \beta_{15} X_{15it} + \alpha_i + \varepsilon_{it}$	789.17	829.91	0.92
FE6(irr1)	$Y_{2it} = \beta_1 X_{1it} + \beta_3 X_{3it} + \beta_5 X_{5it} + \beta_6 X_{6it} + \beta_8 X_{8it} + \beta_{11} X_{11it} + \beta_{12} X_{12it} + \beta_{15} X_{15it} + D_t + \alpha_i + \varepsilon_{it}$	757.11	838.59	0.85

Il criterio AIC e l'indicatore RMSE concordano nello scegliere il modello FE6(irr1), mentre il criterio BIC varia poco tra i due modelli, pertanto il modello migliore è **FE6(irr1)**. I residui si distribuiscono normalmente, sebbene debolmente (p-value>0.01), sia secondo la rappresentazione grafica che secondo i test utilizzati[44]. Essi risultano essere eteroschedastici, pertanto sono state utilizzate le procedure robust, jackknife e bootstrap[45], che hanno confermato i risultati nonostante qualcuno sia lievemente indebolito in termini di significatività. Anche in questo caso, l'analisi dei residui, ha fatto emergere la presenza di correlazione[46]. Si è pertanto proceduto come per i modelli precedenti.

[44]Figura A12 e Tabella A8 in appendice.

[45]Tabella A9 in appendice.

[46] I risultati sono presenti nell'appendice del Capitolo 5.

Tabella 7: Coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE6(irr1)

Variabile	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Densità di popolazione	-0.1326895	0.0123017	-10.79	0.000	-0.1569136 -0.1084655
Istruzione terzo livello	0.2626658	0.0688201	3.82	0.000	0.1271476 0.398184
Tassazione	1.063561	0.2137847	4.97	0.000	0.6425837 1.484539
Imprenditorialità femminile	-0.3133613	0.0894839	-3.50	0.001	-0.48957 -0.1371526
Natalità delle imprese	-0.4137995	0.1650165	-2.51	0.013	-0.7387444 -0.0888546
Dipendenti pubblici	-0.1307712	0.053977	-2.42	0.016	-0.237061 -0.0244815
Disoccupazione giovanile	0.0317894	0.0144974	2.19	0.029	0.0032417 0.0603371
Rapine	2.544598	0.4074637	6.24	0.000	1.742235 3.346961
Dummy2002	0.3158559	0.2857173	1.11	0.270	-0.2467687 0.8784805
Dummy 2003	-0.5036731	0.2884009	-1.75	0.082	-1.071582 -.064236
Dummy2004	-0.8407689	0.2400631	-3.50	0.001	-1.313493 -0.3680447
Dummy2005	Variabileomessa				
Dummy2006	-1.493127	0.3256402	-4.59	0.000	-2.134366 -0.8518871
Dummy2007	-1.414781	0.3935425	-3.59	0.000	-2.189731 -0.6398302
Dummy2008	-1.824629	0.3925646	-4.65	0.000	-2.597654 -1 .051605
Dummy2009	-1.725278	0.4559528	-3.78	0.000	-2.623124 -0.8274314
Dummy2010	-0.9043	0.392586	-2.30	0.022	-1..677367 -0.1312331
Dummy2011	-0.8238243	0.3914522	-2.10	0.036	-1.594659 -0.0529901
Dummy2012	-1.765023	0.5499394	-3.21	0.001	-2.847945 -0.6821017
Dummy2013	-1.65307	0.646006	-2.56	0.011	-2.925163 -0.380977
Dummy2014	-0.9196268	0.6429156	-1.43	0.154	-2.185634 0.3463804
Dummy2015	-1.553007	0.8100863	-1.92	0.056	-3.148201 0.0421869
Cons	17.4161	7.268569	2.40	0.017	3.103085 31.72912

Risultati dell'elaborazione con il software statistico STATA

Ampliando il set informativo con il triennio 2013-2015, il numero dei regressori che concorrono alla definizione del tasso di occupazione irregolare nell'accezione irr1 (numero di occupati) rimane uguale, ossia pari a otto variabili. Cambiano invece la composizione delle stesse. In particolare, concorrono alla definizione della variabile risposta in entrambi i modelli FE3(irr1) e FE6(irr1): la densità di popolazione, l'istruzione di terzo livello, la quota delle entrate tributarie, l'indice di regolamentazione, la disoccupazione giovanile, le rapine denunciate. Il modello ampliato FE6(irr1) acquisisce un gruppo di variabili che analizza la struttura economica delle regioni, ossia l'imprenditorialità femminile e la natalità delle imprese, area che non era stata presa in considerazione nel modello FE3(irr1), con un conseguente guadagno in termini di qualità informativa. Rimangono invece escluse dal modello ampliato FE6(irr1) due variabili sulla criminalità, ossia il numero di omicidi volontari consumati e la percezione della criminalità. Rimane comunque presente la variabile rapine denunciate, quale rappresentante del gruppo sulla criminalità[47].

Anche in questo caso le variabili dummy riferite all'anno, vengono modellizzate secondo una parametrizzazione corner point. L'anno base è ancora il 2005. Per il modello FE6(irr1) gli anni significativamente diversi dall'anno base sono il 2002 e il 2014.

[47] In appendice la Tabella A10 sintetizza i risultati principali dei tre modelli esaminati.

Interpretazione economica dei risultati del modello FE6(irr1)

Il modello FE6(irr1) evidenzia come tutte le variabili esplicative abbiano coefficienti significativi e mostrino un segno atteso coerente con la teoria economica.

In merito alle variabili che definiscono la struttura socio-demografica del Paese, sono risultate con coefficiente significativo la variabile esplicativa che esprime la densità di popolazione e la variabile esplicativa che esprime il tasso di partecipazione all'istruzione di terzo livello. La prima ha un coefficiente con segno negativo, poiché laddove la maggior densità è legata a una necessità lavorativa, tale variabile può essere correlata negativamente all'Economia Sommersa (Morvillo, 2016). L'istruzione di terzo livello ha un segno del coefficiente positivo. Secondo la letteratura economica (Cappariello e Zizza, 2009) l'istruzione, poiché è un elevatore culturale e sociale, dovrebbe avere un ruolo di contrasto rispetto al fenomeno in esame, pertanto il segno adeguato al coefficiente della variabile esplicativa relativa all'istruzione dovrebbe essere negativo. C'è però da osservare che in mercati di lavoro "saturi" una elevata istruzione garantisce sicuramente delle migliori opportunità occupazionali, ma non sempre maggiori possibilità (Lisi, 2010). Alla luce di tale ultima osservazione, il segno del coefficiente risulta essere coerente, in quanto la variabile esprime un livello di istruzione alto (tasso di partecipazione all'istruzione di terzo livello).

Concorrono alla definizione della struttura economica regionale le entrate tributarie, che si pongono l'obiettivo di sintetizzare il contesto economico-istituzionale comune a tutte le 20 regioni, il tasso di imprenditorialità femminile[48] e il tasso di natalità delle imprese. La prima variabile esprime la dotazione di industrie presenti in ogni regione italiana, con particolare riferimento a imprese con donne imprenditrici[49] (Alleva, 2017), mentre la seconda si focalizza sulla nascita di nuove imprese (ISTAT, 2017), riferendosi a unità nate "da zero" (nascite reali) senza cioè il coinvolgimento di altre unità, attraverso ad esempio scorpori e/o fusioni. Il coefficiente della variabile che esprime le entrate tributarie, mostra un segno coerente (+) con la letteratura economica (Zizza, 2002; Amendola e Dell'Anno,2008), infatti, è riconosciuta come la causa principale. Deve esserci coerenza tra la pressione fiscale e i suoi benefici. Se l'onere fiscale è troppo elevato rispetto a ciò che lo Stato fornisce alle imprese e alle persone, queste faranno di tutto per evadere le imposte. Il tasso di imprenditorialità femminile e il tasso di natalità delle imprese mostrano un segno del coefficiente negativo e coerente con la letteratura economica, poiché solo nelle regioni in cui la dotazione di industrie è particolarmente carente o ha una struttura medio piccola ci si attende una maggior diffusione di Economia Sommersa (Zizza, 2002; Daniele e Marani, 2008). Ciò è coerente anche con le stime dell'ISTAT, in base alle quali le imprese, con particolare riferimento al settore dell'industria, mostrano una propensione minore a evadere (ISTAT, 2018).

Tra le variabili di controllo sono presenti, nel modello FE6(irr1), l'indicatore di regolamentazione, utile a fornire una fotografia del contesto istituzionale italiano e la disoccupazione giovanile[50]. In particolare, l'indicatore di regolamentazione non ha un segno coerente con la teoria economica. Infatti, questa fa

[48] Si precisa che tale variabile non è risultata significativa alla correzione congiunta eteroschedasticità/correlazione dei residui attraverso la procedura robust/cluster.

[49] L'identificazione dell'imprenditore all'interno di un'impresa avviene tramite l'applicazione di opportune regole deterministiche parzialmente differenti a seconda della forma giuridica delle imprese:
1) nel caso delle imprese individuali, l'imprenditore corrisponde alla figura del titolare;
2) nelle società di persone l'imprenditore viene identificato tra i soci che posseggono una carica di amministratore (ad esempio nelle società in nome collettivo) o di accomandatario (nelle società ad accomandita semplice). Un caso a parte sono gli studi associati. Per questa tipologia di forma giuridica ogni associato viene definito imprenditore;
3) nelle società di capitale e nelle società cooperative l'imprenditore viene identificato tra i soci, utilizzando informazioni sia sulle cariche sociali, sia sul fatto di detenere o meno, e in che misura, quote azionarie.

[50] Si precisa che tali variabili non sono risultate significative alla correzione congiunta eteroschedasticità/correlazione dei residui attraverso la procedura robust/cluster

generalmente riferimento a indicatori di regolamentazione costruiti tenendo conto dello stock di tutte le leggi in vigore, dello Stato e degli enti locali, relative all'accesso al lavoro, alla sicurezza sociale, alle ore lavorative, alle condizioni di lavoro, all'esercizio dell'attività d'impresa. In questo contesto si è deciso di utilizzare quale indicatore il rapporto tra i dipendenti ascrivibili al pubblico impiego e le forze di lavoro in età 15-64 anni. Tale indicatore, così come costruito, è stato utilizzato da Frey e Weck-Hanneman nel 1984. Nel suddetto studio la relazione trovata dagli autori era risultata positiva. Va però evidenziato che l'analisi era stata applicata su un campione di 17 Paesi OECD, con riferimento all'arco temporale 1960-1978 e con un modello diverso da quello utilizzato nel presente approfondimento. Sulla base di queste considerazioni, il risultato ottenuto appare coerente, ed è pertanto agevole ritenere che nelle zone con una maggiore presenza di dipendenti pubblici il sommerso sia meno radicato. La Pubblica Amministrazione combatte tale fenomeno, e questo risultato fornisce una dimostrazione della positiva opera dei pubblici dipendenti di tutte le istituzioni centrali e periferiche (Morvillo,2016). Per quanto riguarda l'esplicativa relativa alla disoccupazione giovanile, essa ha un coefficiente con segno coerente con la teoria economica (+), in quanto più elevata è la disoccupazione, più elevata sarà l'Economia Sommersa (Lucifora, 2003).

Infine, in rappresentanza del gruppo riguardante la diffusione della criminalità, concorre alla definizione del modello FE6(irr1) l'esplicativa che esprime il numero di rapine denunciate. Il suo coefficiente mantiene il segno atteso (+), pertanto esiste una correlazione positiva tra il numero di delitti (ossia le rapine) e l'Economia Sommersa così come emerge dalla letteratura (Daniele e Marani, 2008; Marini e Turato, 2002).

3.5 Conclusioni

L'approfondimento svolto nel presente Capitolo, prendendo spunto da molti studi esistenti sull'Economia Sommersa, si incardina nel filone di ricerche con approccio modellistico. Il metodo econometrico ha riscosso negli ultimi anni molto successo in quanto è in grado di studiare l'Economia Sommersa attraverso le sue cause, non limitandosi solamente all'analisi degli aspetti puramente fiscali, ma individuando anche fattori di carattere sociale ed economico che in misura diversa influenzano il fenomeno. In accordo con l'ipotesi che il lavoro irregolare è "il principale fattore produttivo su cui si basa il funzionamento dell'economia sommersa" (Lucifora, 2003), la variabile in esame viene in questo contesto identificata con il tasso di irregolarità del lavoro. L'ISTAT calcola il fattore lavoro non solo in termini di unità di lavoro (ULA), ma anche in termini di ore lavorate e occupati. Per ognuno di essi viene calcolata la percentuale dovuta all'attività irregolare. I dati relativi al tasso di irregolarità del lavoro, costruito come rapporto percentuale tra unità di lavoro non regolare e unità di lavoro totali (irr), sono attualmente fermi al 2012. Sono invece stati rilevati dal database on-line dell'ISTAT (I.Stat) i dati aggiornati a dicembre 2018 e relativi al fattore lavoro in termini di occupati irregolari (irr1), disponibili fino al 2015. Poiché quest'ultima variabile, quale proxy dell'Economia Sommersa, non viene usualmente utilizzata negli studi econometrici, appare meritevole di approfondimento non solo in quanto elemento di novità nell'ambito della letteratura dedicata all'argomento, ma anche in considerazione di un facile reperimento di dati aggiornati.

L'analisi è stata applicata su un campione di dati costituito da un panel bilanciato relativo alle 20 regioni d'Italia, con 12 osservazioni annuali comprese tra il 2001 e il 2012, sia per la variabile oggetto di studio nell'accezione irr che nell'accezione irr1. Per entrambe le variabili dipendenti si è analizzato il contributo fornito da 16 variabili esplicative, volte a spiegare la struttura sociodemografica, economica e a definire il fenomeno della criminalità. Avendo a disposizione per la variabile dipendente nell'accezione irr1 anche dati annuali per il triennio 2013-2015, si è ritenuto utile definire un modello che considerasse questa ulteriore informazione.

Lo studio effettuato, oltre a confermare alcune relazioni già esistenti, ha fatto emergere alcuni risultati importanti. Avere un livello di istruzione troppo elevato non sempre è un fenomeno di contrasto all'Economia Sommersa, poiché in mercati di lavoro "saturi" una elevata istruzione garantisce sicuramente delle migliori opportunità di lavoro ma non sempre maggiori possibilità. L'intensità della regolamentazione non sempre aumenta l'Economia Sommersa. Ciò dipende dalla modalità di costruzione dell'indicatore, dal campione di riferimento utilizzato e dalla tecnica di stima applicata. L'interpretazione economica della nuova relazione trovata è perfettamente intuibile considerando la specifica scelta dell'indicatore. È , infatti, agevole ritenere che nelle zone con una maggiore presenza di dipendenti pubblici il sommerso sia meno radicato e ciò a dimostrazione della positiva opera dei pubblici dipendenti di tutte le istituzioni centrali e periferiche. Infine, la relazione tra l'Economia Sommersa e la densità di popolazione mostra un segno negativo, poiché laddove la maggior densità è legata a una necessità lavorativa, tale variabile può essere correlata negativamente all'Economia Sommersa.

In termini di policy, i risultati emersi ci suggeriscono di intervenire in modo più incisivo sul problema della disoccupazione, soprattutto giovanile e sul peso della tassazione, che risultano essere delle variabili che incentivano il fenomeno dell'Economia Sommersa. È importante inoltre tenere sotto controllo il tema della criminalità.

APPENDICE CAPITOLO 3

Tabella A1: Valori del Tasso di irregolarità del lavoro calcolato come percentuale di unità di lavoro irregolari sul totale delle unità di lavoro (irr) in Italia per ripartizione geografica, 2001-2012

Regioni	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012
Piemonte	10,8	9,6	8,4	8,8	9,7	10,1	10,0	10,3	10,8	11,1	11,6	11,3
VDA	10,4	10,0	9,9	10,5	10,9	11,2	10,8	13,4	11,0	11,5	11,3	11,4
Lombardia	9,5	8,2	7,1	7,6	7,5	8,0	8,5	8,2	8,1	7,5	7,3	7,1
TAA	9,2	8,8	8,6	8,6	9,1	8,7	8,6	8,0	8,0	7,5	7,5	7,6
Veneto	10,0	8,9	8,0	8,4	8,4	8,3	8,6	8,4	8,5	8,3	8,3	8,0
FVG	11,4	10,8	10,0	9,8	10,3	10,7	10,8	10,1	10,2	10,4	11,0	10,4
Liguria	13,9	12,0	10,7	11,5	12,5	12,5	12,0	11,6	12,1	12,2	12,7	12,3
ER	9,5	8,6	7,5	7,7	7,9	7,8	8,0	8,3	8,3	8,1	8,2	7,9
Toscana	10,8	9,7	8,6	8,6	9,2	8,9	8,9	9,1	9,1	8,9	9,2	9,0
Umbria	14,8	13,1	11,1	12,0	12,1	12,5	12,6	11,7	11,6	12,0	12,2	12,4
Marche	11,8	10,5	9,8	9,8	9,6	10,0	10,2	9,7	9,9	9,9	9,4	9,4
Lazio	15,0	13,1	11,2	12,2	12,1	11,4	11,3	11,0	11,4	11,3	10,7	10,3
Abruzzo	13,7	13,6	12,2	12,3	13,0	12,7	11,9	12,4	13,1	13,3	13,7	14,0
Molise	18,0	18,5	17,7	16,9	18,2	19,0	19,3	21,6	23,8	22,9	22,9	24,6
Campania	22,9	22,0	21,1	21,0	19,8	19,2	17,7	18,5	18,4	18,4	18,5	19,3
Puglia	18,8	18,1	16,7	15,3	16,5	17,2	17,1	17,5	18,5	17,9	18,0	18,7
Basilicata	18,5	19,2	19,1	18,1	18,2	19,6	18,6	20,0	22,2	20,9	22,4	22,4
Calabria	25,6	25,5	24,2	25,6	27,0	27,7	27,0	26,6	28,6	30,9	28,5	30,9
Sicilia	22,8	21,6	20,9	19,3	21,0	19,7	18,9	18,7	20,3	20,4	20,8	21,3
Sardegna	18,4	17,1	17,7	19,1	18,7	19,4	18,8	18,4	19,4	20,7	21,9	22,9
Italia	13,7	12,6	11,6	11,7	12,0	12,0	11,8	11,8	12,1	12,0	12,0	12,0

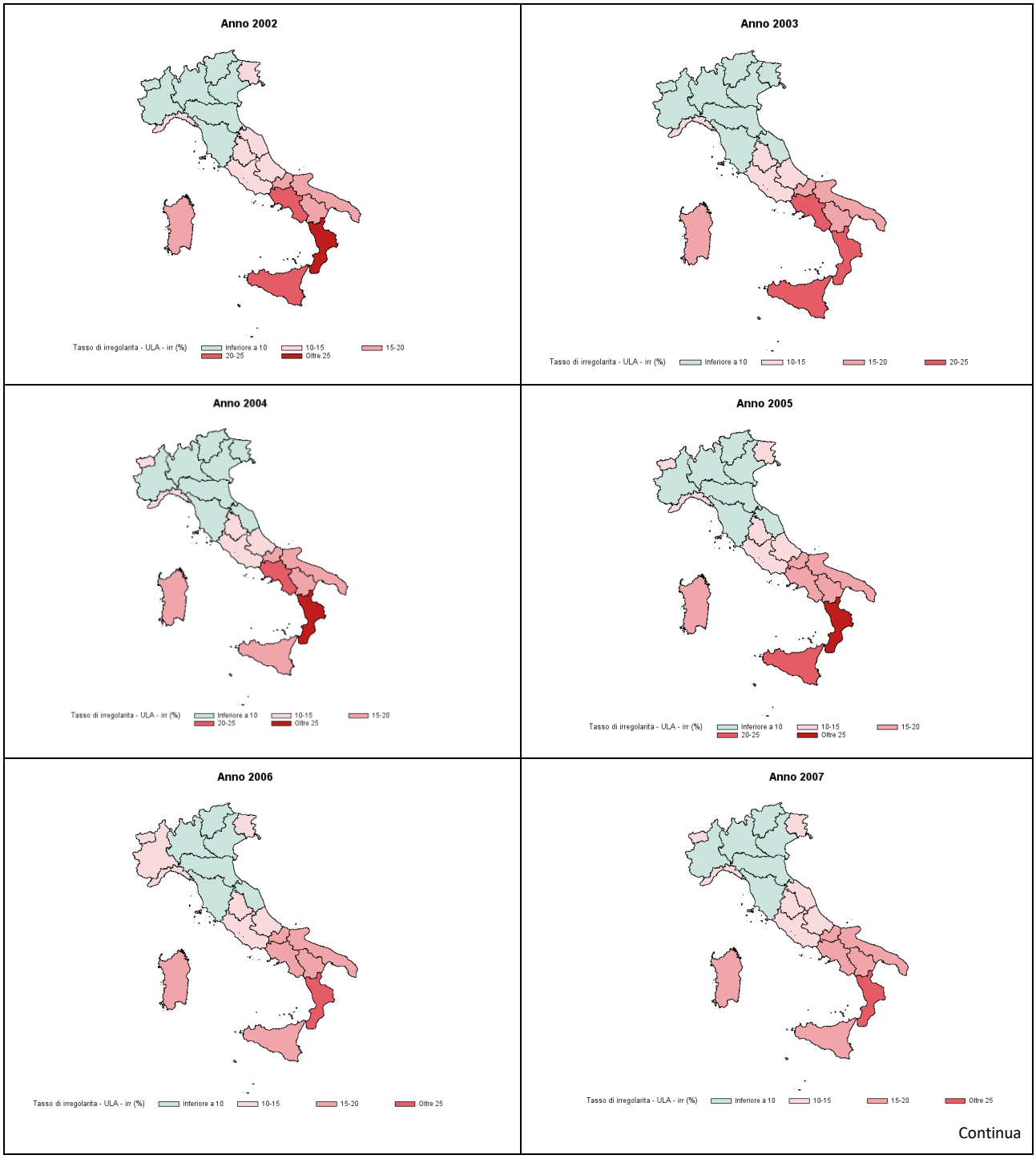
Fonte: Elaborazione su dati ISTAT

Tabella A2: Valori del Tasso di irregolarità del lavoro calcolato come percentuale di occupati irregolari sul totale degli occupati (irr1) in Italia per ripartizione geografica, 2001-2015

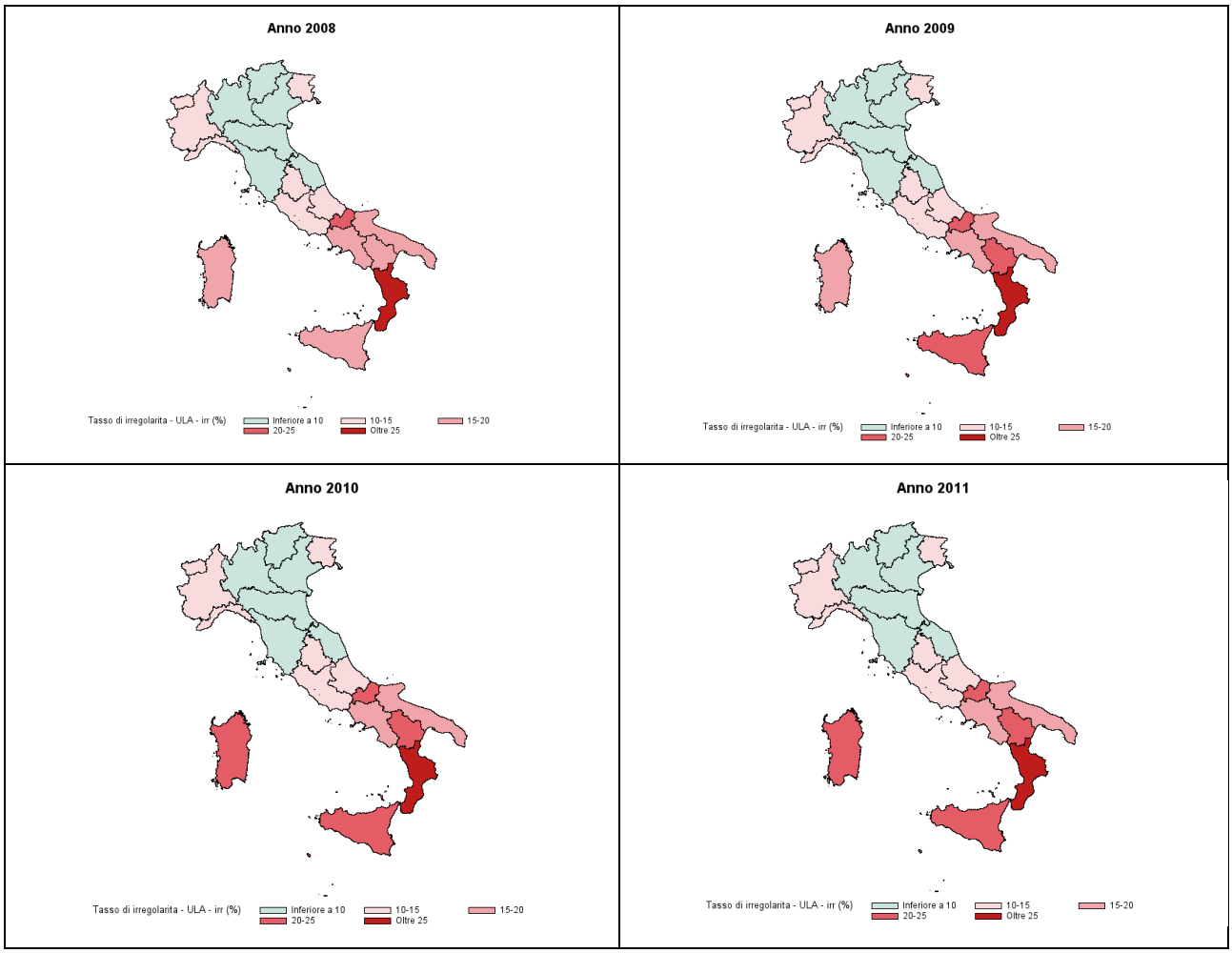
Regione	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Piemonte	7,4	6,6	6,0	6,5	7,1	7,8	7,7	7,7	7,9	8,5	9,3	10,4	10,7	11,0	11,0
VDA	10,1	8,1	6,6	7,1	6,7	6,7	7,1	7,6	7,5	8,3	7,9	9,2	9,8	9,9	11,2
Lombardia	13,8	12,2	10,7	10,9	9,8	10,4	10,8	10,8	10,0	9,2	9,1	9,5	9,5	10,2	10,1
TAA	8,9	8,0	7,1	7,1	7,4	7,5	7,6	7,2	7,3	7,9	8,0	9,1	9,4	9,6	9,8
Veneto	10,0	9,1	8,5	8,6	8,1	8,5	8,6	8,1	8,1	8,0	8,2	8,3	8,3	8,8	9,1
FVG	8,8	8,6	8,8	8,5	8,6	8,7	9,2	8,8	8,8	8,7	9,5	9,9	10,0	10,8	11,0
Liguria	10,3	8,5	7,4	8,3	9,1	9,9	9,3	8,9	9,3	9,6	10,5	11,1	11,5	12,1	12,1
ER	10,7	9,5	8,1	7,8	7,9	8,1	8,4	8,7	8,7	8,5	9,1	9,3	9,7	10,0	10,0
Toscana	12,4	11,2	10,0	9,4	10,0	10,2	11,3	10,8	10,2	9,7	9,7	10,1	10,6	11,1	11,1
Umbria	12,3	11,0	10,2	11,4	11,7	12,4	12,0	11,7	11,2	11,5	11,8	12,5	12,8	12,5	13,3
Marche	11,6	10,0	9,2	9,5	10,0	10,0	10,0	9,5	9,3	9,3	8,7	8,9	9,4	10,2	10,2
Lazio	16,4	15,0	13,7	15,1	14,9	14,0	14,2	13,7	14,2	14,6	13,8	13,8	14,9	16,1	15,8
Abruzzo	14,8	14,7	13,0	13,4	14,0	13,8	12,7	13,5	14,6	14,8	15,3	15,2	15,9	15,7	16,7
Molise	10,0	10,1	10,0	9,9	10,6	11,6	11,3	12,6	14,1	13,7	14,5	15,0	14,9	15,6	15,6
Campania	26,3	26,0	25,0	25,6	24,5	23,6	21,4	22,1	22,4	22,2	22,1	21,3	20,8	21,5	21,0
Puglia	16,9	16,8	15,7	14,6	15,5	16,1	16,0	16,1	16,7	16,7	17,3	17,1	16,5	16,8	17,6
Basilicata	9,0	9,5	9,5	8,8	9,3	10,3	10,2	11,3	12,3	11,8	12,8	13,9	14,2	15,0	15,0
Calabria	20,2	20,6	19,0	19,5	20,5	21,4	20,7	20,9	21,6	23,3	22,8	22,4	22,6	23,0	23,2
Sicilia	20,9	19,9	19,4	18,7	20,4	19,5	18,4	18,0	19,1	19,2	19,4	19,5	19,8	20,3	20,6
Sardegna	9,7	9,3	9,9	10,8	10,8	11,4	11,2	11,2	11,7	13,1	14,0	14,1	14,3	14,8	15,4
Italia	14,2	13,3	12,3	12,4	12,4	12,5	12,4	12,2	12,3	12,3	12,4	12,6	12,8	13,3	13,1

Fonte: Elaborazione su dati ISTAT

Figura A1: Tasso di irregolarità del lavoro nell'accezione irr per regione, 2002-2011

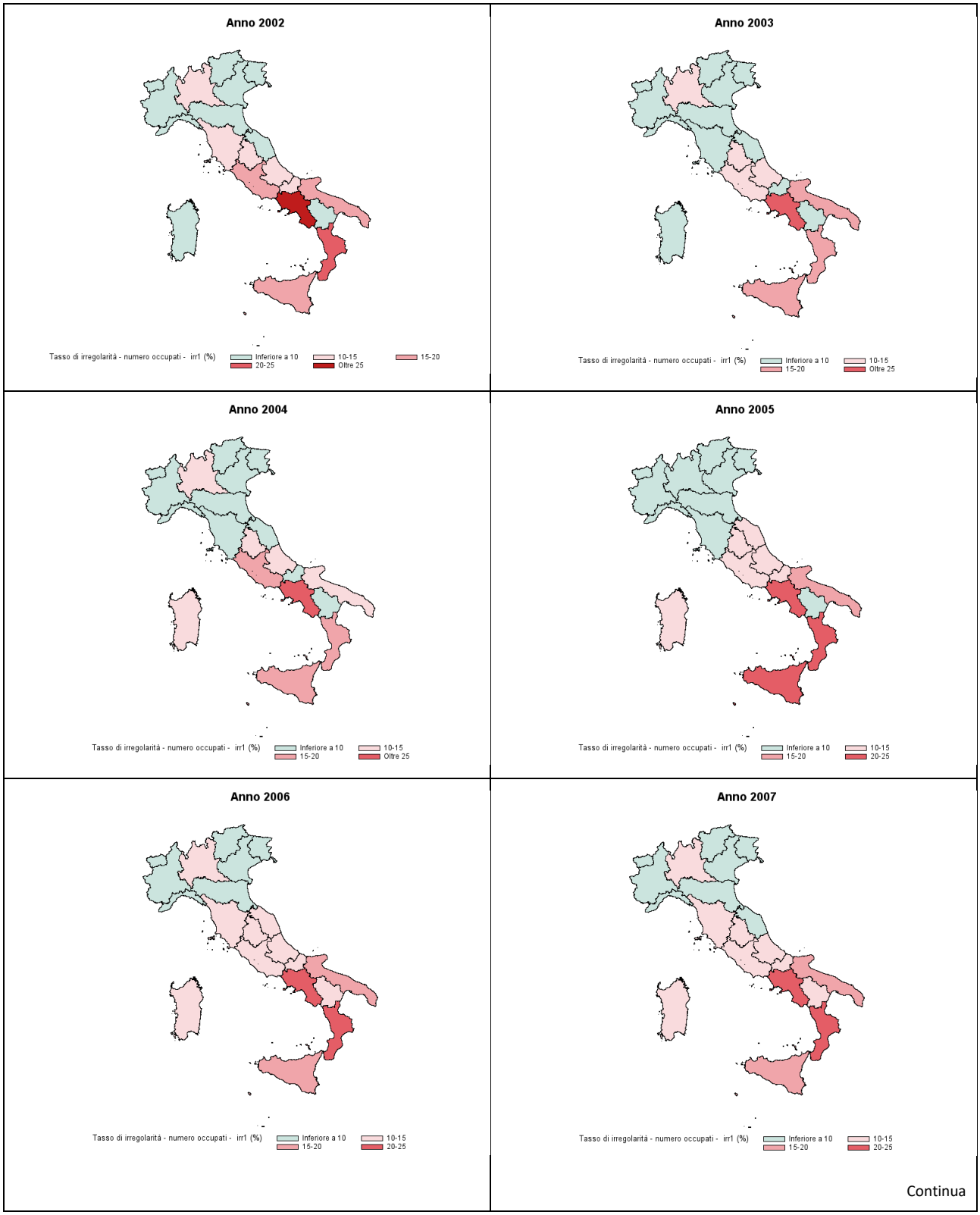


Continua



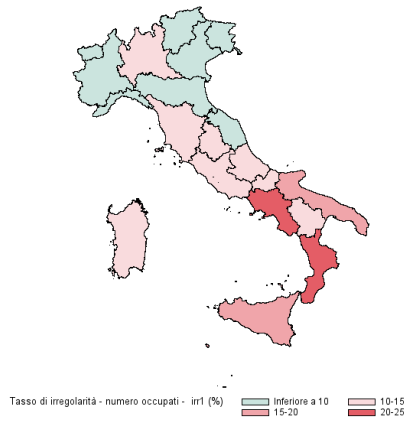
Fonte: Elaborazione su dati ISTAT

Figura A2: Tasso di irregolarità del lavoro nell'accezione irr1 per regione, 2002-2014

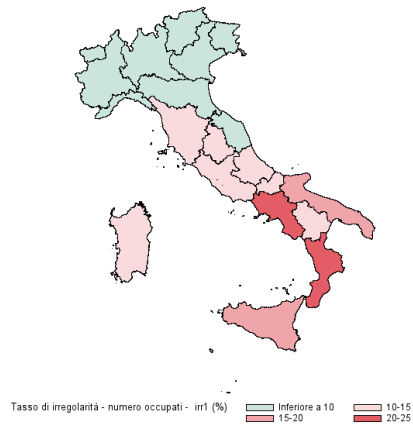


Continua

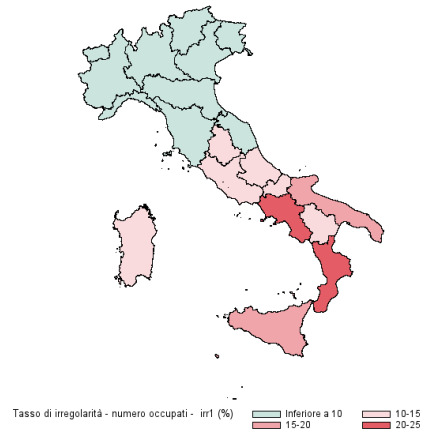
Anno 2008



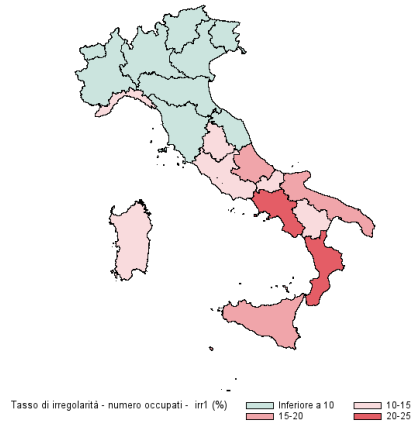
Anno 2009



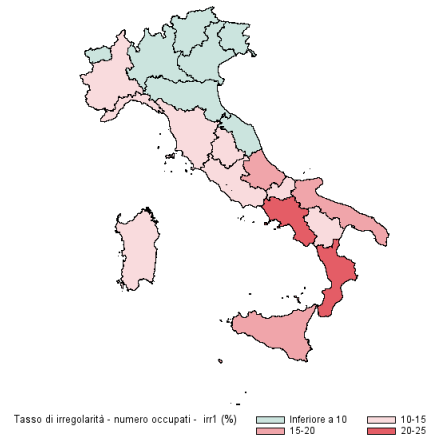
Anno 2010



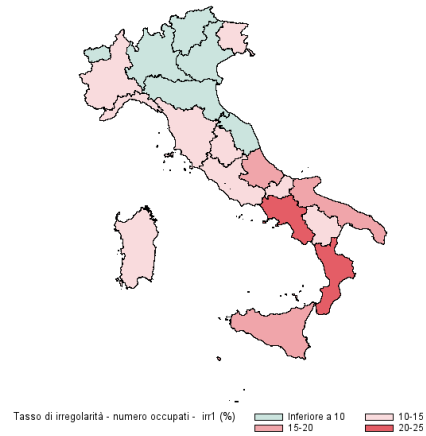
Anno 2011



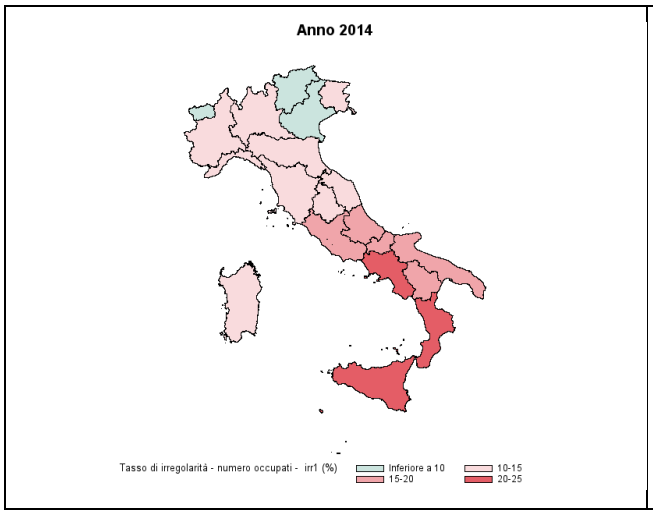
Anno 2012



Anno 2013



Continua



Fonte: Elaborazione su dati ISTAT

Tabella A3: Risultati dell'analisi descrittiva per il periodo 2001-2012[51]

Variabile		Media	Deviazione	Variabile		Media	Deviazione	Variabile		Media	Deviazione
Irr	Overall	13,8	5,6	**Istr**	Overall	94,8	6	**Istr1**	Overall	12,9	2,7
	Between		5,6		Between		5,5		Between		1,9
	Within		1,2		Within		2,6		Within		2
crimperc	Overall	23,8	10,5	**Pilpc**	Overall	24,1	6	**natalitàimp**	Overall	7,2	1,2
	Between		10		Between		6		Between		1
	Within		3,8		Within		1,4		Within		0,6
Dens	Overall	107,8	106,2	**Tax**	Overall	25,6	0,6	**omicidi**	Overall	1	0,8
	Between		108,7		Between		0		Between		0,7
	Within		3,8		Within		0,6		Within		0,4
dippub	Overall	15,3	2,9	**Disocgiov**	Overall	24,5	11,7	**rapine**	Overall	0,5	0,5
	Between		2,8		Between		10,7		Between		0,5
	Within		1,1		Within		5,3		Within		0,1
Fem	Overall	46,6	11,3	**Furti**	Overall	20,9	7,7	**imprengiov**	Overall	7,3	1,5
	Between		11,4		Between		7,6		Between		1,3
	Within		1,7		Within		2,3		Within		0,8
Indu	Overall	19,1	6,9	**Imprenfem**	Overall	26,7	3,4	**irr1**	Overall	12,2	4,6
	Between		7		Between		3,4		Between		4,6
	Within		1		Within		0,5		Within		1,1

Risultati dell'elaborazione con il software statistico STATA

[51] I valori indicati in tabella sono tutti approssimati alla prima cifra decimale.

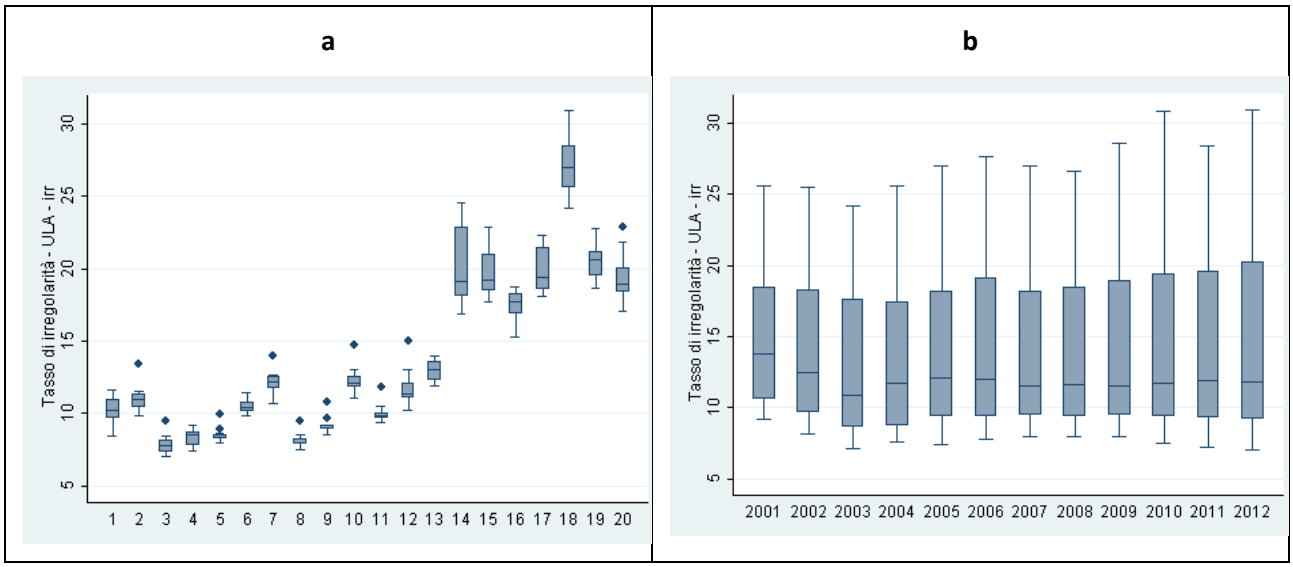
Tabella A4: Risultati dell'analisi descrittiva per il periodo 2001-2015[52]

Variabile		Media	Deviazione	Variabile		Media	Deviazione	Variabile		Media	Deviazione
Irr1	Overall	12,5	4,6	**imprengiov**	Overall	7,0	1,5	**Istr1**	Overall	13,7	3,1
	Between		4,5		Between		1,3		Between		2
	Within		1,3		Within		0,8		Within		2,4
crimperc	Overall	24,7	10,7	**pilpc**	Overall	24,4	6,2	**natalitàimp**	Overall	7,2	1,2
	Between		9,8		Between		6,2		Between		1,1
	Within		4,9		Within		1,8		Within		0,5
Dens	Overall	179,1	107,3	**tax**	Overall	25,8	0,7	**omicidi**	Overall	1	0,8
	Between		109,8		Between		0		Between		0,6
	Within		5,3		Within		0,7		Within		0,6
dippub	Overall	15,1	2,9	**disocgiov**	Overall	27,7	13,4	**rapine**	Overall	0,5	0,5
	Between		2,8		Between		10,7		Between		0,4
	Within		1,1		Within		8,3		Within		0,1
Fem	Overall	46,8	11,4	**furti**	Overall	20,9	7,8				
	Between		11,5		Between		7,5				
	Within		1,7		Within		2,2				
Indu	Overall	18,9	6,9	**imprenfem**	Overall	26,8	3,3				
	Between		7		Between		3,3				
	Within		1		Within		0,6				

Risultati dell'elaborazione con il software statistico STATA

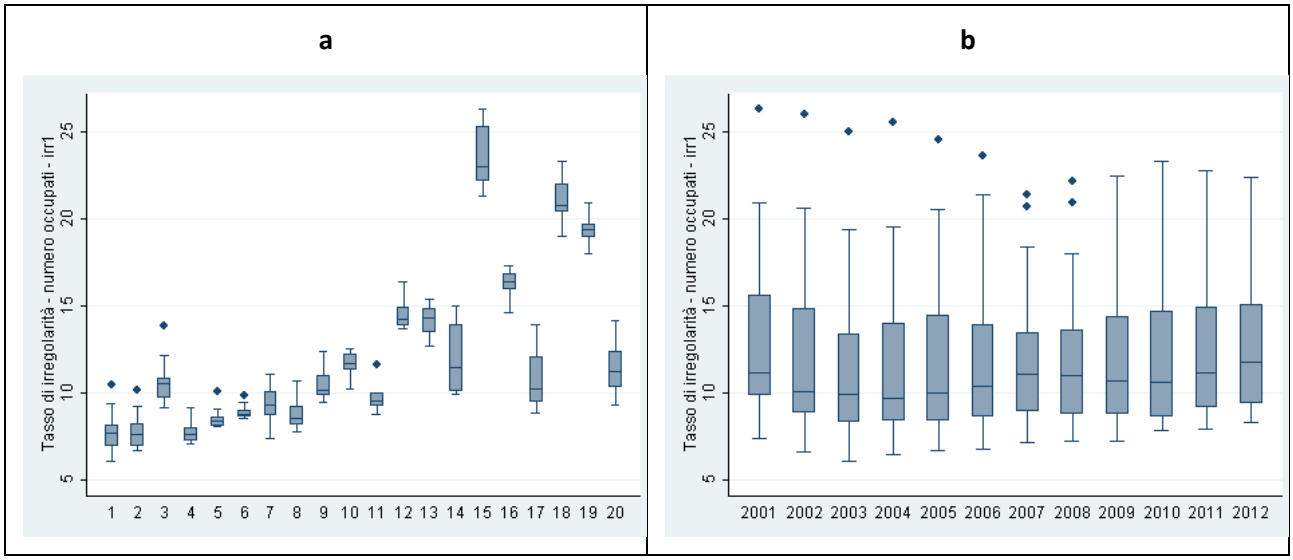
[52] I valori indicati in tabella sono tutti approssimati alla prima cifra decimale.

Figura A3: Rappresentazioni grafiche (box plot) del tasso di irregolarità del lavoro per codice regionale ISTAT (a) e per anno (b) in termini di unità di lavoro (ULA) –irr



Risultati dell'elaborazione con il software statistico STATA

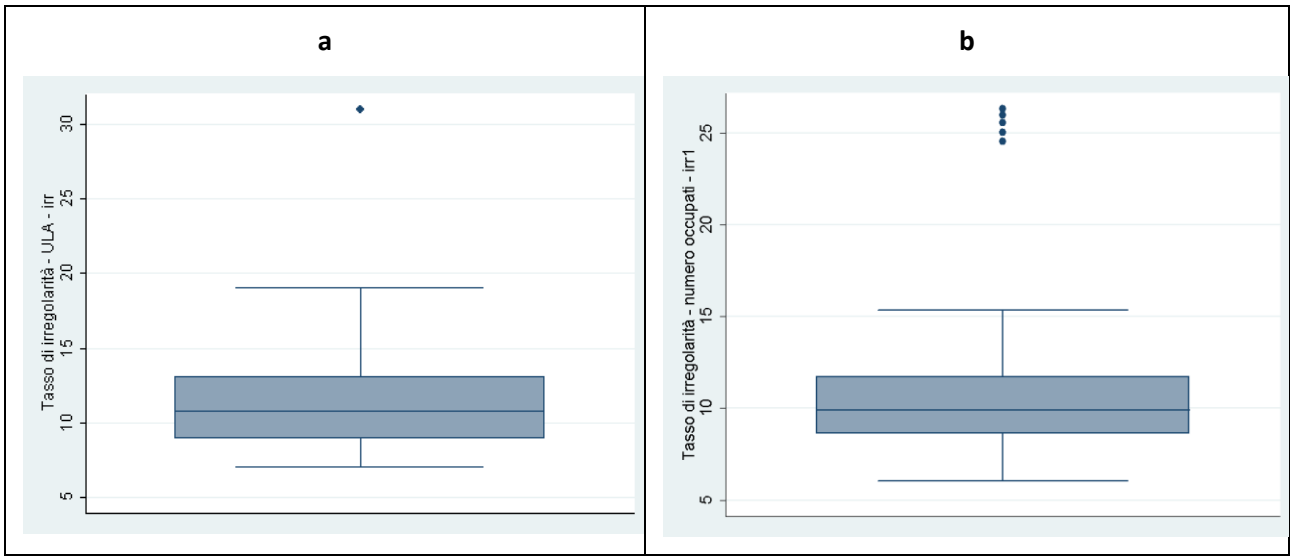
Figura A4: Rappresentazioni grafiche (box plot) del tasso di irregolarità del lavoro per codice regionale ISTAT (a) e per anno (b) in termini di unità di occupati - irr1



Risultati dell'elaborazione con il software statistico STATA

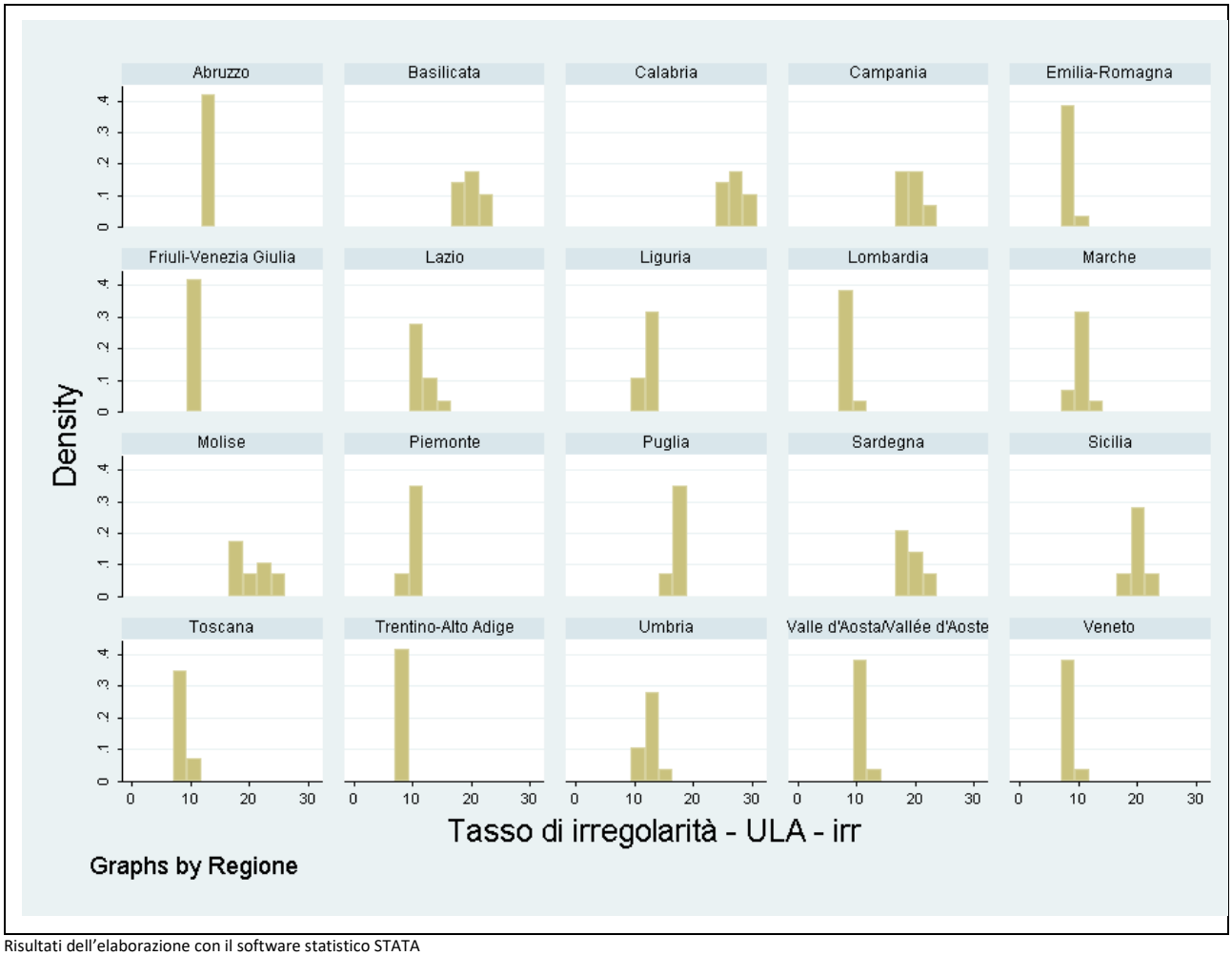
Codice ISTAT	Regione	Codice ISTAT	Regione	Codice ISTAT	Regione	Codice ISTAT	Regione
1	Piemonte	2	Valle d'Aosta	3	Lombardia	4	Trentino Alto Adige
5	Veneto	6	Friuli Venezia Giulia	7	Liguria	8	Emilia Romagna
9	Toscana	10	Umbria	11	Marche	12	Lazio
13	Abruzzo	14	Molise	15	Campania	16	Puglia
17	Bailicata	18	Calabria	19	Sicilia	20	Sardegna

Figura A5: Rappresentazioni grafiche (box plot) del tasso di irregolarità del lavoro sopra l'ottantesimo percentile per entrambe le accezioni irr (a) e irr1 (b)



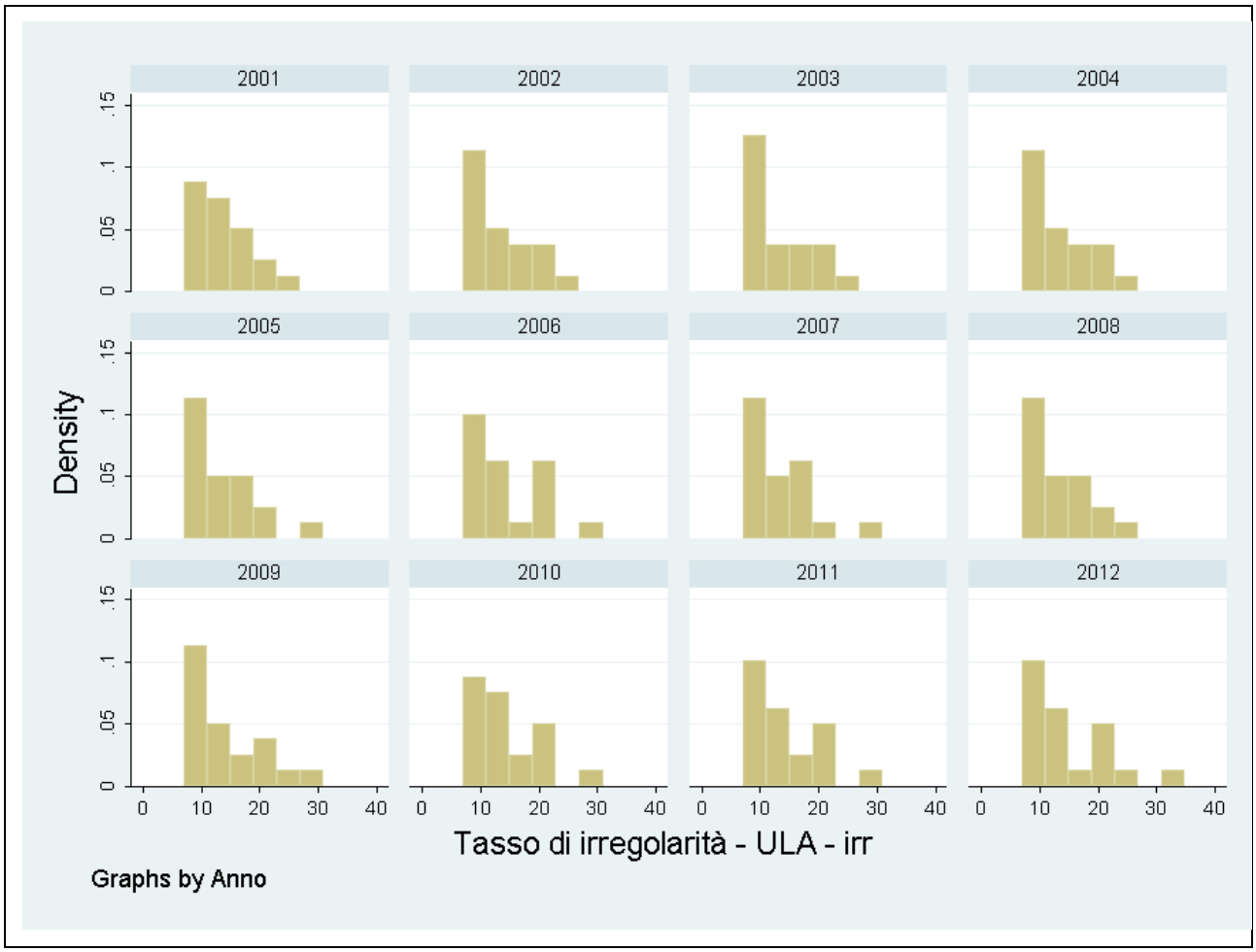
Risultati dell'elaborazione con il software statistico STATA

Figura A6: Istogrammi del tasso di irregolarità del lavoro per regioni in termini di unità di lavoro (ULA) - irr



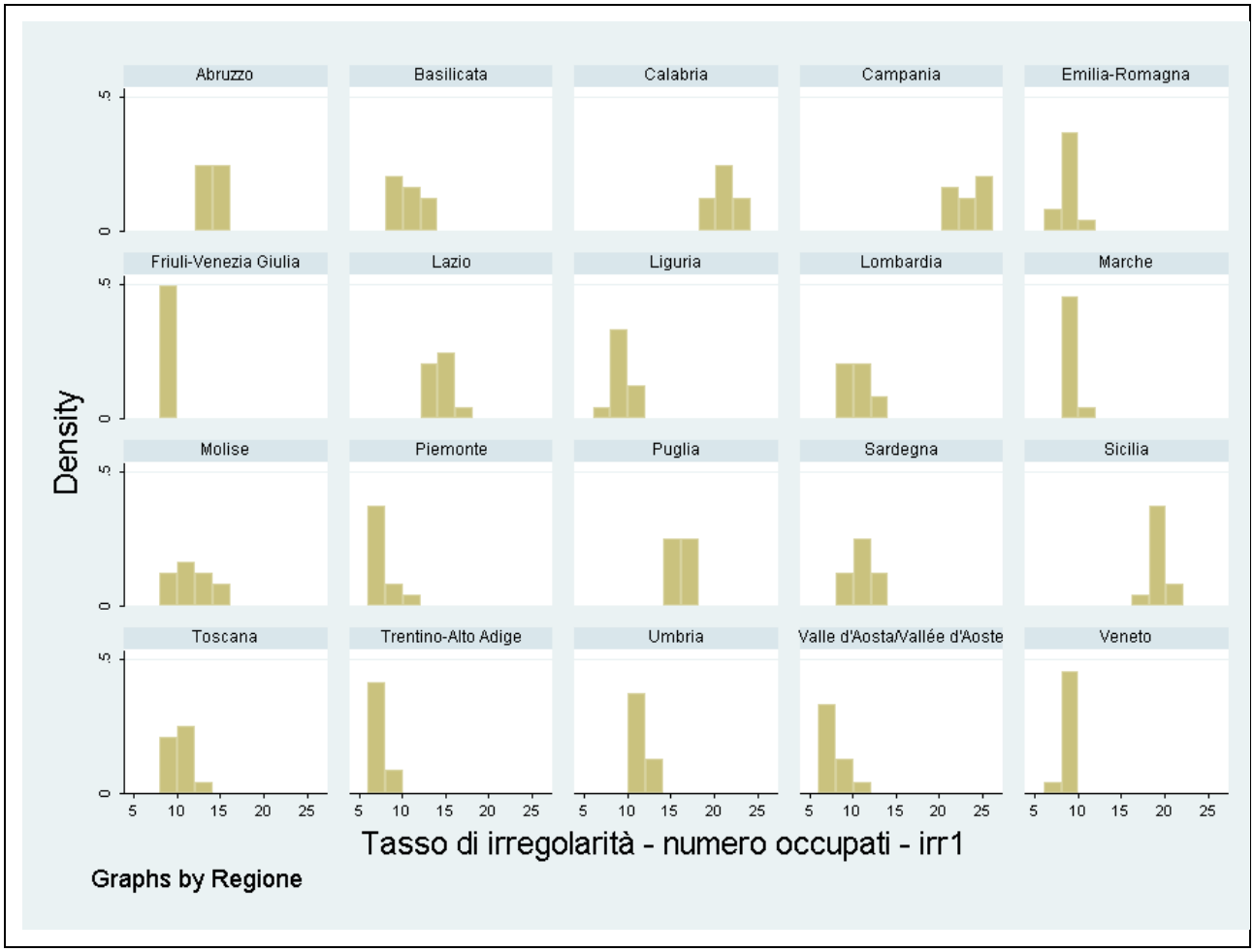
Risultati dell'elaborazione con il software statistico STATA

Figura A7: Istogrammi del tasso di irregolarità del lavoro per anno in termini di unità di lavoro (ULA) - irr



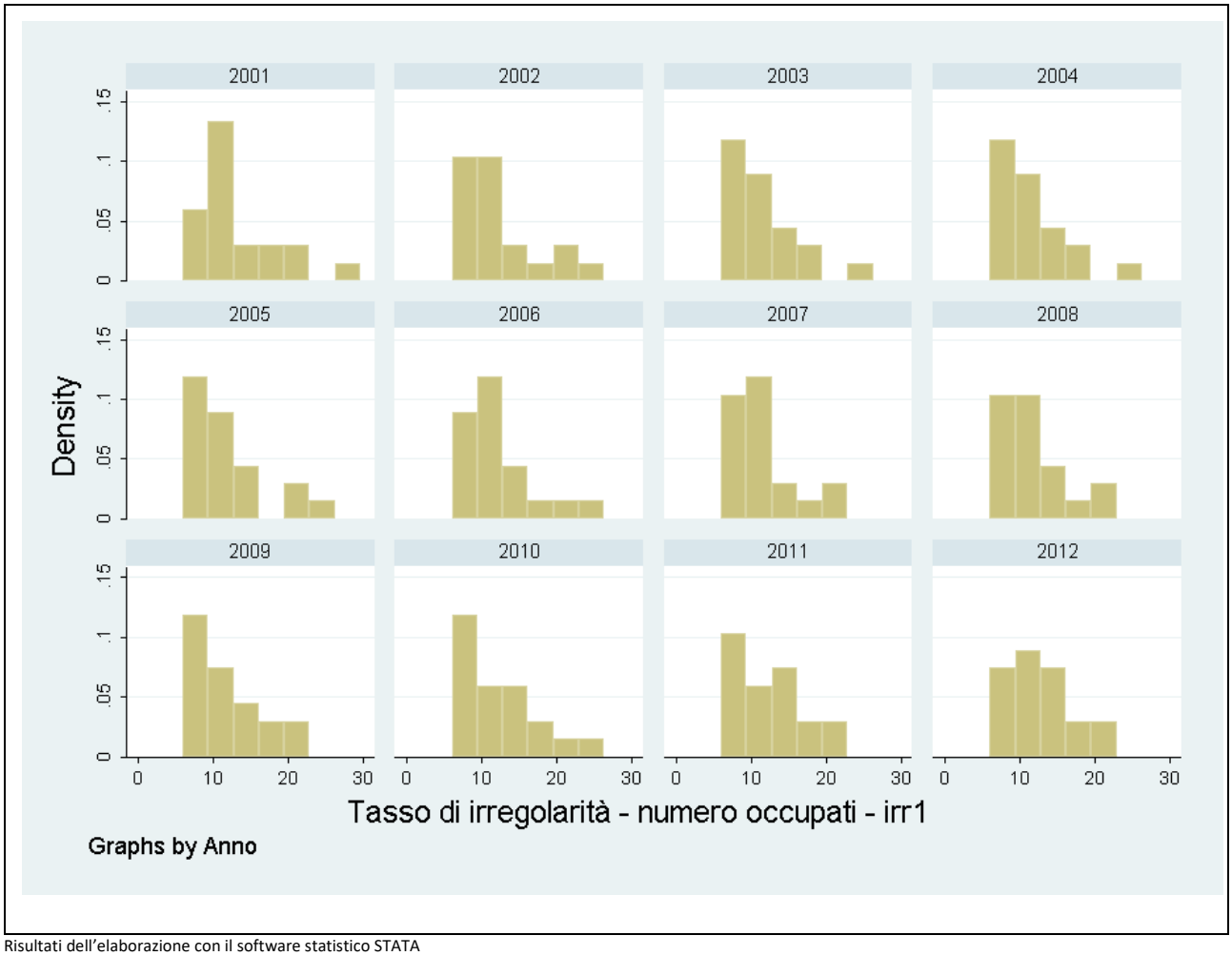
Risultati dell'elaborazione con il software statistico STATA

Figura A8: Istogrammi del tasso di irregolarità del lavoro per regioni in termini di numero di occupati (irr1)



Risultati dell'elaborazione con il software statistico STATA

Figura A9: Istogrammi del tasso di irregolarità del lavoro per anno in termini di numero di occupati (irr1)



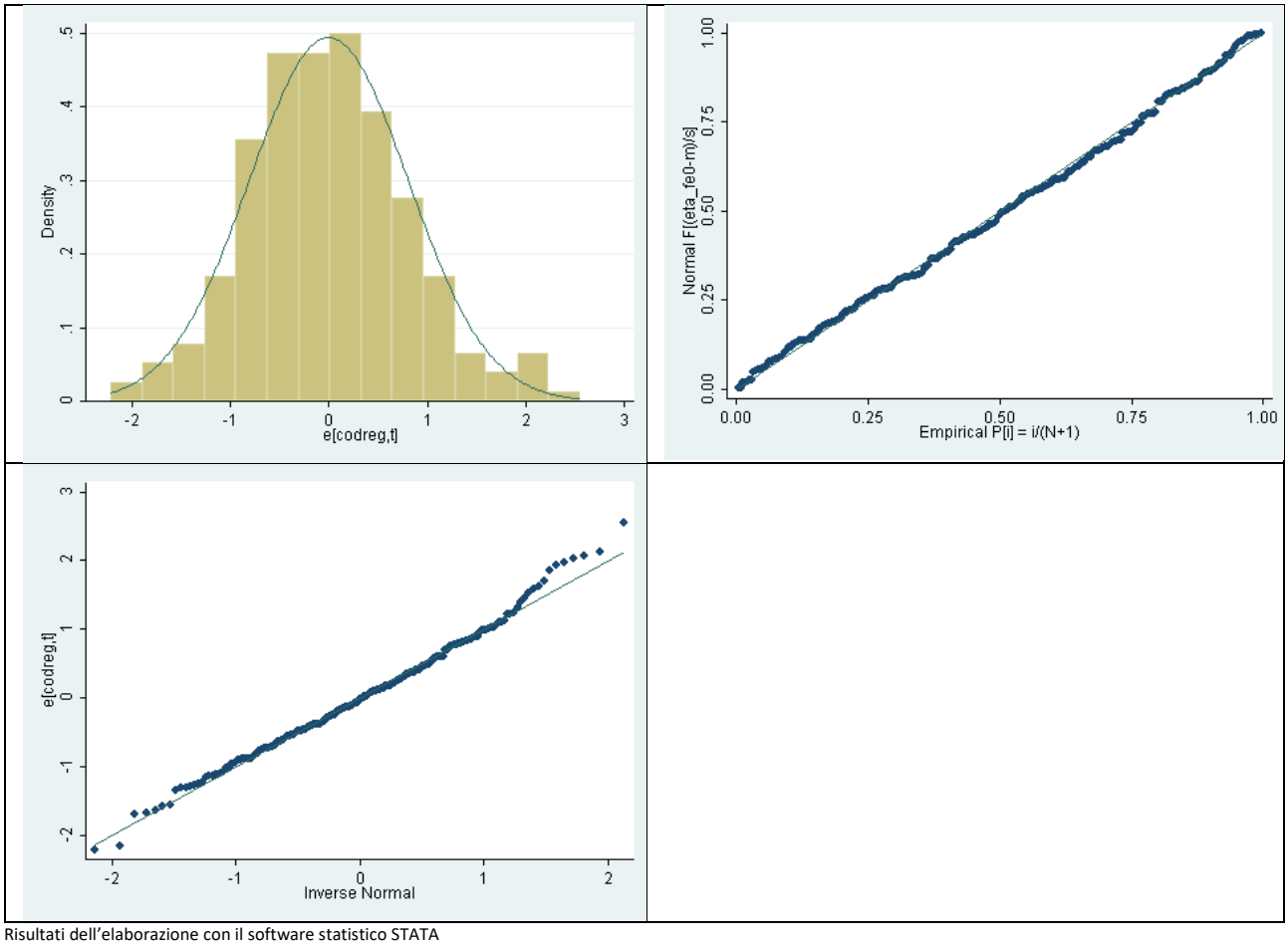
Risultati dell'elaborazione con il software statistico STATA

Tabella A5: Risultati dell'indicatore VIF per i diversi modelli analizzati

Modello 1 e Modello 2		Modello 3	
Variabile	**VIF**	**Variabile**	**VIF**
Fem	(22,41)	Fem	(20,20)
Pilpc	(16,37)	Pilpc	(13,69)
Dippub	6,35	Dippub	5,36
crimperc	5,69	dens	5,18
Dens	5,36	Disocgiov	4,63
disocgiov	5,18	indu	4,46
Furti	5,14	Furti	4,41
imprengiov	4,98	crimperc	4,32
Indu	4,90	natalitàimp	4,29
rapine	4,62	imprengiov	4,21
natalitàimp	4,10	rapine	3,81
istr	3,13	Istr1	3,46
Istr1	3,02	imprenfem	2,64
imprenfem	2,82	tax	2,21
omicidi	2,69	omicidi	2
tax	1,40		

Risultati dell'elaborazione con il software statistico STATA

Figura A10: Risultati della verifica della normalità dei residui – FE1(irr)



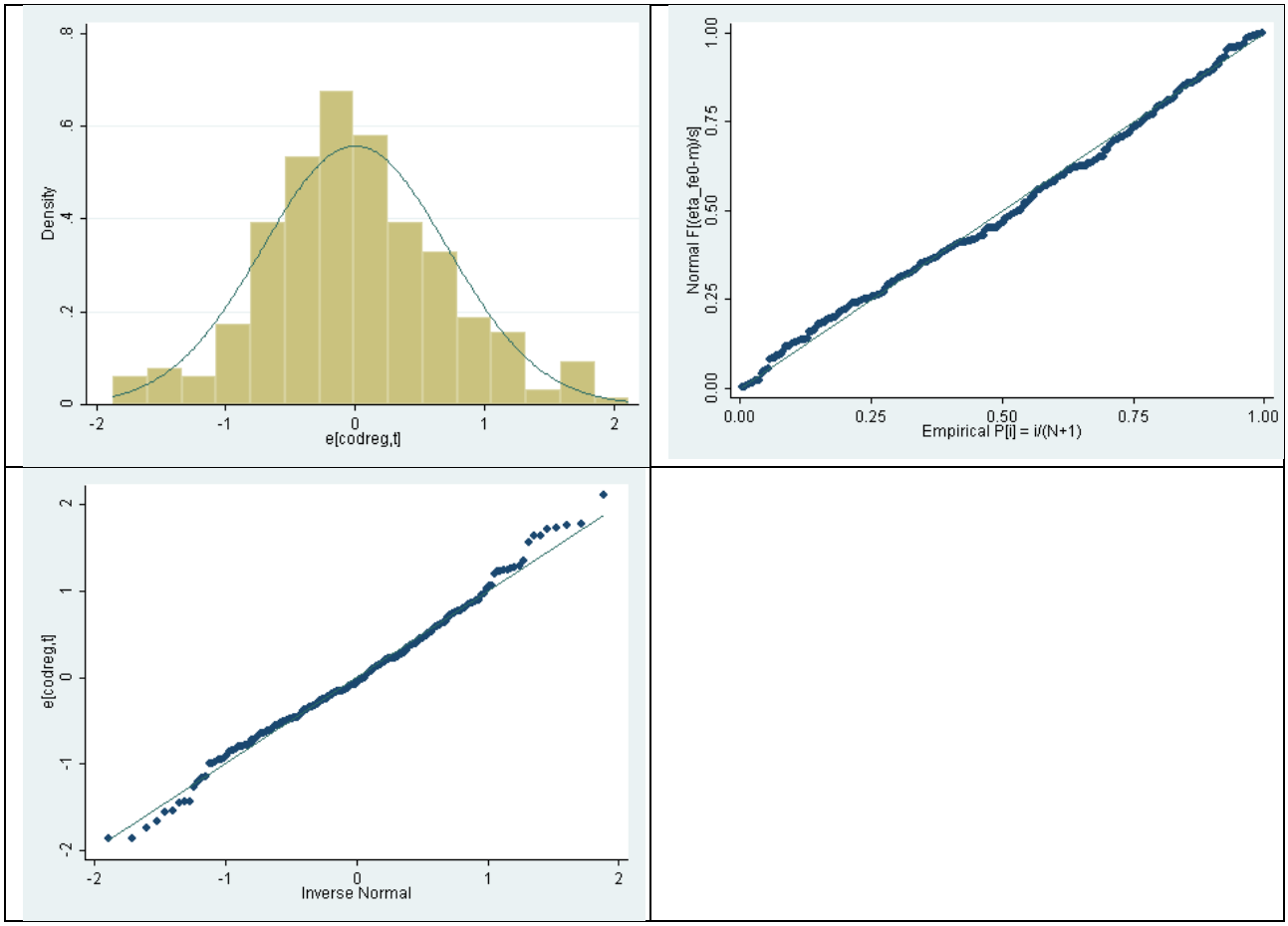
Risultati dell'elaborazione con il software statistico STATA

Tabella A6: Risultati dei test di verifica della normalità dei residui – FE1(irr)

Test di normalità					
Test	Osservazioni	W	V	Z	Prob>z
Shapiro-Wilk	240	0.99316	1.197	0.418	**0.33794**
Test	Osservazioni	W'	V'	Z	Prob>z
Shapiro-Francia	240	0.99290	1.351	0.630	**0.26441**
Test	Osservazioni	Pr(Asimmetria)	Pr(Curtosi)	χ^2	Prob>χ2
Asimmetria/Curtosi	240	0.2893	0.2893	3.93	**0.1402**

Risultati dell'elaborazione con il software statistico STATA

Figura A11: Risultati della verifica della normalità dei residui – FE3(irr1)



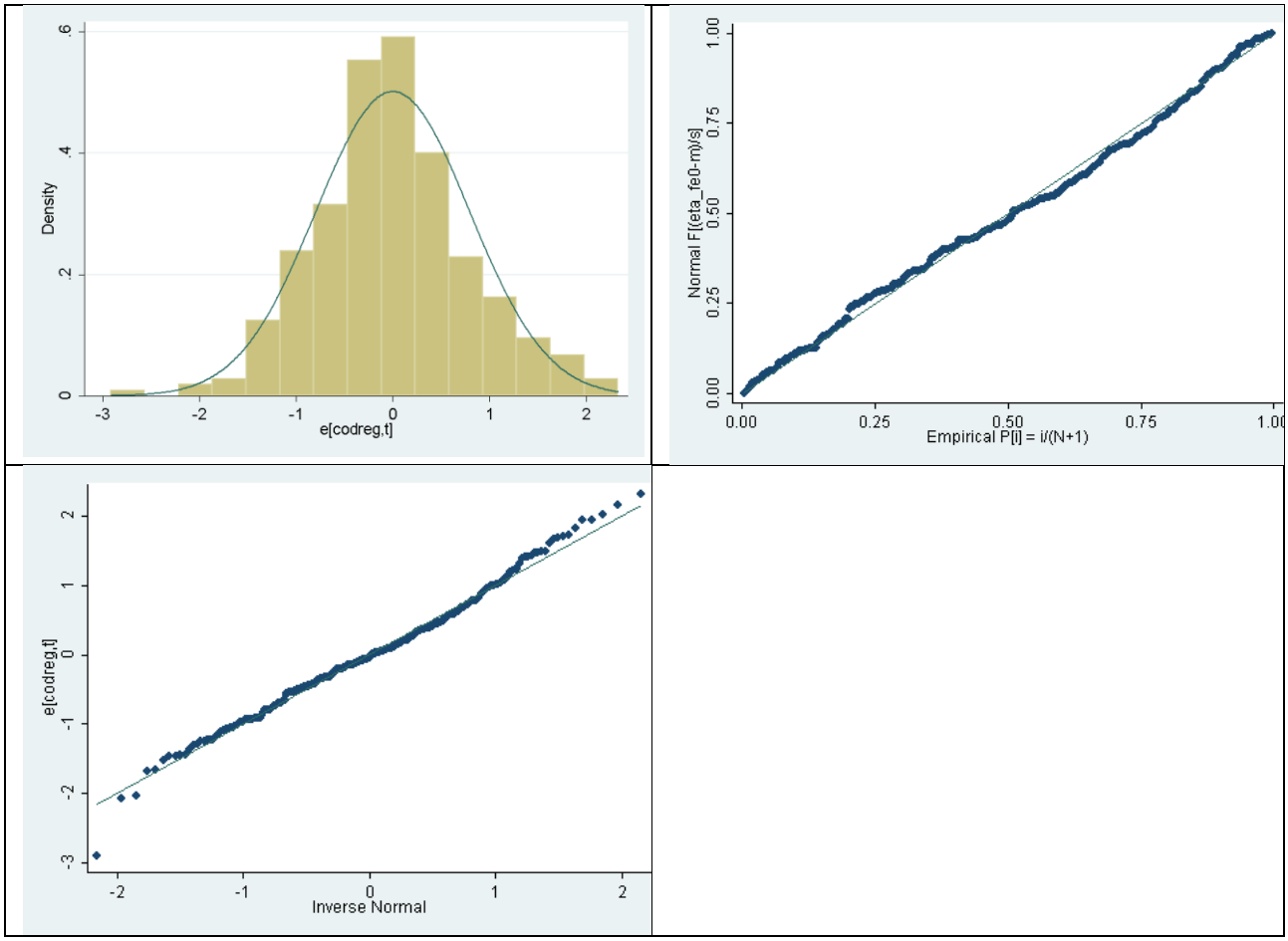
Risultati dell'elaborazione con il software statistico STATA

Tabella A7: Risultati dei test di verifica della normalità dei residui – FE3(irr1)

Test di normalità					
Test	Osservazioni	W	V	Z	Prob>z
Shapiro-Wilk	240	0.99147	1.492	0.929	**0.17648**
Test	Osservazioni	W'	V'	Z	Prob>z
Shapiro-Francia	240	0.99185	1.551	0.918	**0.17922**
Test	Osservazioni	Pr(Asimmetria)	Pr(Curtosi)	χ^2	Prob>χ2
Asimmetria/Curtosi	240	0.2850	0.2882	2.29	**0.3182**

Risultati dell'elaborazione con il software statistico STATA

Figura A12: Risultati della verifica della normalità dei residui – FE6(irr1)



Risultati dell'elaborazione con il software statistico STATA

Tabella A8: Risultati dei test di verifica della normalità dei residui – FE6(irr1)

Test di normalità					
Test	Osservazioni	W	V	z	Prob>z
Shapiro-Wilk	300	0.99204	1.696	1.240	**0.10742**
Test	Osservazioni	W'	V'	z	Prob>z
Shapiro-Francia	300	0.99103	2.070	1.545	**0.06123**
Test	Osservazioni	Pr(Asimmetria)	Pr(Curtosi)	χ^2	Prob>χ2
Asimmetria/Curtosi	300	0.3892	0.0826	3.78	**0.1514**

Risultati dell'elaborazione con il software statistico STATA

Tabella A9: Modello FE6(irr1) con procedure robust[53], jackknife e bootstrap

Variabile	Coefficiente di regressione	p-value	p-value robust	p-value jackknife	p-value bootstrap
Densità di popolazione	-0.1326895	0.000	0.001	0.029	0.013
Livello di istruzione terzo livello	0.2626658	0.000	0.010	0.053	0.009
Tassazione	1.063561	0.000	0.002	0.891	0.860
Imprenditorialità femminile	-0.3133613	0.001	0.164	0.340	0.223
Natalità delle imprese	-0.4137995	0.013	0.030	0.155	0.068
Dipendenti pubblici	-0.1307712	0.016	0.149	0.524	0.316
Disoccupazione giovanile	0.0317894	0.029	0.290	0.396	0.302
Rapine	2.544598	0.000	0.011	0.604	0.313
Dummy2002	0.3158559	0.270	0.337	0.495	0.367
Dummy 2003	-0.5036731	0.082	0.074	0.311	0.177
Dummy2004	-0.8407689	0.001	0.000	0.276	0.135
Dummy2005	-1.86881	Variabile omessa	Variabile omessa	0.483	0.350
Dummy2006	-1.493127	0.000	0.009	0.194	0.058
Dummy2007	-1.414781	0.000	0.041	0.247	0.076
Dummy2008	-1.824629	0.000	0.007	0.135	0.028
Dummy2009	-1.725278	0.000	0.019	0.011	0.001
Dummy2010	-0.9043	0.022	0.090	0.342	0.225
Dummy2011	-0.8238243	0.036	0.208	0.465	0.320
Dummy2012	-1.765023	0.001	0.105	0.322	0.149
Dummy2013	-1.65307	0.011	0.203	0.437	0.253
Dummy2014	-0.9196268	0.154	0.456	0.699	0.595
Dummy2015	-1.553007	0.056	0.277	Variabile omessa	Variabile omessa
Cons	26.732	0.017	0.163	0.143	0.061

Risultati dell'elaborazione con il software statistico STATA

[53] Tale opzione viene applicata insieme all'opzione cluster poiché il raggruppamento produce uno stimatore coerente quando i disturbi non sono identicamente distribuiti o vi è una correlazione seriale.

Tabella A10: Determinanti dell'Economia Sommersa presenti nei tre modelli esaminati con il rispettivo segno del coefficiente di regressione

Variabile	Modello FE1(irr)	Modello FE3(irr1)	Modello FE6(irr1)
Densità di popolazione	-	-	-
Istruzione di terzo livello		+	+
Tassazione	+	+	+
Imprenditorialità femminile	-		-
Imprenditorialità giovanile	-		
Natalità delle imprese			-
Pil pro capite	+		
Dipendenti pubblici	-	-	-
Disoccupazione giovanile	+	+	+
Percezione della criminalità	+	+	
Furti	+		
Rapine	+	+	+
Omicidi	-	-	

Tabella A11: Modello FE1(irr) con procedure robust[54], jackknife e bootstrap

Variabile	Coefficiente di regressione	p-value	p-value robust	p-value jackknife	p-value bootstrap
Densità di popolazione	-0.1726881	0.000	0.000	0.007	0.007
Tassazione	1.828424	0.000	0.005	0.212	0.094
Imprenditorialità femminile	-0.3848288	0.004	0.226	0.365	0.314
Imprenditorialità giovanile	-0.9502937	0.000	0.006	0.037	0.011
Pil pro capite	0.3527608	0.051	0.290	0.418	0.292
Dipendenti pubblici	-0.1486721	0.027	0.109	0.292	0.228
Disoccupazione giovanile	0.0379435	0.037	0.156	0.267	0.163
Percezione della criminalità	0.0424319	0.047	0.125	0.243	0.123
Furti	0.0823281	0.017	0.299	0.467	0.333
Rapine	0.9702864	0.085	0.332	0.842	0.711
Omicidi	-0.3410333	0.045	0.162	0.403	0.208
Dummy2002	1.229681	0.001	0.035	0.198	0.083
Dummy2003	0.2303791	0.483	0.621	0.181	0.072
Dummy2004	-0.8020854	0.002	0.000	0.172	0.065
Dummy2005	-24.33102	Variabile omessa	Variabile omessa	0.190	0.078
Dummy2006	-2.688425	0.000	0.016	0.137	0.043
Dummy2007	-3.798799	0.000	0.018	0.095	0.020
Dummy2008	-3.482437	0.000	0.043	0.156	0.050
Dummy2009	-3.236864	0.000	0.054	0.278	0.137
Dummy2010	-1.64368	0.013	0.176	0.208	0.092
Dummy2011	-1.72572	0.012	0.182	0.214	0.095
Dummy2012	-3.305763	0.000	0.061	Variabile omessa	Variabile omessa
Cons	6.568574	0.614	0.800	0.144	0.049

Risultati dell'elaborazione con il software statistico STATA

[54] Tale opzione viene applicata insieme all'opzione cluster poiché il raggruppamento produce uno stimatore coerente quando i disturbi non sono identicamente distribuiti o vi è una correlazione seriale.

Tabella A12: Modello FE3(irr1) con procedure robust[55], jackknife e bootstrap

Variabile	Coefficiente di regressione	p-value	p-value robust	p-value jackknife	p-value bootstrap
Densità di popolazione	-0.1979079	0.000	0.000	0.000	0.000
Istruzione di terzo livello	0.3120811	0.000	0.007	0.031	0.005
Tassazione	1.003738	0.000	0.001	0.378	0.241
Dipendent ipubblici	-0.0885674	0.097	0.247	0.633	0.528
Disoccupazione giovanile	0.0618856	0.000	0.034	0.061	0.028
Percezione della criminalità	0.0330629	0.075	0.113	0.199	0.130
Rapine	1.926509	0.000	0.055	0.629	0.379
Omicidi	-0.3245098	0.033	0.027	0.095	0.048
Dummy2002	0.4732817	0.082	0.137	0.341	0.204
Dummy2003	-0.1363701	0.606	0.637	0.317	0.181
Dummy2004	-0.6063633	0.007	0.001	0.305	0.167
Dummy2005	-11.70977	Variabile omessa	Variabile omessa	0.338	0.199
Dummy2006	-1.079123	0.001	0.034	0.279	0.137
Dummy2007	-1.497433	0.000	0.029	0.217	0.067
Dummy2008	-1.471014	0.001	0.034	0.214	0.085
Dummy2009	-1.381837	0.004	0.048	0.502	0.353
Dummy2010	-0.4185706	0.269	0.416	0.350	0.224
Dummy2011	-0.3224366	0.392	0.610	0.384	0.249
Dummy2012	-1.590962	0.007	0.150	Variabile omessa	Variabile omessa
cons	16.77847	0.020	0.120	0.273	0.145

Risultati dell'elaborazione con il software statistico STATA

[55] Tale opzione viene applicata insieme all'opzione cluster poiché il raggruppamento produce uno stimatore coerente quando i disturbi non sono identicamente distribuiti o vi è una correlazione seriale.

Tabella A13: Regressione con AR(1) - Coefficienti di regressione, errore standard, t, significatività ed intervalli di confidenza (95%) delle variabili esplicative del modello FE1(irr)

Variabile	Coefficiente di regression	Errore standard	t	p-value	Intervallo di confidenza
Densità di popolazione	-0.0890527	0.0530447	-1.68	0.095	-0.1937221 0.0156168
Tassazione	0.4272536	0.2599928	1.64	0.102	-0.0857722 0.9402795
Imprenditorialità femminile	-0.4177408	0.3154559	-1.32	0.187	-1.040208 0.2047264
Imprenditorialità giovanile	-0.3963066	0.3489966	-1.14	0.258	-1.084957 0.2923442
Pil pro capite	0.6757968	0.1773428	3.81	0.000	0.3258586 1.025735
Dipendenti pubblici	-0.17485	0.0879167	-1.99	0.048	-0.3483299 -0.0013702
Disoccupazione giovanile	0.0022656	0.0150244	0.15	0.880	-0.0273809 0.0319122
Percezione della criminalità	0.0153129	0.0168472	0.91	0.365	-0.0179305 0.0485564
Furti	0.0382991	0.0391037	0.98	0.329	-0.0388616 0.1154598
Rapine	0.024203	0.6326894	0.04	0.970	-1.224239 1.272645
Omicidi	-0.114497	0.1099773	-1.04	0.299	-0.3315076 0.1025136
Dummy2002	0.2458246	0.3808649	0.65	0.519	-0.5057098 0.9973589
Dummy2003	-0.2396288	0.4456436	-0.54	0.591	-1.118986 0.6397289
Dummy2004	-0.3981078	0.4255114	-0.94	0.351	-1.23774 0.4415245
Dummy2005	0.1016572	0.4657318	0.22	0.827	-0.8173391 1.020654
Dummy2006	-0.7396909	0.3895197	-1.90	0.059	-1.508303 0.0289215
Dummy2007	-1.58916	0.4033245	-3.94	0.000	-2.385012 -0.7933074
Dummy2008	-1.33027	0.3604836	-3.69	0.000	-2.041587 -0.6189522
Dummy2009	-0.159552	0.3527005	-0.45	0.652	-.8555116 .5364075
Dummy2010	0.1407186	0.1910451	0.74	0.462	-.2362574 .5176945
Dummy2011	Variabile omessa				
Dummy2012	Variabile omessa				
Cons	17.6421	4.386527	4.02	0.000	8.98647 26.29773

Risultati dell'elaborazione con il software statistico STATA

Tabella A14: Regressione con AR(1) - Coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE3(irr1)

Variabile	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Densità di popolazione	-0.0919009	0.0472762	-1.94	0.053	-0.1851774 0.0013756
Istruzione di terzo livello	0.2027986	0.0760344	2.67	0.008	0.0527818 0.3528153
Tassazione	0.1185143	0.2301832	0.51	0.607	-0.33564 0.5726686
Dipendent ipubblici	-0.1003933	0.0736052	-1.36	0.174	-0.2456172 0.0448306
Disoccupazione giovanile	-0.0028736	0.0121105	-0.24	0.813	-0.0267678 0.0210206
Percezione della criminalità	0.020036	0.0138733	1.44	0.150	-0.0073362 0.0474081
Rapine	0.6610105	0.4992223	1.32	0.187	-0.3239612 1.645982
Omicidi	-0.1442389	0.0890037	-1.62	0.107	-0.3198442 0.0313664
Dummy2002	-0.4851141	0.2609598	-1.86	0.065	-0.9999908 0.0297626
Dummy2003	-1.082004	0.2760726	-3.92	0.000	-1.626699 -0.5373094
Dummy2004	-1.065594	0.2492331	-4.28	0.000	-1.557334 -0.573854
Dummy2005	-0.7443056	0.3157166	-2.36	0.019	-1.367218 -0.1213929
Dummy2006	-0.7241483	0.2905167	-2.49	0.014	-1.297341 -0.1509554
Dummy2007	-0.9996408	0.3044884	-3.28	0.001	-1.6004 -0.3988816
Dummy2008	-0.9839697	0.2772468	-3.55	0.000	-1.530981 -.4369584
Dummy2009	-0.5076975	0.2941421	-1.73	0.086	-1.088043 0.0726482
Dummy2010	-0.2054464	0.1454225	-1.41	0.159	-0.4923667 0.081474
Dummy2011	Variabile omessa				
Dummy2012	Variabile omessa				
cons	24.21955	2.575063	9.41	0.000	19.13892 29.30018

Risultati dell'elaborazione con il software statistico STATA

Tabella A15: Regressione con AR(1) - Coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE6(irr1)

Variabile	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Densità di popolazione	0.0077834	0.0255504	0.30	0.761	-0.0425482 0.058115
Istruzione terzo livello	0.1251445	0.0577127	2.17	0.031	0.0114564 0.2388326
Tassazione	0.2946295	0.1422122	2.07	0.039	0.014486 0.5747729
Imprenditorialità femminile	-0.1052839	0.2074017	-0.51	0.612	-0.5138441 0.3032763
Natalità delle imprese	-0.2167778	0.1029528	-2.11	0.036	-0.4195843 -0.0139714
Dipendenti pubblici	-0.1311252	0.0681367	-1.92	0.055	-0.2653475 0.003097
Disoccupazione giovanile	-0.0065827	0.0093911	-0.70	0.484	-0.0250821 0.0119167
Rapine	0.4468287	0.4387866	1.02	0.310	-0.417536 1.311193
Dummy2002	-0.4728457	0.242294	-1.95	0.052	-0.95014 0.0044487
Dummy 2003	-1.26976	0.3029665	-4.19	0.000	-1.866573 -0.6729468
Dummy2004	-1.260987	0.3097642	-4.07	0.000	-1.87119 -0.6507829
Dummy2005	-0.7879589	0.391376	-2.01	0.045	-1.55893 -0.0169882
Dummy2006	-1.124134	0.3268697	-3.44	0.001	-1.768033 -0.4802338
Dummy2007	-1.203114	0.3576224	-3.36	0.001	-1.907593 -0.4986341
Dummy2008	-1.492914	0.3232986	-4.62	0.000	-2.129779 -0.8560484
Dummy2009	-1.223275	0.297038	-4.12	0.000	-1.808409 -0.6381405
Dummy2010	-0.9319337	0.3077928	-3.03	0.003	-1.538254 -0.3256135
Dummy2011	-0.6864949	0.2944926	-2.33	0.021	-1.266615 -.1063745
Dummy2012	-0.7021757	0.1954538	-3.59	0.000	-1.0872 -0.3171517
Dummy2013	-.5719434	0.1315695	-4.35	0.000	-0.8311219 -0.3127649
Dummy2014	Variabile omessa				
Dummy2015	Variabile omessa				
Cons	8.76134	1.563839	5.60	0.000	5.680737 11.84194

Risultati dell'elaborazione con il software statistico STATA

Tabella A16: Test F e Test di Breusch-Pagan

Modello	**Test F**	**Test Breusch- Pagan**
Modello 1	F(19,205)= 33,16= 186,93; Prob> F = 0.0000	χ^2= 131,06; Prob> χ^2 = 0.0000
Modello 2	F(19,206)= 52,54; Prob> F= 0.0000	χ^2= 289,88; Prob> χ^2 = 0.0000
Modello3	F(19,207)= 46,40; Prob> F= 0.0000	χ^2= 429,81; Prob> χ^2 = 0.0000

Tabella A17: Test di Hausmann e Test di Wooldridge

Modello	Test di Hausmann	Test di Wooldridge
Modello 1	χ^2= 80,36; Prob> χ^2 = 0.0000	χ^2= 135,677; Prob> χ^2 = 0.0000
Modello 2	χ^2= 138,57; Prob> χ^2 = 0.0000	χ^2= 172,544; Prob> χ^2 = 0.0000
Modello 3	χ^2= 152,42; Prob> χ^2 = 0.0000	χ^2= 144,455; Prob> χ^2 = 0.0000

BIBLIOGRAFIA CAPITOLO 3

Alleva, G. (2017). Indagine conoscitiva sulle politiche in materia di parità tra donne e uomini. ISTAT

Alleva, G., Pericolini, F. (2004). L'interpretazione del tasso regionale di irregolarità del lavoro come proxy dell'economia sommersa in Italia: modelli statistici e stimatori panel. 1° Convegno italiano degli utenti di STATA, Roma 25 ottobre 2004

Amendola, A., e Dell'Anno, R. (2008). Istituzioni, Disuguaglianza ed Economia Sommersa: quale relazione? Quaderno n. 24/2008, Dipartimento di Scienze Economiche, Matematiche e Statistiche; Università degli Studi di Foggia.

Baltagi, B.H. (2005). Econometric Analysis of Panel data. John Wiley and Son. Third Edition

Busato, F., e Chiarini, B. (2004). Market and Underground Activities in a Two Sector Dynamic Equilibrium Model. EconomicTheory, 23(4): 831-861.

Cameron, A. C., e Trivedi P. K.(2010). Microeconometrics Using Stata. Rev. ed. College Station, TX: Stata Press.

Cappariello, R., e Zizza, R. (2009). Dropping the Books and Working Off the Books. Temi di discussione, WorkingPapers, Ufficio Studi Banca d'Italia, n. 702.

Cochrane, D. e Orcutt, G.H. (1949). Application of Least Squares Regression to Relationships Containing Auto-Correlated Error Terms. Journal of the American Statistical Association. Volume 44, 1949 - Issue 245

Daniele, V., e Marani, U. (2008). Criminalità e investimenti esteri. Un'analisi per le province italiane. WorkingPaper, Università Magna Graecia di Catanzaro.

Dell'Anno, R. (2003). Stimare l'economia sommersa con un approccio ad equazioni strutturali. Un'applicazioneall'economiaitaliana (1962-2000). SIEP.

Friedman, D.A. (2006). On the so-called "Huber sandwich Estimator" and "Robust Standard Error". Department of Statistics – University of California - Berkeley

Frey, B. - Weck-Hanmeman, H. (1984), "The hidden economy as an unobserved variable". EuropeanEconomicReview n. 26/1.

ISTAT (2017). Imprenditorialità nelle regioni italiane. Caratteri strutturale e socio-economici.

ISTAT, 2018. Economia non osservata nei conti nazionali. Anni 2013-2016.

Lisi, G. (2009). Underground Employment and Unemployment in the Regions of Italy: A panel analysis. University of Cassino (Italy). MPRA Paper No. 18525.

Lisi, G. (2010). Underground Employment and Unemployment in Italy: A panel analysis. University of Cassino (Italy). MPRA Paper No. 22508.

Lucifora, C. (2003). Economia sommersa e lavoro nero. Il Mulino.

Marini. D., e Turato, F. (2002). Nord-Est e Mezzogiorno: nuove relazioni, vecchi stereotipi. Rapporti Formez-Fondazione Nord-Est, aprile.

Medina, L. e Schneider, F. (2017). Shadow Economies Around the World: What Did We Learn Over the Last 20 Years? WP/18/17 – IMF Working Paper

Morvillo, C. (2016). Evoluzione delle determinanti dell'economia sommersa: analisi panel di regioni italiane. Ministero dell'Economia e delle Finanze. Dipartimento del Tesoro. Nota tematica 1/2016.

Schneider, F. (2004). The Size of the Shadow Economies of 145 Countries all over the World: First Results over the Period 1999 to 2003. Papers, No. 1431, Institute for the Study of Labor (IZA), Bonn.

Schneider, F. (2005). Shadow Economies around the World: What do we really know? IAW Diskussionspapiere, No. 16, Institut für Angewandte Wirtschaftsforschung (IAW), Tübingen

Schneider, F. (2013). Size and Development of the Shadow Economy of 31 European and 5 other OECD Countries from 2003 to 2013: A Further Decline.

Schneider, F. (2013). Size and Development of the Shadow Economy of 31 European and 5 other OECD Countries from 2003 to 2012: Some New Facts. ResearchGate

Schneider, F. (2013). The Shadow Economy in Europe.ATKearny.

Schneider, F. e Williams, C. (2013). The Shadow Economy.The Institute of Economic Affairs.

Schneider, F. e Buehn, C. (2016). Estimating the Size of the Shadow Economy: Methods, Problems and Open Questions. Discussion Paper No. 9820.The Institute for the Study of Labor (IZA) in Bonn.

Soliani, L. (2005). Manuale di statistica per la ricerca e la professione. Statistica uni variata e bivariata parametrica e non parametrica per le discipline ambientali e biologiche – Edizione aprile 2005. Università di Parma. http://www.dsa.unipr.it/soliani/soliani.html
Zizza, R. (2002). Metodologie di stima dell'economia sommersa: un'applicazione al caso italiano. Banca d'Italia. Temi di discussione del Servizio Studi.

4. Il caso europeo – Analisi per Stati Membri

L'Unione europea (UE) è un'unione politica ed economica di Stati che comprende 28 Paesi europei: Austria, Belgio, Bulgaria, Cipro, Croazia, Danimarca, Estonia, Finlandia, Francia, Germania, Grecia, Irlanda, Italia, Lettonia, Lituania, Lussemburgo, Malta, Paesi Bassi, Polonia, Portogallo, Regno Unito, Repubblica Ceca, Romania, Slovacchia, Slovenia, Spagna, Svezia, Ungheria. La sua origine risale al secondo dopoguerra, quando sei Stati (Belgio, Francia, Germania, Italia, Lussemburgo e Paesi Bassi) istituirono la Comunità Economica Europea (CEE) per rafforzare la cooperazione economica nel continente europeo. Da allora, altri 22 membri hanno aderito formando un enorme mercato unico. Quella che era nata come un'unione puramente economica è diventata col tempo un'organizzazione attiva in tutta una serie di settori che vanno dal clima all'ambiente, alla salute, alle relazioni esterne e alla sicurezza, alla giustizia e all'immigrazione. Per sottolineare questo cambiamento, nel 1993 il nome di Comunità Economica Europea (CEE) è stato sostituito da Unione europea (UE).

In questo approfondimento viene considerato anche il Regno Unito, sebbene dal 31.01.2020 abbia cessato di essere uno Stato membro dell'Unione europea, in quanto i dati utilizzati arrivano fino a tutto il 2015. L'uscita del Regno Unito dall'Unione europea, nota anche come Brexit (sincrasi formata dall'inglese Britain, "Gran Bretagna", ed exit, "uscita"), è stato il processo che ha posto fine all'adesione del Regno Unito all'Unione europea, secondo le modalità previste dall'articolo 50 del Trattato sull'Unione europea, come conseguenza del referendum sulla permanenza del Regno Unito nell'Unione europea.

L'UE ha introdotto una moneta unica europea, l'euro, usata da oltre 340 milioni di cittadini dell'UE in 19 Paesi: Austria, Belgio, Cipro, Estonia, Finlandia, Francia, Germania, Grecia, Irlanda, Italia, Lettonia, Lituania, Lussemburgo, Malta, Paesi Bassi, Portogallo, Slovacchia, Slovenia e Spagna.

4.1 I dati sull'Economia Sommersa

I dati sull'Economia Sommersa, a livello europeo, sono stati desunti dai lavori di Friedrich Schneider[56] (Medina e Schneider, 2017). La metodologia utilizzata per ottenere tali dati è l'approccio MIMIC (Multiple Indicators, Multiple Causes), introdotta nel Capitolo 1. Consiste in un particolare tipo di modellazione di equazioni strutturali (SEM) (Joreskog e Goldberger, 1975). Nel modello MIMIC viene inizialmente stabilito un approccio teorico che spiega la relazione tra le variabili esogene e la variabile latente (che in questa applicazione è l'Economia Sommersa), per poi studiare l'effetto della variabile latente (Economia Sommersa) sulle variabili dell'indicatore macroeconomico. Pertanto, il modello MIMIC è considerato un metodo di conferma piuttosto che un metodo esplicativo. Esso è costituito da due modelli:

1. **un modello strutturale** che descrive la relazione esistente tra la variabile latente, Economia Sommersa, e le sue cause, formalizzato dalla seguente espressione:

$$\eta=\Gamma X+\xi$$

dove η è la variabile latente Economia Sommersa, X è il vettore (qx1) delle cause dell'Economia Sommersa elencate nello schema sottostante, Γ è la matrice dei coefficienti (1xq) delle cause dell'Economia Sommersa, ξ è il termine di errore, q è un valore pari al numero delle cause dell'Economia Sommersa che vengono considerate;

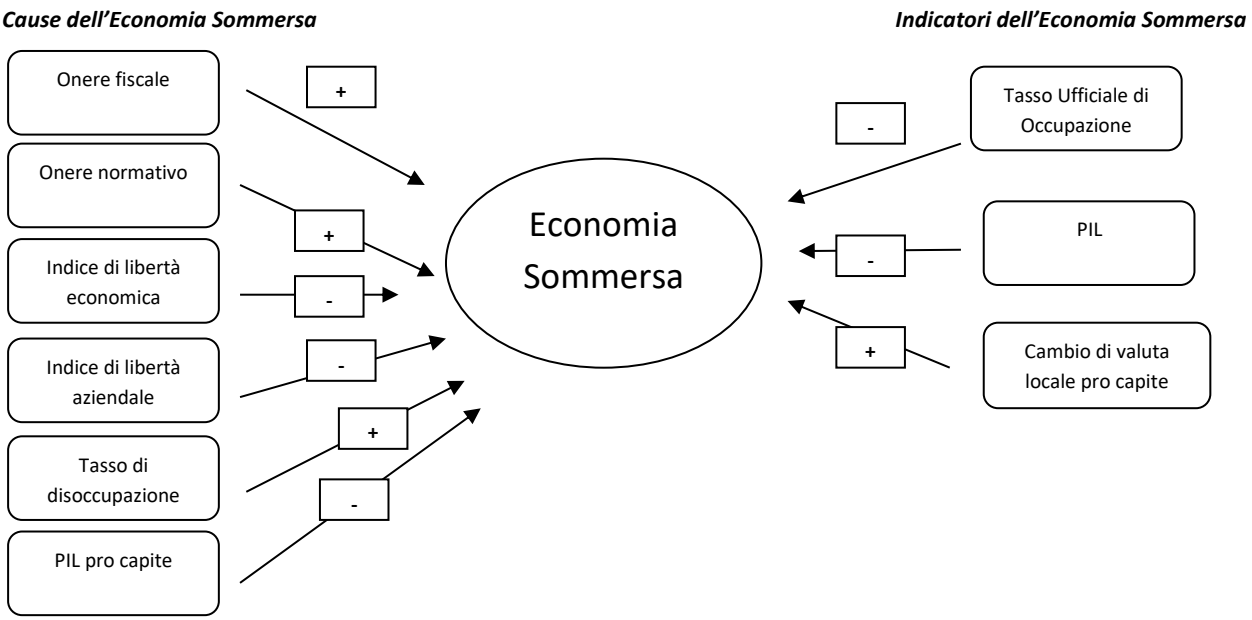
[56] Professore di economia presso l'Università Johannes Kepler di Linz in Austria e dal 2006 ricercatore presso l'Istituto Tedesco di Ricerche Economiche.

2. **un modello di misurazione** che collega la variabile latente Economia Sommersa ai suoi indicatori, espresso formalmente dalla seguente espressione:

$$Y=\Lambda_Y\eta+\varepsilon$$

dove Y è il vettore di indicatori (px1), Λ_Y è la matrice dei coefficienti (px1), η è la variabile latente Economia Sommersa, ε è un vettore (px1) di componenti di errori e p è un numero pari al numero di indicatori utilizzati.

Schema del modello MIMIC di Friedrich Schneider (Medina e Schneider, 2017)



Per ottenere i valori dell'Economia Sommersa si procede secondo i seguenti due step:

1. Nel primo step l'Economia Sommersa rimane un fenomeno non osservato (variabile latente) che viene stimato utilizzando cause di comportamento illecito, come ad esempio l'onere fiscale e l'onere normativo o l'intensità della regolamentazione, e indicatori che riflettono attività illecite, come ad esempio la domanda di valuta (indice di libertà economica) e l'orario di lavoro ufficiale (indice di libertà aziendale). Questa procedura "produce" solo stime relative della dimensione dell'Economia Sommersa.
2. Nel secondo step il metodo della domanda di valuta (Tanzi, 2013) viene utilizzato per calibrare le stime relative in stime assolute utilizzando i valori assoluti del metodo della domanda di valuta[57] come valori iniziali per l'Economia Sommersa.

4.2 Analisi descrittiva della variabile oggetto di studio

Si procede all'analisi attraverso i seguenti passi[58]:

A. analisi descrittiva e grafica della variabile dipendente;

B. definizione di un modello guideline;

[57]Il metodo della valuta è stato ideato da Cagan, ripreso da Gutmann e infine perfezionato da Tanzi (Tanzi, 2013). Questo metodo, che stima la domanda di moneta circolante rispetto alla domanda di depositi a vista, è sicuramente uno dei metodi più conosciuti (Capitolo 1).
[58]Tutte le elaborazioni sono state sviluppate con il software statistico di elaborazione dati STATA.

C. analisi della correlazione e dell'indice VIF per la verifica di eventuali multicollinearità;

D. confronto dei diversi modelli panel analizzati e scelta del modello più idoneo;

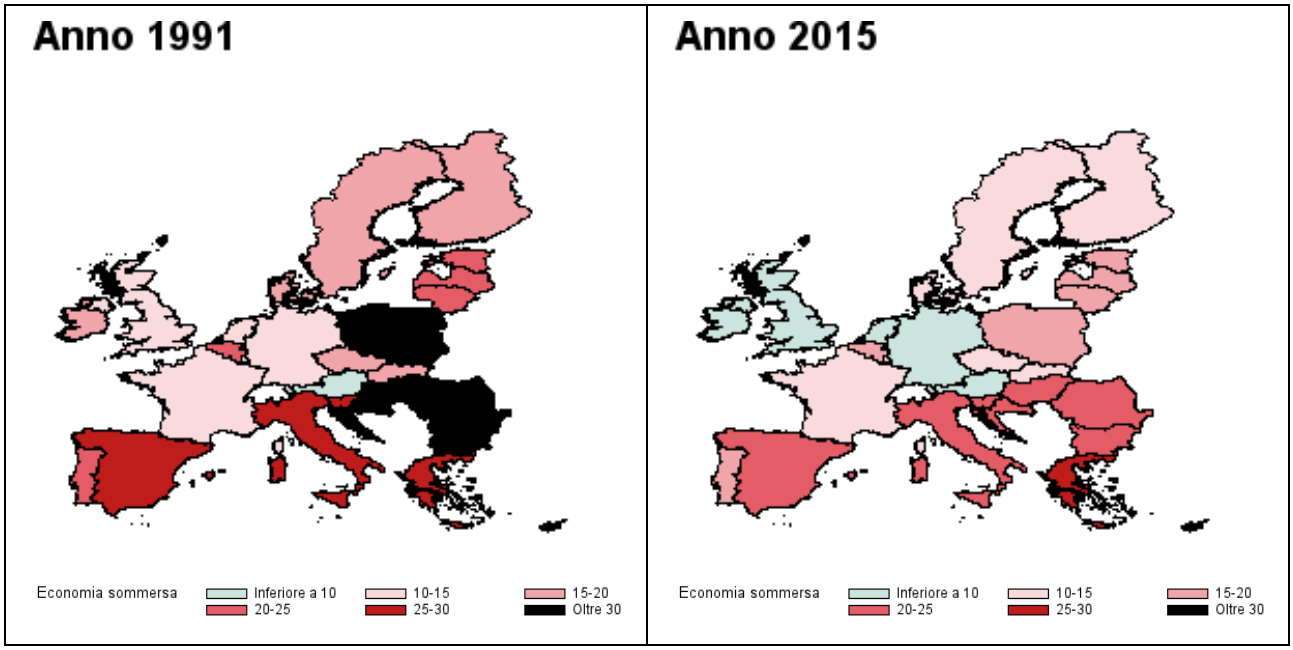
E. affinamento del modello scelto;

F. test per la diagnosi del modello (omoschedasticità e normalità).

In questo paragrafo si riporta un'analisi esplorativa del fenomeno dell'Economia Sommersa così come definito nel paragrafo precedente. I dati sono disponibili dal 1991 al 2015 ed espressi in percentuali di PIL (Tabella B1 in appendice).

Una prima rappresentazione con cartografia fornisce l'immediata contezza della diffusione del fenomeno nell'Unione europea, per ogni singolo anno (Figura 1 e Figura B1 in appendice). I Paesi con più alto livello di Economia Sommersa, sia all'inizio che alla fine del periodo, sono Cipro, Romania, Malta e Croazia, mentre i Paesi con il più basso livello sono Austria, Lussemburgo, Paesi Bassi, Regno Unito e Germania. In particolare, l'Austria rimane sempre e costantemente a un livello bassissimo (<10% di PIL) e nel corso del tempo le si affiancano la Germania, l'Irlanda, i Paesi Bassi e il Regno Unito; mentre a un livello altissimo (oltre 30% di PIL) inizialmente abbiamo molti Paesi della zona dell'Est Europa che sono entrati nell'UE dopo il 2004 (Bulgaria, Cipro, Croazia, Polonia, Slovacchia e Ungheria). Alla fine del periodo (2015) solo Cipro ha un valore superiore al 30% (32,2%).

Figura 1: Rappresentazione cartografica dell'Economia Sommersa (% PIL) per Paese membro dell'Unione europea, Anni 1991 e 2015

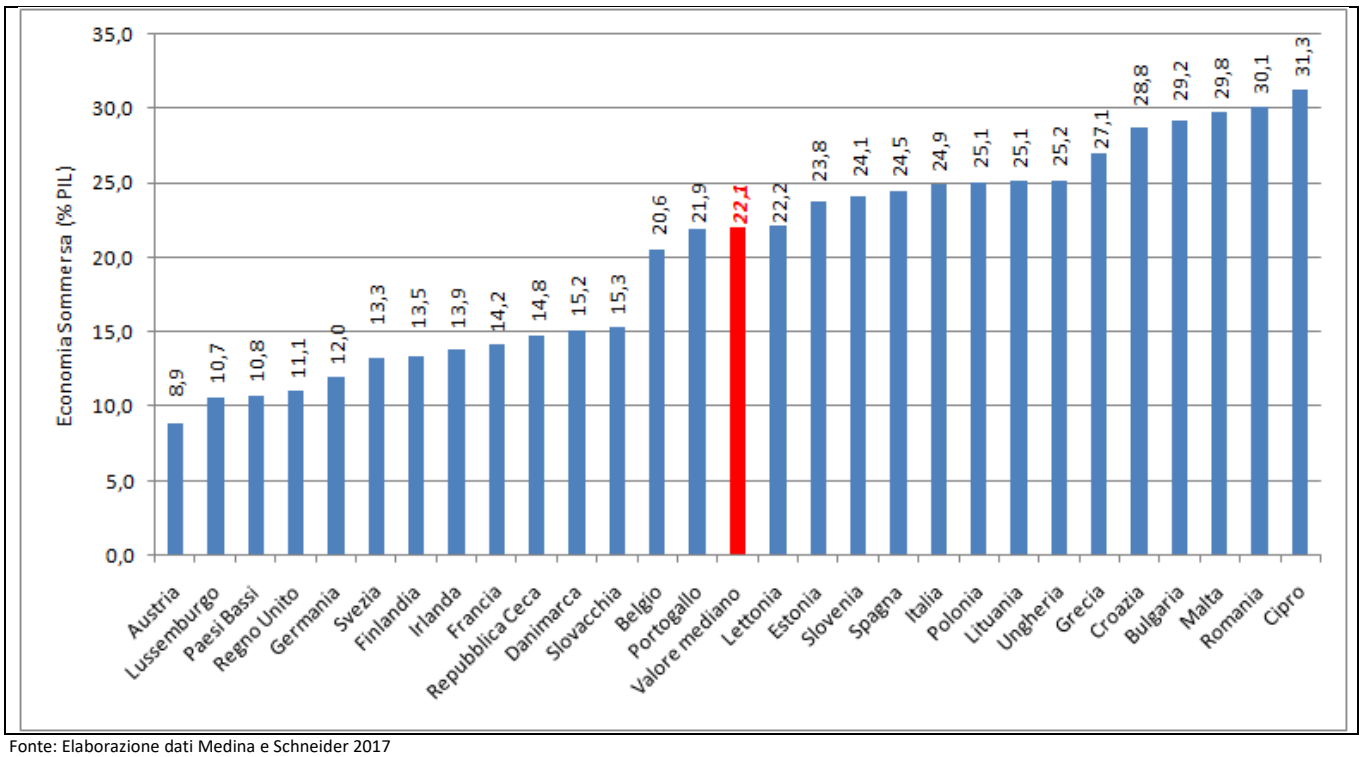


Fonte: Elaborazionedati Medina e Schneider 2017 - © EuroGeographics per i confini amministrativi

Quanto emerge dalle cartografie, viene confermato dal diagramma a barre (Figura 2), nel quale è rappresentato il valore medio dell'Economia Sommersa per il periodo a disposizione (1991-2015) espresso come percentuale del PIL.

I Paesi UE con minore Economia Sommersa sono quelli che fanno parte dell'Unione europea da più tempo, ossia dalla sua costituzione (1957) fino al 1995, ad esclusione della Repubblica Ceca e della Slovacchia; i Paesi con maggiore Economia Sommersa sono quelli che hanno aderito all'Unione europea recentemente, ossia dal 2004, eccetto Spagna, Italia e Grecia.

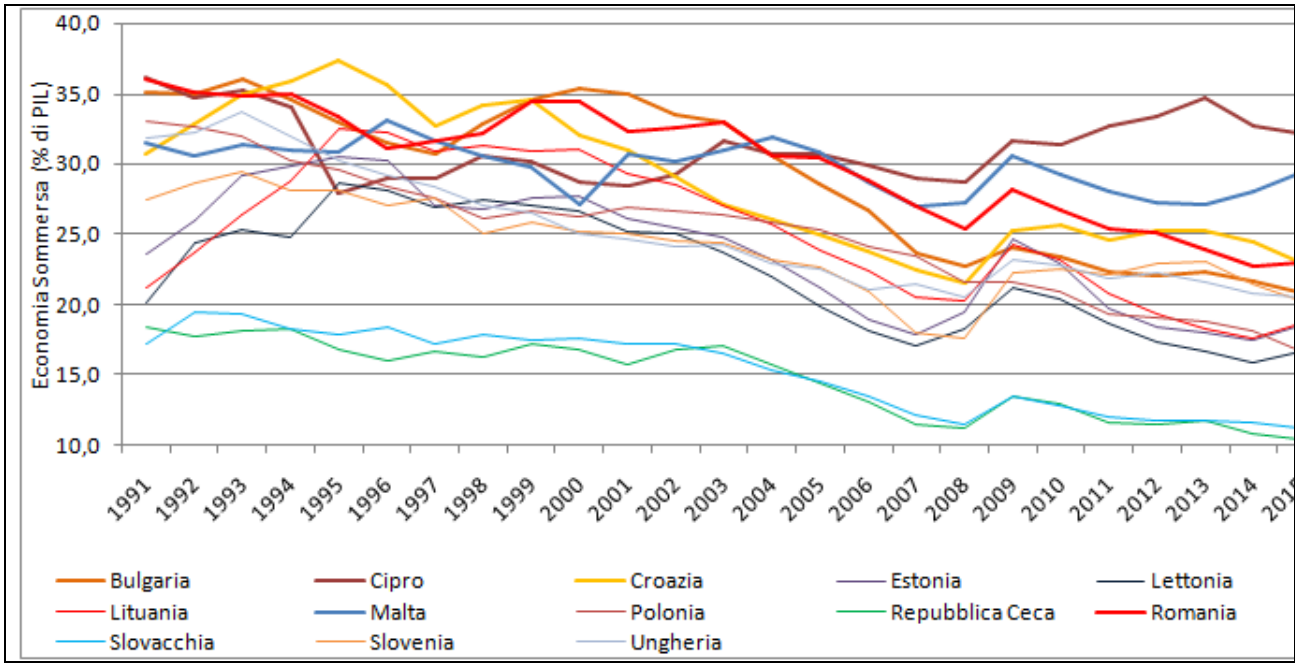
Figura 2: Economia Sommersa per Paese membro (% sul PIL), Valore medio periodo 1991-2015



Fonte: Elaborazione dati Medina e Schneider 2017

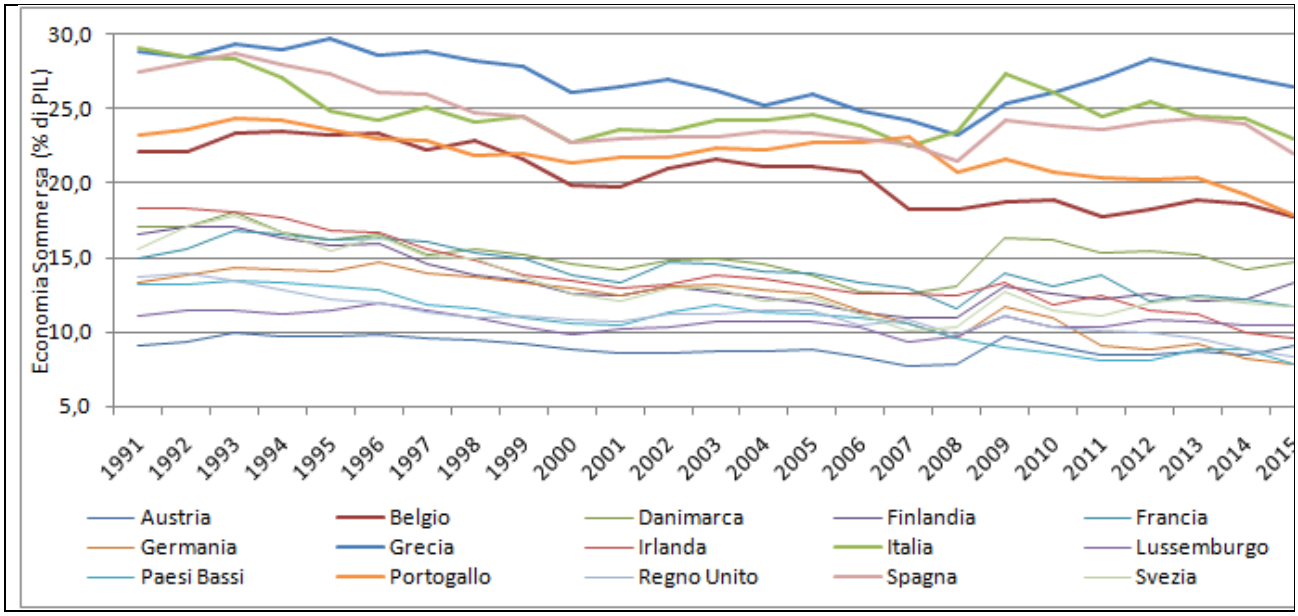
I risultati appena emersi e la letteratura economica esaminata (Achim, Borlea, Gaban e Cuceu, 2018; Mahadeo et al., 2012; Goez, 2001; Falkner e Trieb, 2008) suggeriscono di affiancare lo studio per tutti i 28 Paesi membri dell'UE, distinguendo l'analisi per i 15 Paesi appartenenti all'Unione europea dalla sua costituzione al 1995 (OLD UE) e per i 13 Paesi che hanno aderito dal 2004 (NEW UE) (Tabella B2 in appendice). Infatti, i due gruppi possiedono caratteristiche specifiche diverse tra loro. Innanzitutto, hanno differenze significative in termini di sviluppo economico, a tal punto che i Paesi appartenenti al gruppo OLD UE (nella maggior parte dei casi) vengono ritenuti dagli economisti "sviluppati", mentre i restanti (NEW UE) vengono definiti emergenti. Ciò perché gli studi sopra citati hanno reso evidente come una serie di fattori, quali ad esempio la privatizzazione, il contesto legislativo e l'orientamento nelle esportazioni, si sviluppano diversamente nei due gruppi. Gli economisti (Achim, Borlea, Gaban e Cuceu, 2018; Mahadeo et al., 2012; Goez, 2001; Falkner e Trieb, 2008), inoltre, sostengono che il processo di europeizzazione viene applicato in modo differente nei due insiemi, in quanto il livello di adeguamento alle regole europee è più consolidato nei Paesi OLD UE, mentre è ancora incerto nei Paesi NEW UE.

Figura 3: Andamento dell'Economia Sommersa nel periodo 1991-2015 nei Paesi NEW UE



Fonte: Elaborazione dati Medina e Schneider 2017

Figura 4: Andamento dell'Economia Sommersa nel periodo 1991-2015 nei Paesi OLD UE



Fonte: Elaborazione dati Medina e Schneider 2017

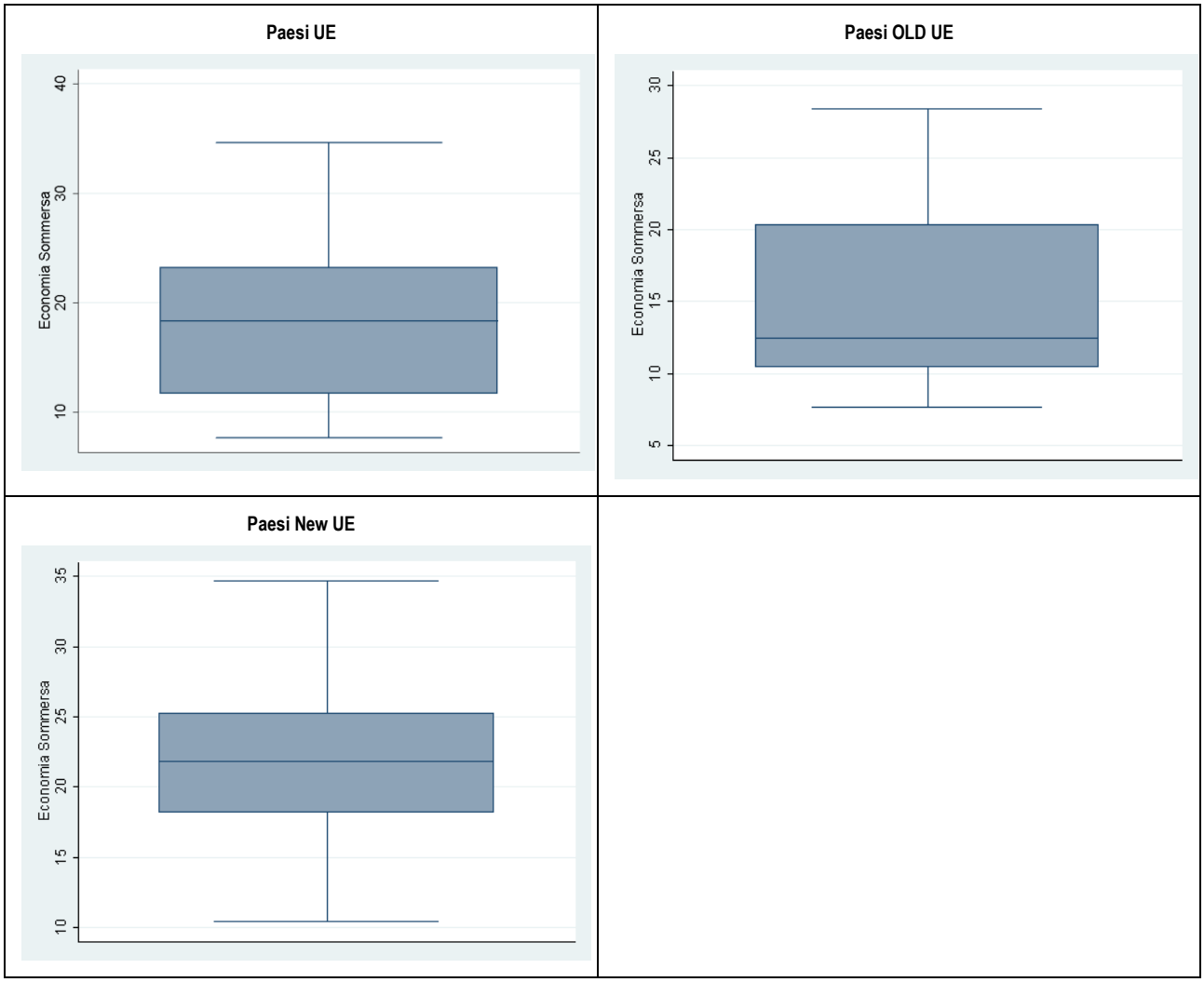
Il confronto grafico tra i due gruppi NEW UE e OLD UE (Figura 3 e Figura 4), mostra come i Paesi del primo gruppo presentino valori dell'Economia Sommersa più elevati (compresi tra il 10 e il 40 per cento di PIL) rispetto al secondo gruppo (compresi tra il 5 e il 30 per cento di PIL). In entrambi i casi, il fenomeno ha un andamento decrescente, meno spiccato per alcuni Paesi OLD UE (Grecia, Lussemburgo, Austria).

Procedendo con l'analisi grafica della variabile risposta, effettuata attraverso i box plot, si nota una distribuzione asimmetrica. Tale caratteristica viene esplorata con una analisi grafica per anno, per Paese e sopra l'ottantesimo percentile[59], che conferma l'asimmetria, evidenzia la presenza di alcuni valori anomali e

[59]Si veda la Figura B2 in appendice.

la maggiore variabilità del fenomeno per Paese UE. In particolare, tutti i 28 Paesi Ue mostrano una asimmetria positiva, con indice di asimmetria[60] pari a 0,21, i 15 Paesi OLD UE una asimmetria positiva (indice di asimmetria 0,72), mentre i 13 Paesi NEW UE evidenziano una lieve asimmetria negativa (indice di asimmetria -0,06). Nei primi due casi, Paesi UE e Paesi OLD UE, vi è una maggiore frequenza dei valori medio-bassi e un conseguente spostamento verso il basso della scatola.

Figura 5: Rappresentazione grafica (box plot) della variabile dipendente Economia Sommersa



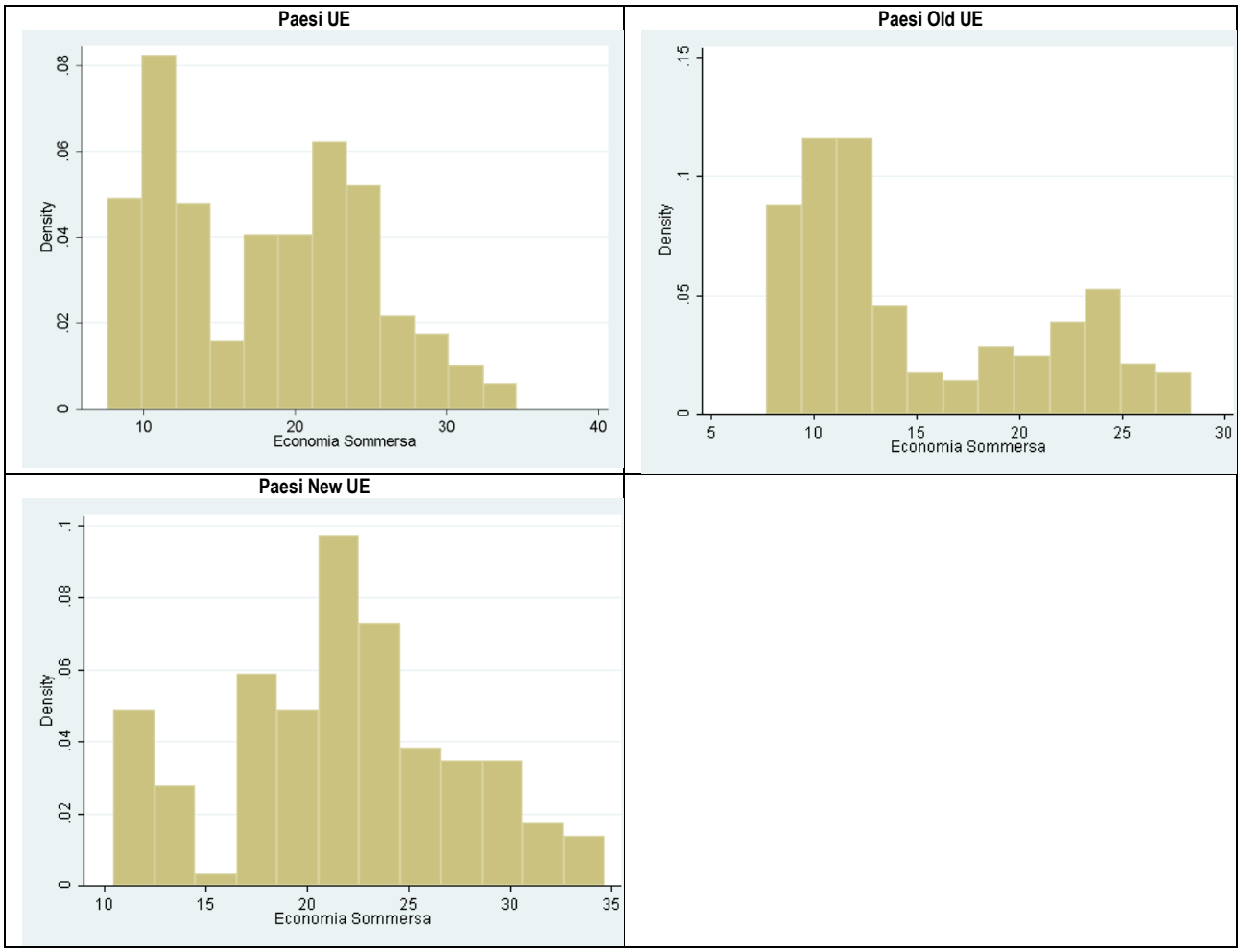
Risultati dell'elaborazione con il software statistico STATA

Continuando lo studio descrittivo con una analisi grafica tramite istogrammi si nota una tendenza alla bimodalità. Tale particolarità viene investigata attraverso un approfondimento con istogrammi al livello di Paese e per anno[61], dal quale si desume una scarsa variabilità nel tempo dell'indicatore e una variabilità più spiccata per Paese UE.

[60] L'indice di asimmetria utilizzato è $m_3.m_2^{-3/2}$.

[61] Si veda la Figura B3 e B4 in appendice.

Figura 6: Istogrammi dell'Economia Sommersa



Risultati dell'elaborazione con il software statistico STATA

4.3 I modelli panel statici per lo studio dell'Economia Sommersa

La scelta delle variabili esplicative da includere nell'analisi è stata suggerita dallo studio della letteratura economica in materia di Economia Sommersa[62] e dalla disponibilità dei dati reperibili on line (EUROSTAT, OCSE, AMECO):

- **Struttura socio-demografica.** Fanno parte di questo gruppo la densità di popolazione (**dens**), utile a definire l'eterogeneità della distribuzione della popolazione in ambito europeo, il tasso di partecipazione all'istruzione tra il secondo e terzo livello (**istr**) e di terzo livello (**istr1**), che hanno generalmente un ruolo di contrasto rispetto al fenomeno in esame, sebbene avere un livello di istruzione troppo elevato garantisce sicuramente delle migliori opportunità di lavoro ma non sempre maggiori possibilità (Cappariello e Zizza, 2009; Lisi, 2010);
- **Struttura economica per Paese membro**. Per sintetizzare il contesto economico-istituzionale dei Paesi membri, vengono considerate le entrate tributarie da tassazione diretta e indiretta (**taxdiretta** e **taxindiretta**) (Amendola e Dell'Anno, 2008; Dell'Anno, 2003) e il tasso di industrializzazione (**indu**),

[62] La letteratura di riferimento è riportata nel Capitolo 3

utile per verificare l'ipotesi che nei Paesi in cui la dotazione di industrie è particolarmente carente ci sia una maggior diffusione di Economia Sommersa (Daniele e Marani, 2008);

- **Variabili di controllo**. Rappresentano questo gruppo il PIL pro capite (**pilpc**), in grado di fornire un'indicazione della dimensione della crescita economica locale (Dell'Arno, 2003; Busato e Chiarini, 2004), la partecipazione femminile al mercato del lavoro (**fem**), la disoccupazione giovanile (**disocgiov**) (Lucifora, 2003) e l'intensità della regolamentazione, indicatore espresso in questo studio dal rapporto tra il numero dei dipendenti dell'amministrazione centrale di ogni Paese e il numero di lavoratori nello stesso Paese (**dippub**), adeguato a fornire una fotografia del quadro istituzionale europeo (Frey e Weck-Hanmeman, 1984 ; Zizza, 2002; Morvillo, 2016).
- **Qualità della vita**. In questo approfondimento si tiene conto anche di un filone della letteratura economica (Achim, Borlea, Gaban, Cuceu, 2015; Schneider e Klinglmair, 2004; Bergheim, 2007; Thieben, 2010) che ha lo scopo di studiare l'incidenza del benessere e della felicità degli individui sul livello dell'Economia Sommersa. Secondo questi approfondimenti se le persone sono più felici e stanno bene sono più propense ad avere rispetto della legge e quindi a pagare le tasse. Tali studi sono iniziati con Frey e Stutzer (2012) e hanno evidenziato come la felicità possa influenzare decisioni economiche molto importanti come il lavoro, il consumo di beni e servizi e gli investimenti. Alcuni autori ritengono infatti che l'aspetto psicologico delle persone sia la determinante più importante dell'Economia Sommersa (Voicu, 2012). Concetti quali la tax moral e la tax compliance[63], introdotti da Schmolders nel 1960, sono stati accentuati dalla letteratura economica recente, sottolineando l'importanza della situazione morale e delle condizioni sociali degli individui per spiegare il comportamento degli stessi nei confronti delle attività nascoste (OECD, 2012). Nel dataset EUROSTAT sono stati scelti due indicatori in grado di rappresentare al meglio le variabili esplicative del benessere e della felicità: il numero di reati criminali (**criminalità**) e la situazione abitativa (**abitazione**).

[63]Il concetto di tax morale è legato al sentimento di consapevolezza civica diffusa tra i cittadini, e alla motivazione a pagare le tasse. Il concetto di tax compliance è legato a quanto i contribuenti si conformano o non alle norme fiscali del loro proprio Paese, dichiarando correttamente i loro redditi e pagando l'imposta dovuta nei termini di legge. Pertanto, nel primo caso vi è la motivazione a rispettare le leggi (attitude), mentre nel secondo caso viene messa in risalto la conformità del comportamento (behavior).

Tabella1: Elenco dettagliato delle variabili utilizzate e delle fonti utilizzate

Sigla		Descrizione	Fonte	Segno atteso
ES	***Variabile oggetto di studio***	Economia Sommera	Medina e Schneider (2017), (1991-2015)	
Dens	***Struttura demografica dei Paesi UE***	Densità abitativa calcolata come rapporto tra numero di residenti e superficie territoriale in kmq	EUROSTAT (2000-2015)	**-**
Istr		Tasso di partecipazione all'istruzione di terzo livello	EUROSTAT (2000-2015)	**-/+**
istr1		Tasso di partecipazione all'istruzione tra il secondo e il terzo livello	EUROSTAT (2000-2015)	**-/+**
Indu	***Struttura economica dei Paesi UE***	Indice del volume di produzione (attività estrattiva; attività manifatturiere; fornitura di energia elettrica, gas, vapore e aria condizionata; costruzioni) (2015=100)	EUROSTAT, (2000-2015)	**-**
Taxdiretta		Quota delle imposte dirette sul prodotto interno lordo	EUROSTAT (2005-2015)	**+**
Taxindiretta		Quota delle imposte indirette sul prodotto interno lordo	EUROSTAT (2005-2015)	**+**
Fem	***Variabili di controllo***	Tasso di occupazione femminile pari al rapporto tra il numero di donne occupate in età 15-64 anni sulla popolazione	EUROSTAT, (2000-2015)	**+**
Pilpc		Prodotto Interno Lordo ai prezzi di mercato (prezzi correnti) per abitante	EUROSTAT, (2000-2015)	**+/-**
Dippub		Rapporto tra i dipendenti della pubblica amministrazione centrale e le forze di lavoro in età 15-64 anni	EUROSTAT (2004-2015)	**-**
Disocgiov		Percentuale Persone in cerca di occupazione in età 15-24 anni su forze di lavoro stessa classe di età	EUROSTAT, (2000-2015)	**+**
Criminalità	***Qualità della vita***	Criminalità, violenze e vandalismo (EU-SILC survey)	EUROSTAT, (2003-2015)	**+**
Abitazione		Percentuale della popolazione che abita in situazioni disagiate sulla popolazione totale (EU-SILC survey)	EUROSTAT, (2003-2015)	**+**

Di seguito vengono esposte le risultanze dell'approfondimento svolto sui dati a disposizione. Il database è costituito da un panel bilanciato relativo ai 28 Paesi UE, composto dalla variabile dipendente, ES, che rappresenta l'Economia Sommersa, disponibile per l'arco temporale 1991-2015. Le variabili esplicative a

disposizione sono 12 (Tabella 1) tutte disponibili per archi temporali diversi. Pertanto, si prende come riferimento l'arco temporale 2005-2015, periodo disponibile per tutte le variabili considerate.

B. Definizione del modello guideline

Si procede inizialmente alla stima del modello pooled OLS, inserendo tutte le variabili a disposizione. Come anticipato nel Capitolo 2, il modello pooled OLS nella maggior parte dei casi è improbabile che sia adeguato, ma offre una *guideline* per un confronto con modelli più complessi.

Sulla base delle considerazioni emerse nel paragrafo 4.2, vengono studiati quattro modelli tutti per l'arco temporale 2005-2015, ma con un numero diverso di Paesi:

- 28 Paesi UE (UE);
- 15 Paesi UE appartenenti all'Unione europea dalla sua costituzione (OLD UE);
- 13 Paesi UE appartenenti all'Unione europea dal 2004 (NEW UE);
- 19 Paesi appartenenti all'Area dell'Euro (AE).

Indicando con:

Y_{it} la variabile dipendente, dove con i indichiamo il Paese UE e con t l'anno;

D_t le dummy temporali;

α_i l'effetto individuale;

ε_{it} l'errore residuo;

X_{hit} le p covariate (h=1,…,p) nella regione i e nell'anno t, così definite:

Variabile	Formalizzazione della covariata
Densità di popolazione	X_{1it}
Istruzione di secondo livello	X_{2it}
Istruzione di terzo livello	X_{3it}
Industrializzazione	X_{4it}
Tassazione diretta	X_{5it}
Tassazione indiretta	X_{6it}
Occupazione femminile	X_{7it}
Pil pro capite	X_{8it}
Dipendenti pubblici	X_{9it}
Disoccupazione giovanile	X_{10it}
Criminalità	X_{11it}
Abitazione	X_{12it}

La formalizzazione dei modelli è data dalle seguenti espressioni:

Modello UE: $Y_{it}=\Sigma_h\beta_h X_{hit}+ D_t+\alpha_i+ \varepsilon_{it}$ i=1,…..,28; t=1,….,11; h=1,…,12

Modello OLD UE: $Y_{it}=\Sigma_h\beta_h X_{hit}+ D_t+\alpha_i+ \varepsilon_{it}$ i=1,…..,15; t=1,….,11; h=1,…,12

Modello NEW UE: $Y_{it}=\sum_h \beta_h X_{hit}+ D_t+\alpha_i+ \varepsilon_{it}$ i=1,…..,13; t=1,….,11; h=1,…,12

Modello AE: $Y_{it}=\sum_h \beta_h X_{hit}+ D_t+\alpha_i+ \varepsilon_{it}$ i=1,…..,19; t=1,….,11; h=1,…,12

Dopo aver applicato ai modelli sopraindicati i passi C-F[64], si ottengono i seguenti quattro risultati[65]:

MODELLO UE

Tabella 2: Coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE UE2[66]

Variabile	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Abitazione	0.0717061	0.0132637	5.41	0.000	0.0455892 0.0978231
Densità di popolazione	0.0659411	0.0090143	7.32	0.000	0.0481914 0.0836907
Disoccupazione giovanile	0.0587464	0.0156699	3.75	0.000	0.0278914 0.0896014
Occupazione femminile	-0.1595814	0.04313	-3.70	0.000	-0.244507 -0.0746558
Industrializzazione	-0.0644933	0.0074376	-8.67	0.000	-0.0791385 -0.0498482
Istruzione di terzo livello	-0.0953173	0.0363544	-2.62	0.009	-0.1669012 -0.0237333
Istruzione di secondo livello	-0.2439303	0.342482	-7.12	0.000	-0.3113671 -0.1764934
Pil pro capite	0.0364602	0.0097987	3.72	0.000	0.017166 0.0557544
Tassazione indiretta	0.2786544	0.0694307	4.01	0.000	0.1419412 0.4153676
Dummy2006	-0.534475	0.1999065	-2.67	0.008	-0.9281027 -0.1408472
Dummy 2007	-0.9076177	0.205679	-4.41	0.000	-1.312612 -0.5026234
Dummy2008	-1.245085	0.2051578	-6.07	0.000	-1.649053 -0.8411165
Dummy2010	-0.6323865	0.2143799	-2.95	0.003	-1.054513 -0.2102598
Dummy2011	-1.341617	0.2301296	-5.83	0.000	-1.794756 -0.8884783
Dummy2012	-1.511694	0.2555158	-5.92	0.000	-2.01482 -1.008568
Dummy2013	-1.64616	0.2789872	-5.90	0.000	-2.195502 -1.096817
Dummy2014	-2.113448	0.3059804	-6.91	0.000	-2.715941 -1.510954
Dummy2015	-2.060994	0.3259784	-6.32	0.000	-2.702865 -1.419123
Cons	29.17707	3.099852	9.41	0.000	23.07328 35.28087

Risultati dell'elaborazione con il software statistico STATA

[64]Maggiori dettagli sono presenti in appendice (Tabella B20). L'indicatore VIF ha suggerito di eliminare le variabili densità e istruzione di secondo livello solo nella simulazione relativa ai Paesi NEW UE. Per maggiori dettagli si veda la tabella B3.

[65]Per ogni modello vengono esposte informazioni di dettaglio in appendice.

[66]Si veda la Tabella B4, B5 e B6 e Figura B5 in appendice per i maggiori dettagli.

MODELLO OLD UE

Tabella 3: Coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE OLDUE3[67]

Variabile	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Disoccupazionegiovanile	0.0718857	0.0245891	2.92	0.004	0.0232527 0 .1205187
Occupazionefemminile	-0.1768448	0.0598368	-2.96	0.004	-0.2951916 -0.0584979
Industrializzazione	-0.0367062	0.0117049	-3.14	0.002	-0.0598564 -0.013556
Istruzione di terzolivello	-0.0703668	0.0376759	-1.87	0.064	-0.1448832 0.0041495
Istruzione di secondo livello	-0.2636953	0.0401316	-6.57	0.000	-0.3430686 -0.1843221
Pil pro capite	0.0401843	0.0106839	3.76	0.000	0.0190533 0.0613153
Tassazioneindiretta	0.2870704	0.1010259	2.84	0.005	0.0872588 0.4868819
Dummy2006	-0.4995376	0.2590279	-1.93	0.056	-1.01185 0.0127744
Dummy 2007	-0.812765	0.2659885	-3.06	0.003	-1.338844 -0.2866861
Dummy2008	-1.10298	0.2614292	-4.22	0.000	-1.620041 -0.5859183
Dummy2010	-0.5338964	0.2643266	-2.02	0.045	-1.056688 -0.0111044
Dummy2011	-0.9485709	0.2806618	-3.38	0.001	-1.503671 -0.3934708
Dummy2012	-0.97377754	0.3126438	-3.11	0.002	-1.59213 -0.3554203
Dummy2013	-0.9297056	0.3402797	-2.73	0.007	-1.60272 -0.2566918
Dummy2014	-1.319977	0.3821524	-3.45	0.001	-2.075808 -0.5641463
Dummy2015	-1.407537	0.3945428	-3.57	0.001	-2.187874 -0.6272
Cons	33.27638	5.333794	6.24	0.000	22.72706 43.82569

Risultati dell'elaborazione con il software statistico STATA

[67] Si veda la Tabella B7, B8 e B9 e Figura B6 in appendice per i maggiori dettagli.

MODELLO NEW UE

Tabella 4: Coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE NEWUE3[68]

Variabile	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Abitazione	0.103939	0.0150503	6.91	0.000	0.0741272 0.1337508
Disoccupazione giovanile	0.0467055	0.0217824	2.14	0.034	0.0035587 0.0898523
Industrializzazione	-0.0665542	0.0096253	-6.91	0.000	-0.0856201 -0.0474882
Istruzione di terzo livello	0.1257048	0.0630756	1.99	0.049	0.0007642 0.2506454
Tassazione indiretta	0.2917889	0.0934864	3.12	0.002	0.1066103 0.4769674
Tassazione diretta	0.2613215	0.0977495	2.67	0.009	0.0676985 0.4549445
Dummy2006	-0.962972	0.2894394	-3.33	0.001	-1.536296 -0.3896482
Dummy 2007	-1.63062	0.3174727	-5.14	0.000	-2.259472 -1.001767
Dummy2008	-1.909833	0.316558	-6.03	0.000	-2.536874 -1.282793
Dummy2010	-0.6370982	0.3318017	-1.92	0.057	-1.294334 0.0201372
Dummy2011	-1.66057	0.3716948	-4.47	0.000	-2.396825 -0.9243138
Dummy2012	-2.261546	0.4124398	-5.48	0.000	-3.07851 -1.444582
Dummy2013	-2.667136	0.458078	-5.82	0.000	-3.574501 -1.759772
Dummy2014	-3.221298	0.4902786	-6.57	0.000	-4.192446 -2.250151
Dummy2015	-3.16162	0.5445508	-5.81	0.000	-4.240271 -2.08297
Cons	17.63092	2.67731	6.59	0.000	12.32769 22.93416

Risultati dell'elaborazione con il software statistico STATA

[68]Si veda la Tabella B10, B11 e B11 e Figura B7 in appendice per i maggiori dettagli.

MODELLO AE

Tabella 5: Coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE AE2[69]

Variabile	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Densità	0.0779926	0.0110749	7.04	0.000	0.0561314 0.0998538
DipendentiPubblici	-0.1493823	0.077449	-1.93	0.055	-0.3022616 0.0034969
Disoccupazionegiovanile	0.075813	0.0188913	4.01	0.000	0.0385229 0.1131031
Occupazionefemminile	-0.2361157	0.0548427	-4.31	0.000	-0.3443716 -0.1278598
Industrializzazione	-0.0515396	0.009064	-5.69	0.000	-0.0694314 -0.0336477
Istruzione di terzolivello	-0.0922295	0.043239	-2.13	0.034	-0.1775803 -0.0068786
Istruzione di secondo livello	-0.2678007	0.0438837	-6.10	0.000	-0.3544243 -0.1811771
PIL pro capite	0.0327438	0.0129882	2.52	0.013	0.0071059 0.0583816
Tassazione indiretta	0.252781	0.0921538	2.74	0.007	0.0708755 0.4346866
Tassazionediretta	0.189606	0.0916442	2.07	0.040	0.0087065 0.3705055
Dummy 2006	-0.5960326	0.2486265	-2.40	0.018	-1.086805 -0.1052603
Dummy 2007	-1.201451	0.2668414	-4.50	0.000	-1.728179 -0.6747241
Dummy2008	-1.293342	0.2685115	-4.82	0.000	-1.823366 -0.7633183
Dummy2010	-0.5296468	0.2701329	-1.96	0.052	-1.062871 0.0035778
Dummy2011	-1.205034	0.2877245	-4.19	0.000	-1.772984 -0.6370854
Dummy2012	-1.507129	0.3247279	-4.64	0.000	-2.14812 -0.8661376
Dummy2013	-1.647671	0.3603249	-4.57	0.000	-2.358929 -0.9364136
Dummy2014	-2.174703	0.3995348	-5.44	0.000	-2.963358 -1.386048
Dummy2015	-2.10728	0.4227716	-4.98	0.000	-2.941803 -1.272757
Cons	28.38126	3.839385	7.39	0.000	20.80256 35.95995

Risultati dell'elaborazione con il software statistico STATA

Le Tabelle 2, 3, 4 e 5 mostrano i risultati principali dei modelli ritenuti migliori sulla base degli indici di bontà di adattamento del modello AIC e BIC e la capacità predittiva tramite l'indicatore RMSE[70]. In tutti i quattro casi esaminati, il test di Hausman e di Wooldridge hanno ritenuto lo stimatore FE consistente[71] e i dati relativi al coefficiente di determinazione e all'esistenza della correlazione tra i residui e le variabili esplicative hanno confermato la scelta del modello fixed effect[72]. I residui si distribuiscono normalmente, sia secondo la rappresentazione grafica che secondo i test utilizzati[73]. I coefficienti dei regressori sono tutti significativi. Sono state utilizzate le procedure robust, jackknife e bootstrap[74]. Infine, l'analisi dei residui, ha fatto emergere la presenza di correlazione[75]. I risultati vengono confermati, sebbene un po' indeboliti in termini di significatività dalle versioni robuste scelte[76].

[69]Si veda la Tabella B14, B15 e B16 e Figura B8 in appendice per i maggiori dettagli.

[70]Maggiori dettagli sono presenti in appendice (Tabella B5, B8, B11, B14).

[71]I dettagli sono illustrati alla nota 48.

[72]Maggiori dettagli sono presenti in appendice (Tabella B21)

[73]Figura B5, B6, B7 e B8 in appendice

[74]Tabella B6, B9, B12 e B15 in appendice.

[75] I risultati sono presenti nell'appendice del Capitolo 5. Tale criticità è stata opportunamente risolta attraverso la trasformazione dei dati di Cochrane-Orcutt (metodo esposto nel paragrafo D2). Il problema congiunto eteroschedasticità/correlazione dei residui è stato superato applicando alla procedura robust l'opzione 'cluster', poiché il raggruppamento produce uno stimatore coerente quando vi è una correlazione seriale. Inoltre, è stato sviluppato un modello panel dinamico per il modello FEUE2 (sviluppato nel Capitolo 5).

[76] Tabella B16, B17, B18, B19.

La Tabella 6 riassume le determinanti dell'Economia Sommersa presenti nei quattro modelli esaminati e il relativo segno del coefficiente di regressione. In particolare, nel modello FEUE2 contribuiscono alla determinazione del fenomeno oggetto di studio nove variabili esplicative (non sono presenti la tassazione diretta, i dipendenti pubblici e la criminalità), mentre i modelli che esaminano i due sottogruppi OLDUE e NEWUE utilizzano, rispettivamente, sette e sei variabili esplicative. Le variabili esplicative in comune tra i modelli FE UE2, FE OLDUE3 e FE NEWUE3 sono l'istruzione di terzo livello, l'industrializzazione, la tassazione indiretta e la disoccupazione giovanile, mentre la tassazione diretta è presente solo nel modello FE NEWUE3. Infine, il modello FE AE2 considera dieci variabili esplicative tra le dodici considerate. L'unica esplicativa non esaminata da alcun modello è la variabile criminalità.

I risultati mostrano un diverso comportamento del modello FE OLDUE rispetto al modello FE NEWUE. Hanno infatti in comune solo tre variabili (industrializzazione, tassazione indiretta e disoccupazione giovanile) con segno atteso uguale ed una con segno atteso contrario (istruzione di terzo livello). Ciò conferma che lo studio separato tra i Paesi UE che hanno aderito all'Unione europea fino al 1995 e quelli entrati dopo il 2004 è giustificato. Inoltre, il modello FE UE è coerente con il modello FE AE.

Tabella 6: Determinanti dell'Economia Sommersa presenti nei quattro modelli esaminati con il rispettivo segno del coefficiente di regressione

Variabile	Modello FE UE2	Modello FE OLDUE3	Modello FE NEWUE3	Modello FE AE2
Densità di popolazione	+			+
Istruzione di secondo livello	-	-		-
Istruzione di terzo livello	-	-	+	-
Industrializzazione	-	-	-	-
Tassazione diretta			+	+
Tassazione indiretta	+	+	+	+
Occupazione femminile	-	-		-
Pil pro capite	+	+		+
Dipendenti pubblici				-
Disoccupazione giovanile	+	+	+	+
Criminalità				
Abitazione	+		+	

Interpretazione economica dei risultati dei quattro modelli FE UE2, FE OLDUE3, FE NEWUE3 e FE AE2

Il **modello FE UE2** evidenzia come tutte le variabili esplicative abbiano coefficienti (Tabella 2) significativi e mostrino un segno atteso coerente con la teoria economica. In particolare, le variabili che definiscono la struttura socio-demografica dell'Unione europea sono le esplicative che esprimono la densità di popolazione, il tasso di partecipazione all'istruzione di terzo livello e di secondo livello. La prima variabile ha un coefficiente con segno positivo, ossia tanto più un Paese è densamente popolato quanto più l'Economia Sommersa aumenta. Tale relazione appare logica, sebbene in contrasto con quanto sostenuto nello studio del caso italiano[77]. Ciò perché la realtà italiana, pur incardinandosi nell'ambito europeo, si comporta in modo diverso. L'istruzione di secondo e terzo livello hanno un segno del coefficiente negativo. Infatti, secondo la letteratura

[77]Tra la densità di popolazione e l'economia sommersa è stata rilevata una relazione inversa (Capitolo 3) in quanto laddove la maggior densità è legata ad una necessità lavorativa, tale variabile può essere correlata negativamente all'economia sommersa (Morvillo, 2016)

economica (Cappariello e Zizza, 2009) l'istruzione ha un ruolo di contrasto rispetto al fenomeno in esame. Concorrono alla definizione della struttura economica la quota delle imposte indirette sul PIL, che sintetizza il contesto economico-istituzionale UE, e il volume di produzione che esprime il livello di industrializzazione. Il coefficiente della variabile relativo alle imposte indirette mostra un segno coerente (+) con la letteratura economica (Amendola e Dell'Anno, 2008), poiché tanto più elevato è il prelievo tributario tanto più elevata sarà l'Economia Sommersa. Il livello di industrializzazione mostra un segno del coefficiente negativo e conforme alla letteratura economica, poiché nei Paesi in cui la dotazione di industrie è particolarmente carente ci si attende una maggior diffusione di Economia Sommersa (Daniele e Marani, 2008).

Tra le variabili di controllo sono presenti nel modello FE UE2 l'occupazione femminile, il PIL pro capite e la disoccupazione giovanile. Il PIL pro capite, in grado di fornire un'indicazione della dimensione della crescita economica, presenta nella letteratura empirica una notevole eterogeneità di opinioni rispetto alle relazioni esistenti tra questa variabile e gli indicatori di Economia Sommersa. Pertanto, vi è una situazione di ambiguità rispetto al segno della relazione (Dell'Arno, 2003; Busato e Chiarini, 2004)[78]. Per le variabili che esprimono la partecipazione femminile al mercato del lavoro e la disoccupazione giovanile, i coefficienti hanno segno atteso conforme con la letteratura economica (Lucifora, 2003)[79]. Infine, in rappresentanza del gruppo della qualità della vita, concorre alla definizione del modello FE UE2 l'esplicativa che esprime il tipo di abitazione. Il suo coefficiente mantiene il segno atteso (+), pertanto secondo il modello FEUE2 quando la qualità della vita non è buona, l'Economia Sommersa aumenta. Infatti, è logico ritenere che vi sia una correlazione positiva tra la percentuale della popolazione che abita in situazioni disagiate e l'Economia Sommersa. Il modello non fa emergere la relazione inversa tra le due variabili.

Il **modello FE OLDUE3** conferma quanto emerso nel modello FE UE2 per le variabili istruzione di secondo e terzo livello, industrializzazione e tassazione indiretta, occupazione femminile, PIL pro capite e disoccupazione giovanile.

Il **modello FE NEWUE3** fornisce risultanze coerenti con il modello FE OLDUE3 per le variabili industrializzazione, tassazione indiretta e disoccupazione giovanile. La tassazione diretta non era una determinante nei modelli precedenti, ma concorre comunque in modo conforme alla teorica economica (+) alla definizione della variabile oggetto di studio. Il livello di istruzione troppo elevato non sempre è un fenomeno di contrasto dell'Economia Sommersa, in quanto in mercati di lavoro "saturi" una elevata istruzione garantisce sicuramente delle migliori opportunità di lavoro ma non sempre maggiori possibilità (Lisi, 2010). L'interpretazione del coefficiente della variabile che esprime il tipo di abitazione è coerente con il modello FEUE2.

Infine, il **modello FE AE3** concorda con le risultanze emerse nel modello FE UE2, ad esclusione delle variabili scelte per rappresentare la qualità della vita, che non sono presenti. L'indicatore di regolamentazione, utile a fornire una fotografia del contesto istituzionale europeo, ha un comportamento coerente con il caso italiano.[80]

[78]Infatti, le attività irregolari sono da considerarsi, per alcuni studi, anticicliche. Più precisamente il settore sommerso appare con caratteristiche anticicliche (Chiarini, 2004) indicando che questo può esercitare un qualche ruolo di "copertura" per famiglie e imprese durante le fasi negative del ciclo. D'altra parte, tali attività contribuiscono al reddito e alla produzione nazionale e quindi ci si potrebbe aspettare un segno positivo della relazione.

[79]"I paesi con maggiori livelli di partecipazione femminile al mercato del lavoro presentano dimensioni minori di economia sommersa. Ci si attende un effetto negativo sulle dimensioni del sommerso. Tuttavia, controllando per la disoccupazione una maggior partecipazione può aumentare alcuni tipi di lavoro irregolare."

[80]L'indice di regolamentazione fa generalmente riferimento ad indicatori di regolamentazione costruiti tenendo conto dello stock di tutte le leggi in vigore, dello Stato e degli enti locali, relative all'accesso al lavoro, alla sicurezza sociale, alle ore lavorative, alle condizioni di lavoro, all'esercizio dell'attività d'impresa. In questo contesto si è deciso di utilizzare quale indicatore il rapporto tra i dipendenti ascrivibili al pubblico impiego e le forze

4.4 Conclusioni

L'approfondimento del presente Capitolo estende lo studio effettuato in ambito italiano (Capitolo 3) al caso europeo. La variabile in esame viene in questo contesto identificata con l'Economia Sommersa, così come la definisce Friedrich Schneider[81] (Medina e Schneider, 2017). La metodologia utilizzata per ottenere tali informazioni è l'approccio MIMIC (Multiple Indicators, Multiple Causes). In questo modello viene inizialmente stabilito un approccio teorico che spiega la relazione tra le variabili esogene e la variabile latente (nel nostro caso l'Economia Sommersa), per poi studiare l'effetto della variabile latente (Economia Sommersa) sulle variabili dell'indicatore macroeconomico. I risultati emersi dall'analisi descrittiva e la letteratura economica esaminata hanno suggerito di affiancare l'analisi per tutti i 28 Paesi membri dell'UE, con l'analisi per i 15 Paesi membri appartenenti all'Unione europea dalla sua costituzione al 1995 (OLD UE) e per i 13 Paesi che hanno aderito dal 2004 (NEWUE). Infatti, i due gruppi possiedono specifiche caratteristiche. Innanzitutto, hanno differenze significative in termini di sviluppo economico, a tal punto che i Paesi appartenenti al gruppo OLD UE vengono ritenuti dagli economisti "sviluppati", mentre i restanti vengono definiti emergenti. Questo perché una serie di fattori, quali ad esempio la privatizzazione, il contesto legislativo e l'orientamento nelle esportazioni, si comportano diversamente nei due gruppi. Gli economisti inoltre sostengono che il processo di europeizzazione viene applicato in modo differente nei due gruppi, in quanto il livello di adeguamento alle regole europee è più consolidato nei Paesi OLD UE, mentre è ancora incerto nei Paesi NEW UE. Viene, infine, effettuata una analisi che considera i 19 Paesi UE che utilizzano la moneta unica europea, l'euro.

Il database è costituito da un panel bilanciato relativo ai 28 Paesi UE, composto dalla variabile dipendente, ES, che rappresenta l'Economia Sommersa, disponibile per l'arco temporale 1991-2015. Le variabili esplicative a disposizione sono 12, volte a spiegare la struttura socio-demografica, economica e la qualità della vita, tutte fruibili per archi temporali diversi. Pertanto, come periodo di riferimento è stato considerato l'intervallo 2005-2015, utilizzabile per tutte le variabili considerate.

I modelli esaminati in questo studio hanno confermato quanto già sostenuto nel caso italiano per il PIL, la tassazione, l'industrializzazione, l'occupazione femminile, la disoccupazione giovanile, l'intensità di regolamentazione, il livello di istruzione medio. Solo per il caso europeo è stato considerato il concetto di qualità della vita ed è risultato espresso in modo significativo e coerente nel modello con tutti e 28 i Paesi europei attraverso l'esplicativa che esprime il tipo di abitazione. Il suo coefficiente mantiene il segno atteso (+), infatti, è logico ritenere che vi sia una correlazione positiva tra la percentuale della popolazione che abita in situazioni disagiate e l'Economia Sommersa, pertanto in situazioni in cui la qualità della vita non è buona, l'Economia Sommersa aumenta. La relazione tra il livello di istruzione elevato e l'Economia Sommersa, per i modelli FE UE, FE OLDUE, FE AE, ha un segno del coefficiente negativo, in quanto secondo la letteratura economica l'istruzione ha un ruolo di contrasto rispetto al fenomeno in esame. Viceversa, per il modello FE NEWUE la variabile esplicativa ha un segno del coefficiente positivo. Infatti, un livello di istruzione troppo elevato non sempre è un fenomeno di contrasto dell'Economia Sommersa, poiché in mercati di lavoro "saturi" una elevata istruzione garantisce sicuramente delle migliori opportunità occupazionali ma non sempre maggiori possibilità. Infine, la relazione tra l'Economia Sommersa e la densità di popolazione mostra un segno positivo, ossia tanto più un Paese è densamente popolato quanto più l'Economia Sommersa

di lavoro in età 15-64 anni. Tale indicatore, così come costruito, è stato utilizzato da Frey e Weck-Hanneman nel 1984. Nel suddetto lavoro la relazione trovata dagli autori era risultata positiva. Va però evidenziato che l'analisi veniva applicata su un campione di 17 paesi OECD, con riferimento all'arco temporale 1960-1978 e con un modello diverso da quello utilizzato nel presente studio. Sulla base delle considerazioni appena fatte il risultato ottenuto è quindi coerente ed è pertanto agevole ritenere che nelle zone con una maggiore presenza di dipendenti pubblici il sommerso sia meno radicato e ciò a dimostrazione della positiva opera dei pubblici dipendenti di tutte le istituzioni centrali e periferiche (Morvillo,2016).

[81]Professore di economia presso l'università Johannes Kepler di Linz in Austria e dal 2006 ricercatore presso l'istituto tedesco di ricerche economiche

aumenta. Tale relazione appare logica, sebbene in contrasto con quanto sostenuto nello studio del caso italiano. Ciò perché la realtà italiana, pur incardinandosi nell'ambito europeo, si comporta in modo diverso.

APPENDICE CAPITOLO 4

Tabella B1: Valori dell'Economia Sommersa in percentuale di PIL per i Paesi membri dell'Unione europea, 1991-2003

	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
Austria	9,0	9,3	10,0	9,7	9,7	9,9	9,6	9,5	9,2	8,8	8,5	8,5	8,7
Belgio	22,1	22,1	23,3	23,5	23,2	23,4	22,2	22,9	21,6	19,9	19,8	20,9	21,7
Bulgaria	35,1	35,0	36,1	34,6	32,9	31,5	30,7	32,8	34,6	35,3	34,9	33,5	33,0
Cipro	36,2	34,7	35,3	34,1	27,9	28,9	29,0	30,5	30,1	28,7	28,4	29,3	31,6
Croazia	30,7	32,9	34,9	35,9	37,3	35,6	32,6	34,2	34,6	32,0	30,9	29,1	27,1
Danimarca	17,1	17,0	18,1	16,7	16,2	16,5	15,2	15,5	15,2	14,6	14,2	14,8	14,9
Estonia	23,5	26,0	29,1	29,8	30,5	30,2	27,0	26,8	27,6	27,7	26,2	25,4	24,8
Finlandia	16,5	17,1	17,0	16,3	15,7	15,9	14,5	13,8	13,4	12,5	12,5	13,0	12,7
Francia	15,0	15,6	16,8	16,6	16,2	16,3	16,0	15,3	14,9	13,8	13,3	14,7	14,6
Germania	13,3	13,8	14,3	14,2	14,1	14,6	14,0	13,7	13,3	12,9	12,5	13,0	13,2
Grecia	28,8	28,5	29,4	28,9	29,8	28,6	28,9	28,2	27,8	26,1	26,5	27,0	26,2
Irlanda	18,4	18,3	18,1	17,7	16,8	16,7	15,5	14,8	13,8	13,4	12,9	13,2	13,8
Italia	29,1	28,5	28,3	27,2	24,8	24,2	25,1	24,1	24,5	22,7	23,6	23,5	24,3
Lettonia	20,1	24,4	25,3	24,8	28,7	28,1	27,0	27,4	27,1	26,7	25,2	25,1	23,7
Lituania	21,2	23,8	26,4	28,8	32,5	32,2	30,9	31,3	30,9	31,1	29,3	28,5	27,0
Lussemburgo	11,1	11,4	11,4	11,2	11,4	12,0	11,4	10,9	10,4	9,8	10,2	10,3	10,7
Malta	31,5	30,6	31,4	31,0	30,9	33,1	31,7	30,6	29,7	27,1	30,7	30,2	31,0
Paesi Bassi	13,2	13,1	13,4	13,3	13,0	12,8	11,8	11,5	10,9	10,5	10,4	11,3	11,8
Polonia	33,1	32,7	32,0	30,2	29,5	28,4	27,6	26,1	26,7	26,2	26,9	26,7	26,4
Portogallo	23,3	23,7	24,4	24,2	23,6	23,0	22,8	21,9	22,0	21,4	21,8	21,7	22,4
Regno Unito	13,7	13,9	13,4	12,8	12,1	12,0	11,3	11,0	11,1	10,8	10,7	11,2	11,2
Repubblica Ceca	18,4	17,8	18,2	18,2	16,8	16,1	16,7	16,3	17,2	16,8	15,8	16,8	17,1
Romania	36,0	35,1	34,8	35,0	33,4	31,1	31,7	32,2	34,5	34,4	32,3	32,5	33,0
Slovacchia	17,2	19,5	19,3	18,3	17,9	18,5	17,2	17,9	17,4	17,6	17,2	17,2	16,6
Slovenia	27,4	28,6	29,5	28,2	28,2	27,0	27,5	25,0	25,9	25,2	25,0	24,5	24,4
Spagna	27,5	28,0	28,7	28,0	27,4	26,1	26,0	24,8	24,5	22,7	23,0	23,1	23,1
Svezia	15,5	17,0	17,9	16,7	15,4	16,4	15,1	14,9	13,7	12,6	12,1	12,9	12,9
Ungheria	31,9	32,3	33,7	32,0	30,2	29,2	28,4	27,1	26,6	25,1	24,7	24,1	24,2

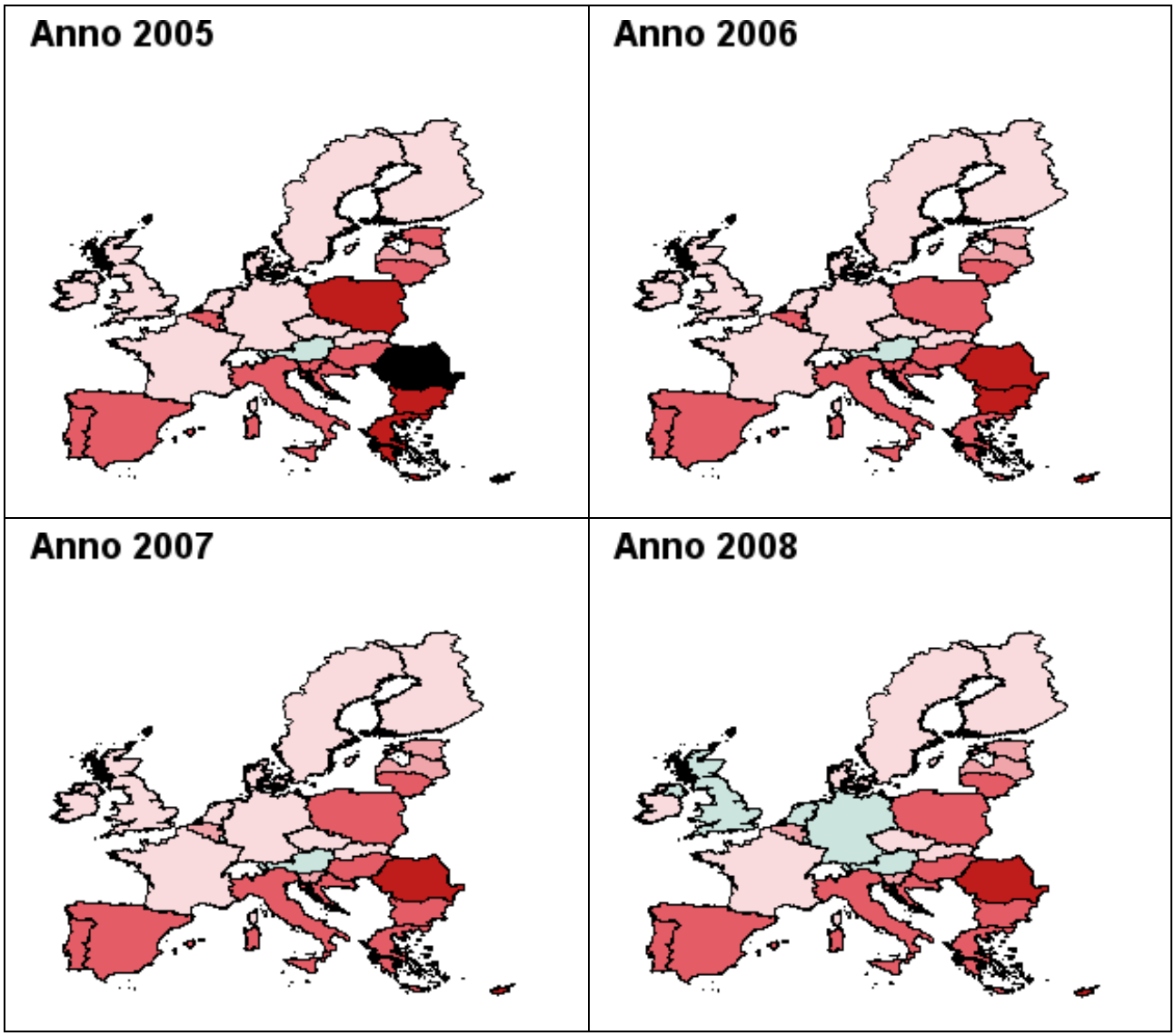
Fonte: Elaborazione dati Medina e Schneider 2017

Tabella B1-bis: Valori dell'Economia Sommersa in percentuale di PIL per i Paesi membri dell'Unione europea, 2004-2015

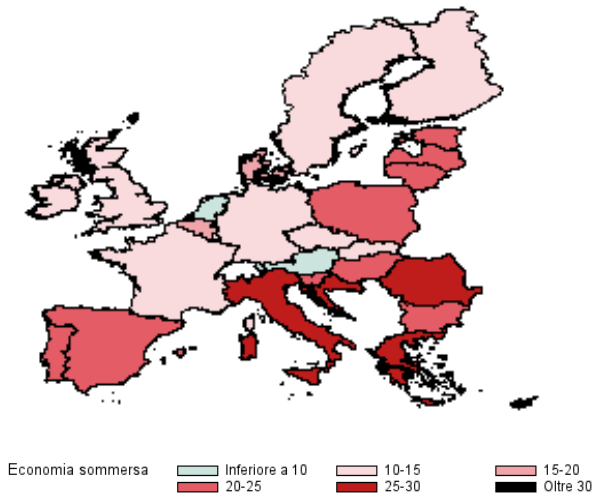
	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015
Austria	8,7	8,9	8,3	7,7	7,8	9,7	9,1	8,5	8,4	8,7	8,4	9,0
Belgio	21,1	21,1	20,7	18,3	18,3	18,7	18,8	17,7	18,3	18,8	18,6	17,8
Bulgaria	30,6	28,6	26,8	23,7	22,8	24,1	23,4	22,4	22,1	22,4	21,6	20,8
Cipro	30,7	30,8	29,9	29,0	28,8	31,6	31,4	32,7	33,3	34,7	32,7	32,2
Croazia	26,1	25,0	23,8	22,5	21,6	25,3	25,6	24,6	25,3	25,3	24,5	23,0
Danimarca	14,6	13,8	12,7	12,5	13,0	16,3	16,2	15,3	15,5	15,2	14,1	14,7
Estonia	23,2	21,3	19,0	17,8	19,4	24,6	23,0	19,7	18,3	18,0	17,5	18,5
Finlandia	12,3	12,0	11,3	11,0	11,0	13,1	12,5	12,2	12,6	12,1	12,1	13,3
Francia	14,0	14,0	13,3	12,9	11,6	13,9	13,1	13,8	12,1	12,4	12,1	11,7
Germania	12,8	12,6	11,4	10,6	9,6	11,7	10,9	9,1	8,9	9,2	8,2	7,8
Grecia	25,3	26,0	24,9	24,2	23,2	25,3	26,2	27,1	28,4	27,8	27,1	26,5
Irlanda	13,5	13,1	12,6	12,6	12,5	13,4	11,8	12,5	11,4	11,1	9,9	9,6
Italia	24,2	24,6	23,8	22,4	23,5	27,3	26,1	24,5	25,5	24,5	24,3	23,0
Lettonia	22,1	19,9	18,1	17,0	18,3	21,2	20,4	18,7	17,3	16,7	15,9	16,6
Lituania	25,7	23,9	22,4	20,6	20,3	24,3	23,1	20,9	19,3	18,3	17,6	18,7
Lussemburgo	10,7	10,7	10,3	9,4	9,7	11,0	10,4	10,3	10,8	10,7	10,4	10,4
Malta	31,9	30,8	28,7	27,0	27,3	30,6	29,2	28,1	27,3	27,2	28,1	29,4
Paesi Bassi	11,4	11,1	10,9	10,6	9,6	8,9	8,6	8,1	8,1	8,8	8,8	7,8
Polonia	25,8	25,3	24,2	23,5	21,7	21,6	20,9	19,3	19,0	18,9	18,1	16,7
Portogallo	22,3	22,7	22,7	23,1	20,7	21,7	20,8	20,4	20,2	20,4	19,3	17,8
Regno Unito	11,4	11,4	10,4	10,8	9,8	11,0	10,3	10,1	9,9	9,6	8,8	8,3
Repubblica Ceca	15,8	14,5	13,1	11,5	11,2	13,5	13,0	11,7	11,5	11,8	10,8	10,5
Romania	30,6	30,5	28,9	27,0	25,4	28,2	26,8	25,4	25,1	24,0	22,7	22,9
Slovacchia	15,4	14,5	13,5	12,2	11,5	13,5	12,8	12,0	11,8	11,8	11,6	11,2
Slovenia	23,3	22,7	20,9	18,0	17,6	22,2	22,5	22,2	22,9	23,0	21,5	20,2
Spagna	23,5	23,3	23,0	22,7	21,5	24,2	23,9	23,7	24,1	24,4	24,0	22,0
Svezia	12,1	12,3	11,1	10,1	10,3	12,7	11,5	11,1	11,9	12,3	11,9	11,7
Ungheria	22,9	22,5	21,1	21,4	20,6	23,2	22,8	21,9	22,3	21,6	20,8	20,5

Fonte: Elaborazione dati Medina e Schneider 2017

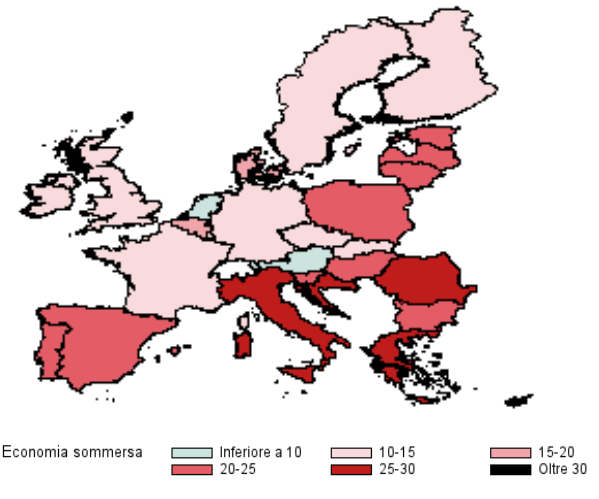
Figura B1: Rappresentazione cartografica dell'Economia Sommersa (% PIL) per Paese membro dell'Unione europea, dal 2005 al 2014 Fonte: Elaborazione dati Medina e Schneider 2017 - © EuroGeographics per i confini amministrativi



Anno 2009

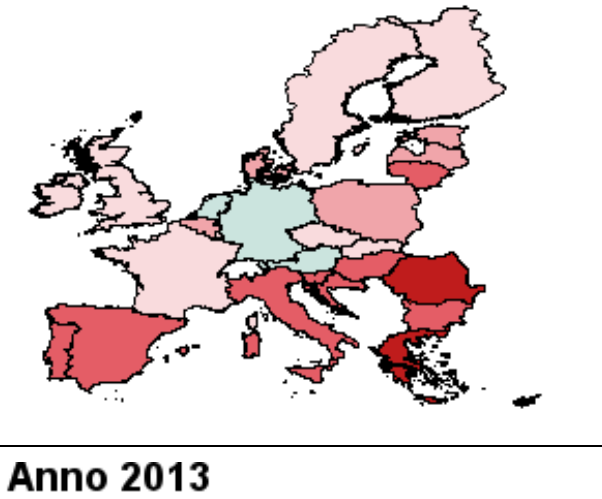


Anno 2010

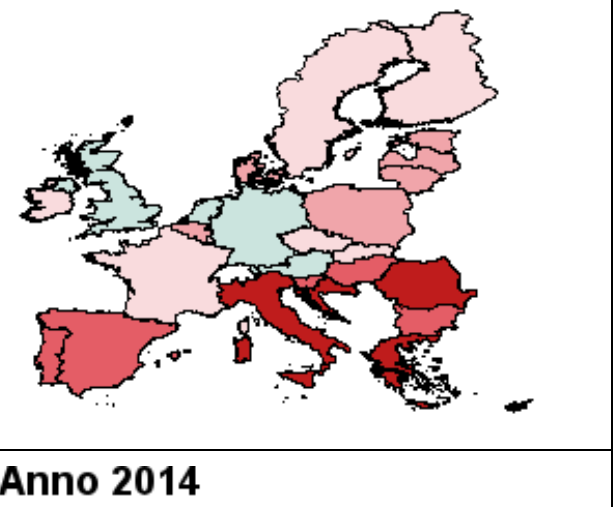


Continua

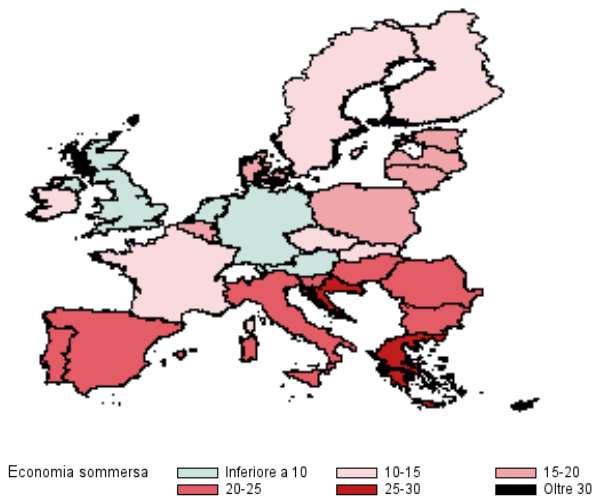
Anno 2011



Anno 2012



Anno 2013



Anno 2014

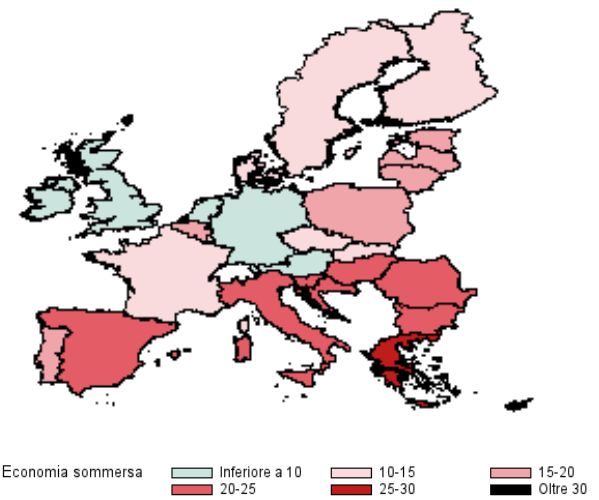


Tabella B2: Informazione sui Paesi dell'Unione europea

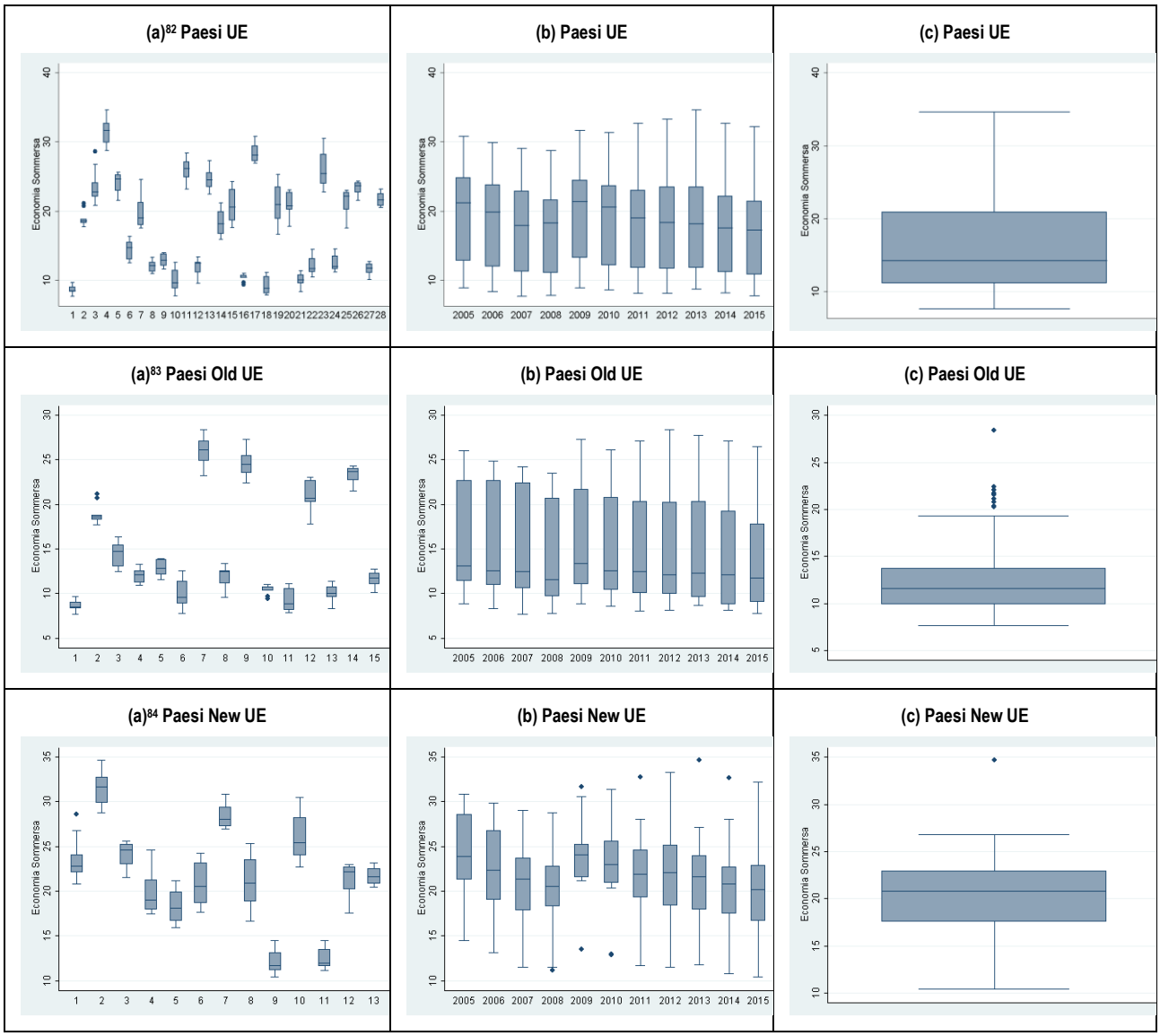
	Anno di adesione UE	Valuta	OLD/NEW UE	Area Euro
Austria	01/01/1995	Euro	OLD UE	AE
Belgio	25/03/1957	Euro	OLD UE	AE
Bulgaria	01/01/2007	Lev Bulgaro	NEWUE	
Cipro	01/05/2004	Euro	NEW UE	AE
Croazia	01/07/2013	Kuna croata	NEW UE	
Danimarca	01/01/1973	Corona danese	OLD UE	
Estonia	01/05/2004	Euro	NEW UE	AE
Finlandia	01/01/1995	Euro	OLD UE	AE
Francia	25/03/1957	Euro	OLDUE	AE
Germania	25/03/1957	Euro	OLD UE	AE
Grecia	01/01/1981	Euro	OLD UE	AE
Irlanda	01/01/1973	Euro	OLD UE	AE
Italia	25/03/1957	Euro	OLD UE	AE
Lettonia	01/05/2004	Euro	NEW UE	AE
Lituania	01/05/2004	Euro	NEW UE	AE
Lussemburgo	25/03/1957	Euro	OLD UE	AE
Malta	01/05/2004	Euro	NEW UE	AE
Paesi Bassi	25/03/1957	Euro	OLD UE	AE
Polonia	01/05/2004	Kloty	NEW UE	
Portogallo	01/01/1986	Euro	OLD UE	AE
Regno Unito	01/01/1973	Sterlina britannica	OLD UE	
Repubblica Ceca	01/05/2004	Corona ceca	NEW UE	
Romania	01/01/2007	Leu rumeno	NEW UE	
Slovacchia	01/05/2004	Euro	NEW UE	AE
Slovenia	01/05/2004	Euro	NEW UE	AE
Spagna	01/01/1986	Euro	OLD UE	AE
Svezia	01/01/1995	Corona svedese	OLD UE	
Ungheria	01/05/2004	Fiorino ungherese	NEW UE	

Tabella B3: Risultati dell'indicatore VIF per i quattro modelli analizzati

Modello UE		Modello OLD UE		Modello NEW UE		Modello AE	
Variabile	**VIF**	**Variabile**	**VIF**	**Variabile**	**VIF**	**Variabile**	**VIF**
Densità	4.43	Abitazione	3.38	Occ.fem.	5.48	Dippub	5.48
Dippub	3.72	Istr	3.00	pilpc	3.91	densità	5.47
Istr1	3.07	Disoc.giov.	3.00	taxdiretta	3.68	Occ.fem	2.89
Pilpc	2.75	Istr 1	2.66	dippub	3.42	taxdiretta	2.84
Occ.fem.	2.74	Occ.fem.	2.60	istr	3.32	Istr1	2.63
taxdiretta	2.67	taxindiretta	2.55	indu	2.45	pilpc	2.41
istr	2.43	taxdiretta	2.29	Disoc.giov.	1.79	Disoc.giov.	2.22
Disoc.giov.	2.00	Densità	2.28	taxindiretta	1.76	istr	1.91
Indu	1.55	Pilpc	2.26	Abitazione	1.48	indu	1.75
taxindiretta	1.44	Dippub	1.71	Criminalità	1.39	taxindiretta	1.67
abitazione	1.36	Indu	1.70			abitazione	1.59
criminalità	1.18	Criminalità	1.47			criminalità	1.23

Risultati dell'elaborazione con il software statistico STATA

Figura B2: Rappresentazioni grafiche (box plot) dell'Economia Sommersa per codice Paese (a), per anno (b) e sopra l'ottantesimo percentile (c)



Risultati dell'elaborazione con il software statistico STATA

[82] Codice Paese: 1=Austria; 2=Belgio; 3=Bulgaria; 4=Cipro; 5=Croazia; 6=Danimarca; 7=Estonia; 8=Finlandia; 9=Francia; 10=Germania; 11=Grecia; 12=Irlanda; 13=Italia; 14=Lettonia; 15=Lituania; 16=Lussemburgo; 17=Malta; 18=Paesi Bassi; 19=Polonia; 20=Portogallo;21=Regno Unito; 22=Repubblica Ceca; 23=Romania; 24=Slovacchia; 25=Slovenia; 26=Spagna; 27=Svezia; 28=Ungheria.

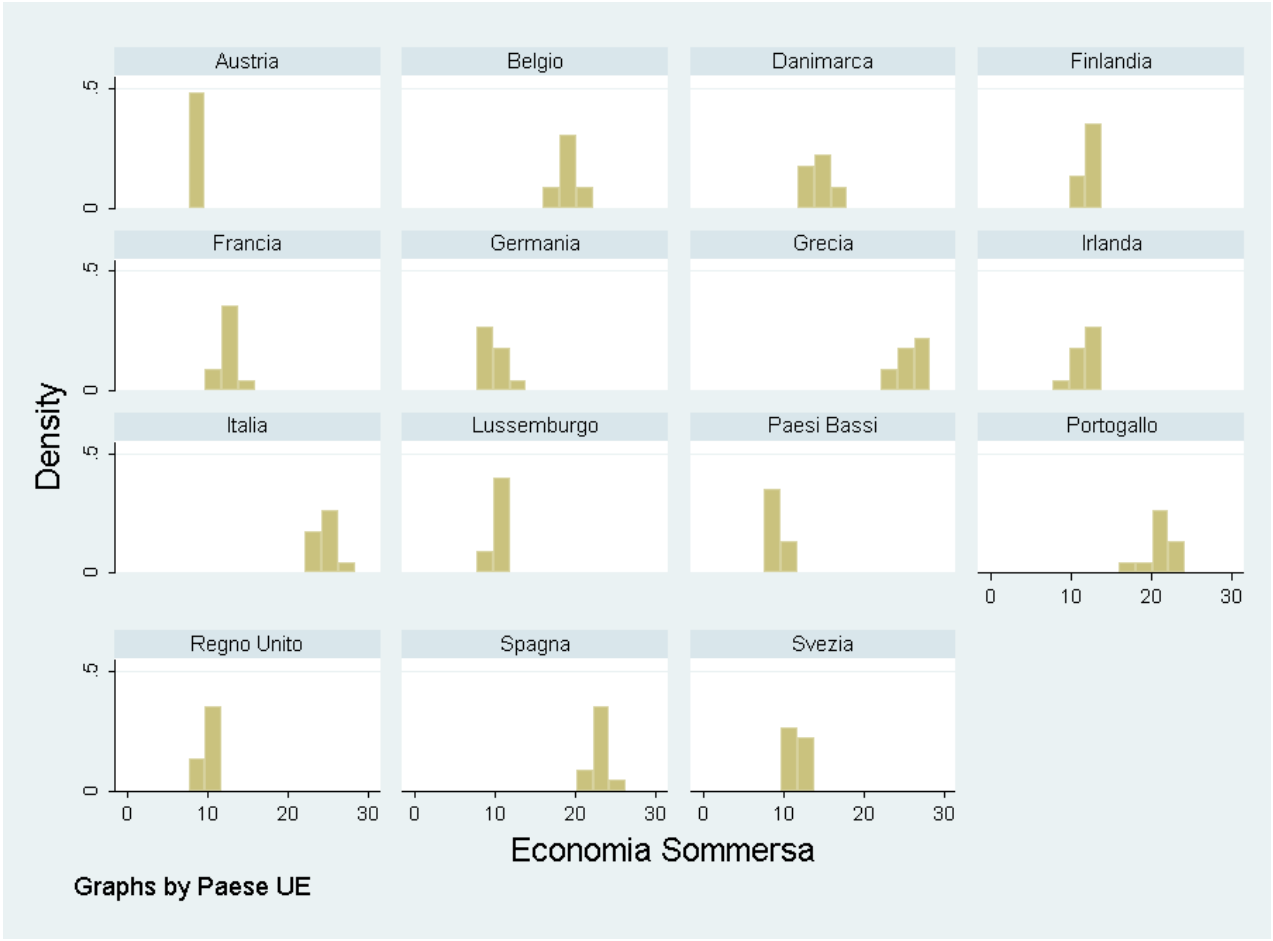
[83] Codice Paese: 1=Austria; 2=Belgio; 3= Danimarca; 4=Finlandia; 5=Francia; 6=Germania; 7=Grecia; 8=Irlanda; 9=Italia; 10=Lussemburgo; 11=Paesi Bassi; 12=Portogallo;13=Regno Unito; 14=Spagna; 15=Svezia.

[84] Codice Paese: 1=Bulgaria; 2=Cipro; 3=Croazia; 4=Estonia; 5=Lettonia; 6=Lituania; 7=Malta; 8=Polonia; 9=Repubblica Ceca; 10=Romania; 11=Slovacchia; 12=Slovenia; 13=Ungheria.

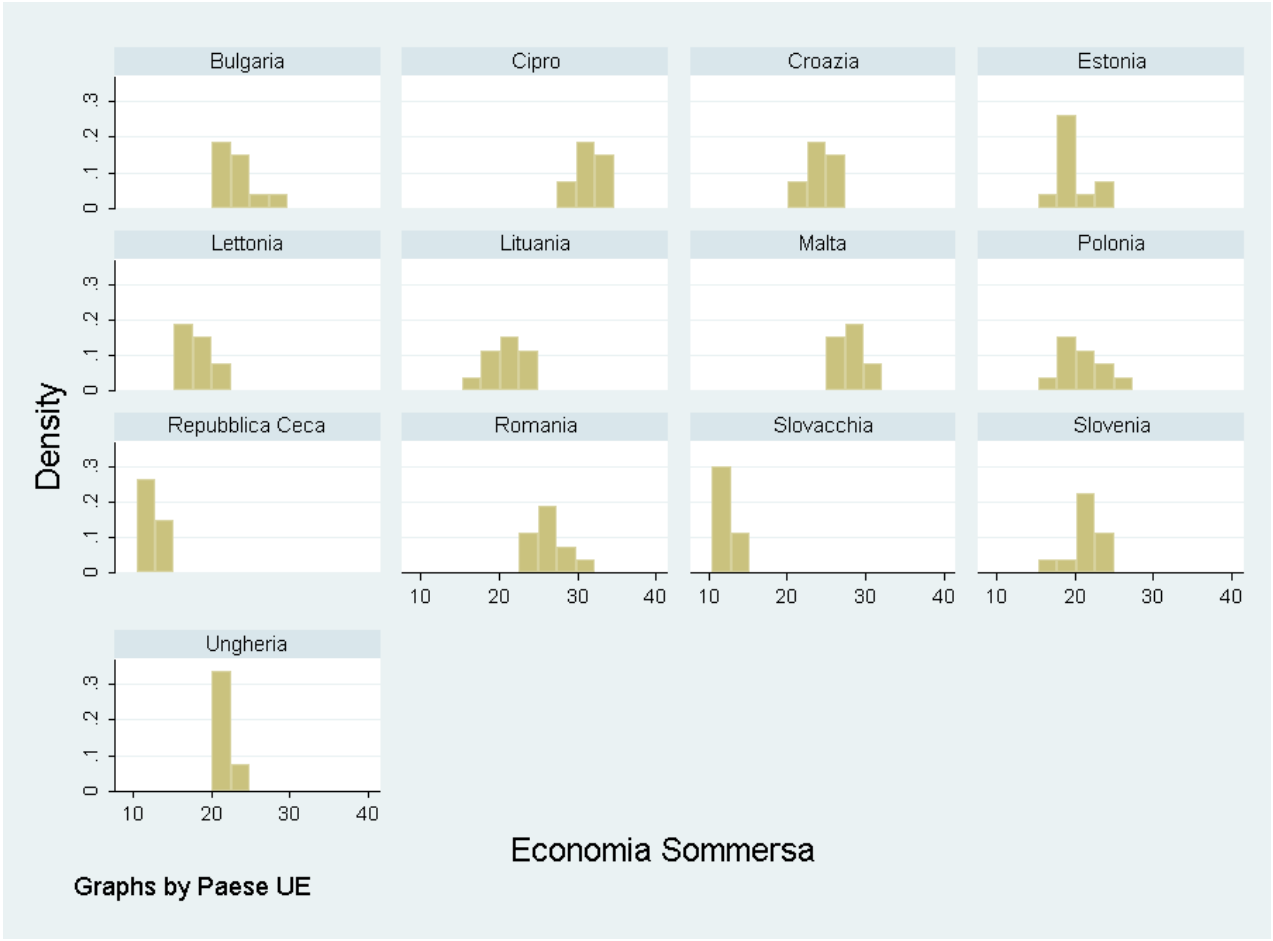
Figura B3: Istogrammi della variabile dipendente Economia Sommersa per Paese

Paesi UE

Paesi Old UE


Austria
Belgio
Danimarca
Finlandia
Francia
Germania
Grecia
Irlanda
Italia
Lussemburgo
Paesi Bassi
Portogallo
Regno Unito
Spagna
Svezia
Density
0
.5
0
10
20
30
Economia Sommersa
Graphs by Paese UE

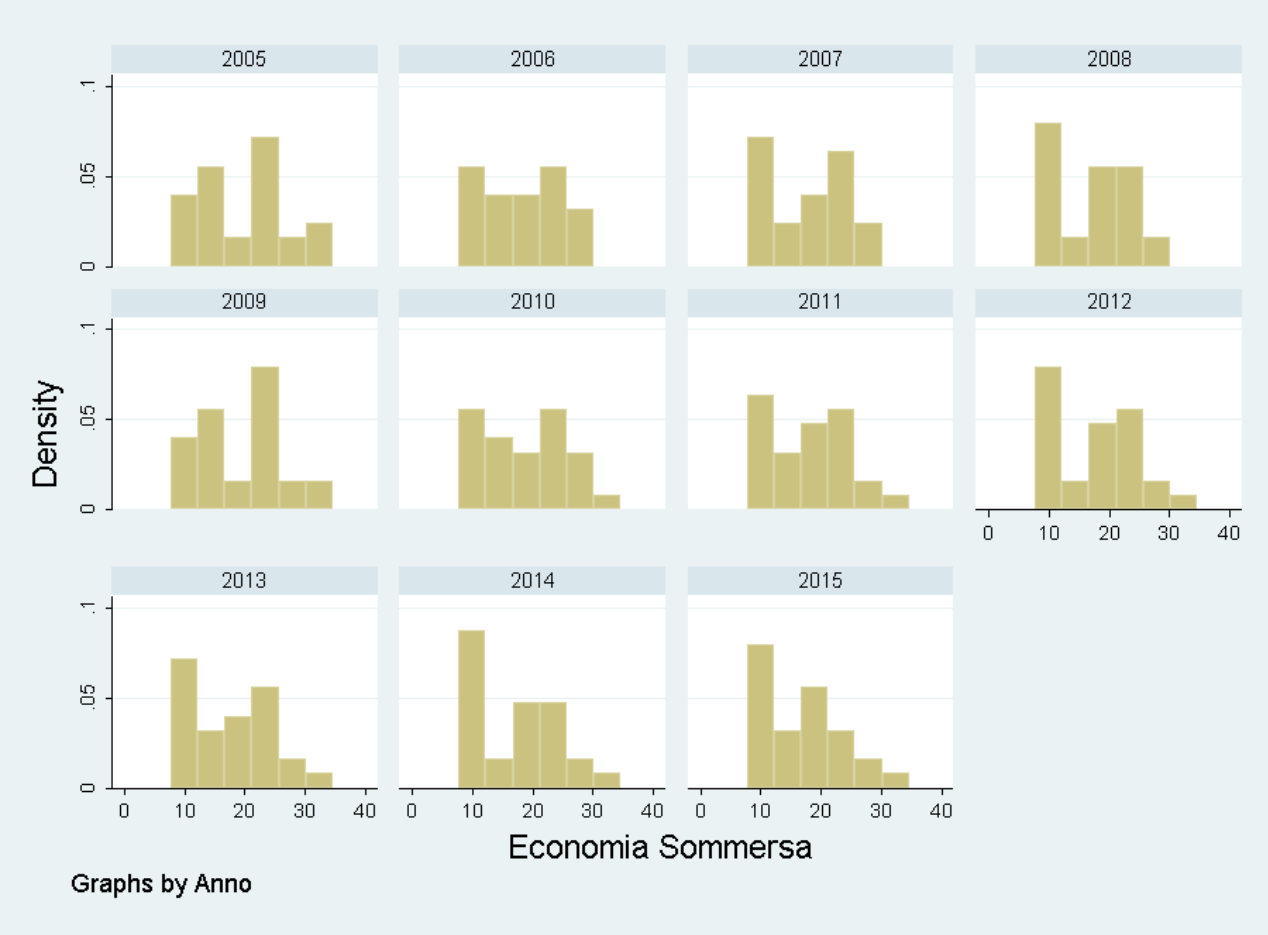
Paesi New UE



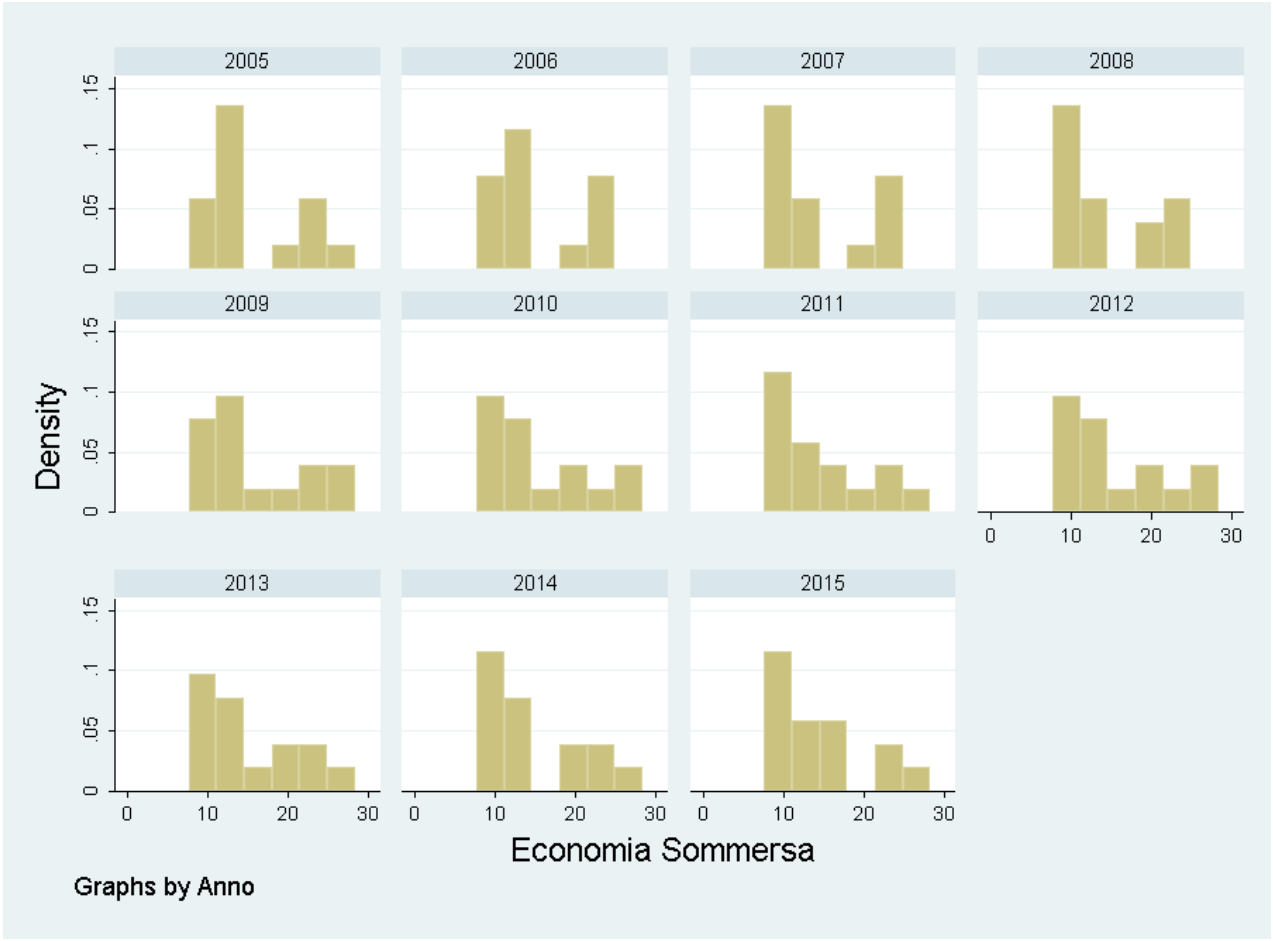
Risultati dell'elaborazione con il software statistico STATA

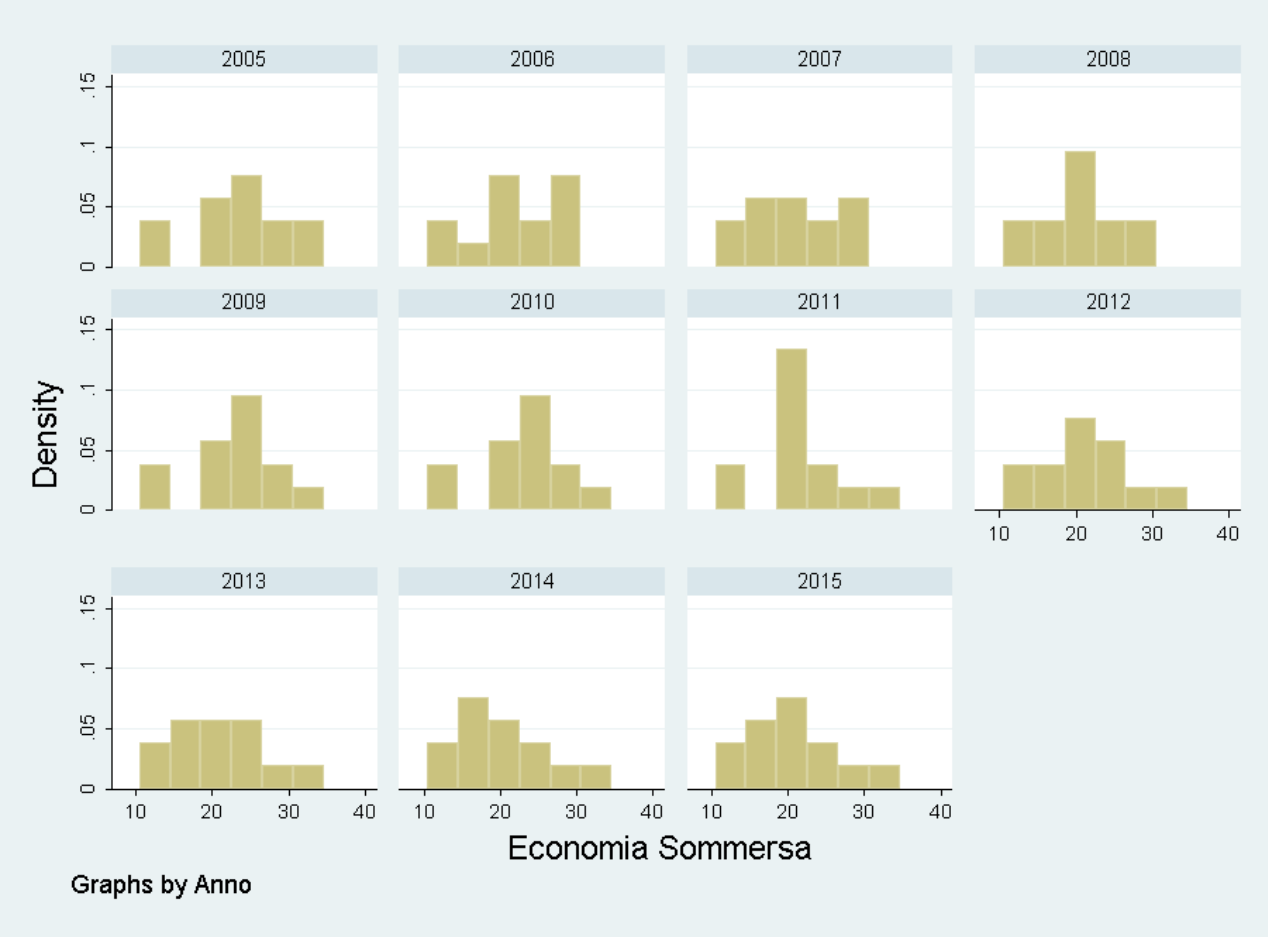
Figura B4: Istogrammi della variabile Economia Sommersa per anno

Paesi UE



Paesi Old UE


2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Density
0
.05
.1
.15
0
10
20
30
Economia Sommersa
Graphs by Anno

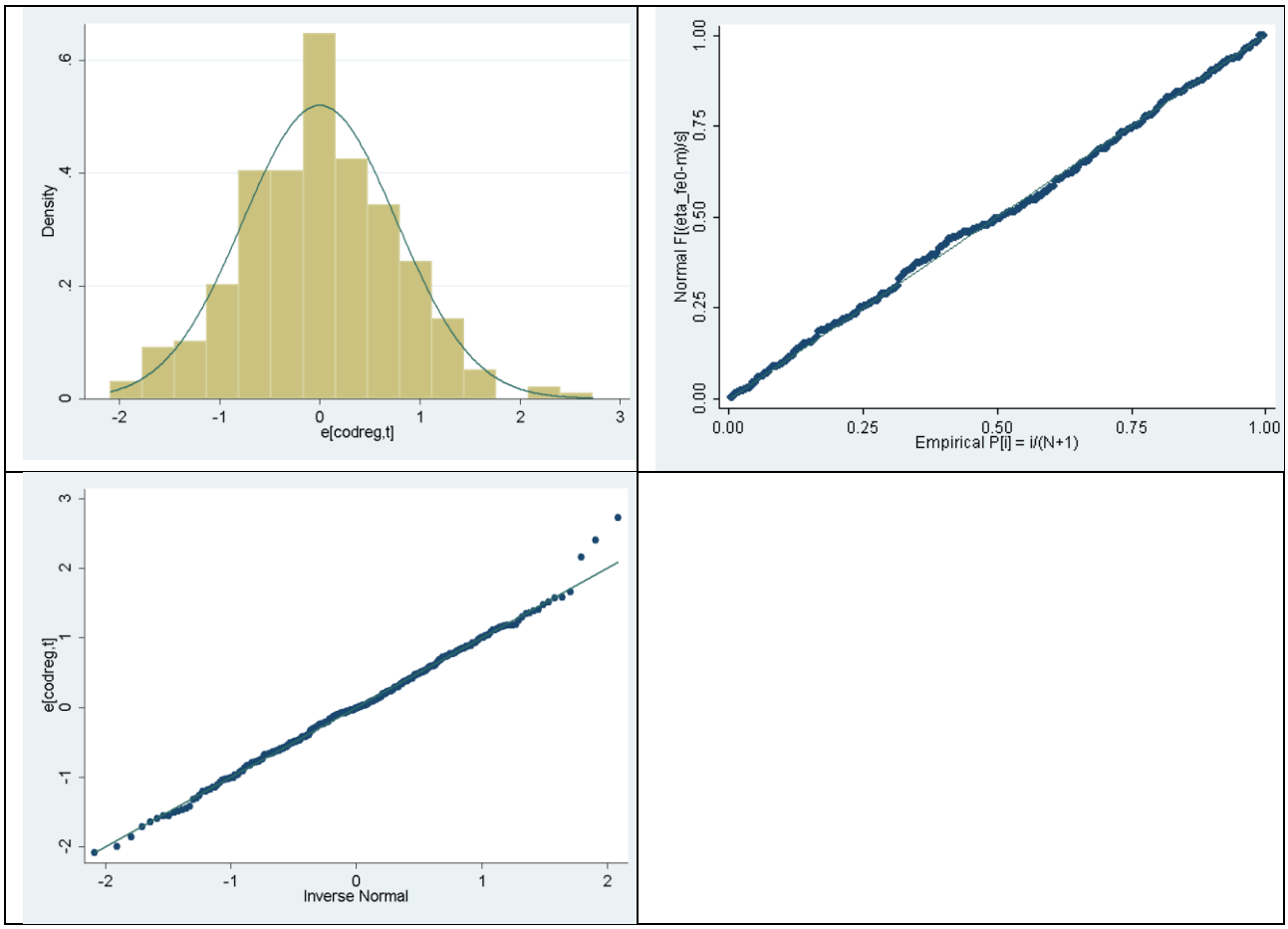


Risultati dell'elaborazione con il software statistico STATA

Tabella B4: Valutazione dei modelli Fixed Effect (FE) per i Paesi UE, tramite gli indicatori AIC, BIC e RMSE

Nome Modello	Modello	AIC	BIC	RMSE
FEUE1	$Y_{it} = \beta_{12}x_{12it} + \beta_1 x_{1it} + \beta_{10} x_{10it} + \beta_7 x_{7it} + \beta_4 x_{4it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_8 x_{8it} + \beta_6 x_{6it} + \alpha_i + \varepsilon_{it}$	808.6	845.9	0.93
FEUE2	$Y_{it} = \beta_{12}x_{12it} + \beta_1 x_{1it} + \beta_{10} x_{10it} + \beta_7 x_{7it} + \beta_4 x_{4it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_8 x_{8it} + \beta_6 x_{6it} + D_t + \alpha_i + \varepsilon_{it}$	747.6	818.5	0.83
FEUE3	$Y_{it} = \beta_{12}x_{12it} + \beta_9 x_{9it} + \beta_{10} x_{10it} + \beta_2 x_{2it} + \beta_6 x_{6it} + D_t + \alpha_i + \varepsilon_{it}$	856.9	912.8	0.99

Figura B5: Risultati della verifica della normalità dei residui – FE UE2



Risultati dell'elaborazione con il software statistico STATA

Tabella B5: Risultati dei test di verifica della normalità dei residui – FE UE2

Test di normalità					
Test	Osservazioni	W	V	z	Prob>z
Shapiro-Wilk	308	0.99598	0.877	0.309	**0.62122**
Test	Osservazioni	W'	V'	z	Prob>z
Shapiro-Francia	308	0.99540	1.086	0.176	**0.43021**
Test	Osservazioni	Pr(Asimmetria)	Pr(Curtosi)	χ^2	Prob>χ2
Asimmetria/Curtosi	308	0.4519	0.2200	2.08	**0.3530**

Risultati dell'elaborazione con il software statistico STATA

Tabella B6: Modello FE UE2 con procedure robust[85], jackknife e bootstrap

Variabile	Coefficiente di regressione	p-value	p-value robust	p-value jackknife	p-value bootstrap
Abitazione	0.717061	0.000	0.000	0.001	0.000
Densità di popolazione	0.659411	0.000	0.000	0.004	0.009
Disoccupazione giovanile	0.587464	0.000	0.015	0.034	0.016
Occupazionefemminile	-0.1595814	0.000	0.009	0.019	0.011
Industrializzazione	-0.0644933	0.000	0.000	0.000	0.000
Istruzione di terzolivello	-0.0953173	0.009	0.077	0.197	0.170
Istruzione di secondo livello	-0.2439303	0.000	0.005	0.038	0.005
Pil pro capite	0.0364602	0.000	0.000	0.001	0.001
Tassazioneindiretta	0.2786544	0.000	0.016	0.033	0.025
Dummy2006	-0.534475	0.008	0.001	0.001	0.000
Dummy 2007	-0.9076177	0.000	0.000	0.000	0.000
Dummy2008	-1.245085	0.000	0.000	0.000	0.000
Dummy2010	-0.6323865	0.003	0.000	0.001	0.000
Dummy2011	-1.341617	0.000	0.000	0.000	0.000
Dummy2012	-1.511694	0.000	0.000	0.000	0.000
Dummy2013	-1.64616	0.000	0.000	0.000	0.000
Dummy2014	-2.113448	0.000	0.000	0.000	0.000
Dummy2015	-2.060994	0.000	0.000	0.000	0.000
Cons	26.17707	0.000	0.000	0.000	0.000

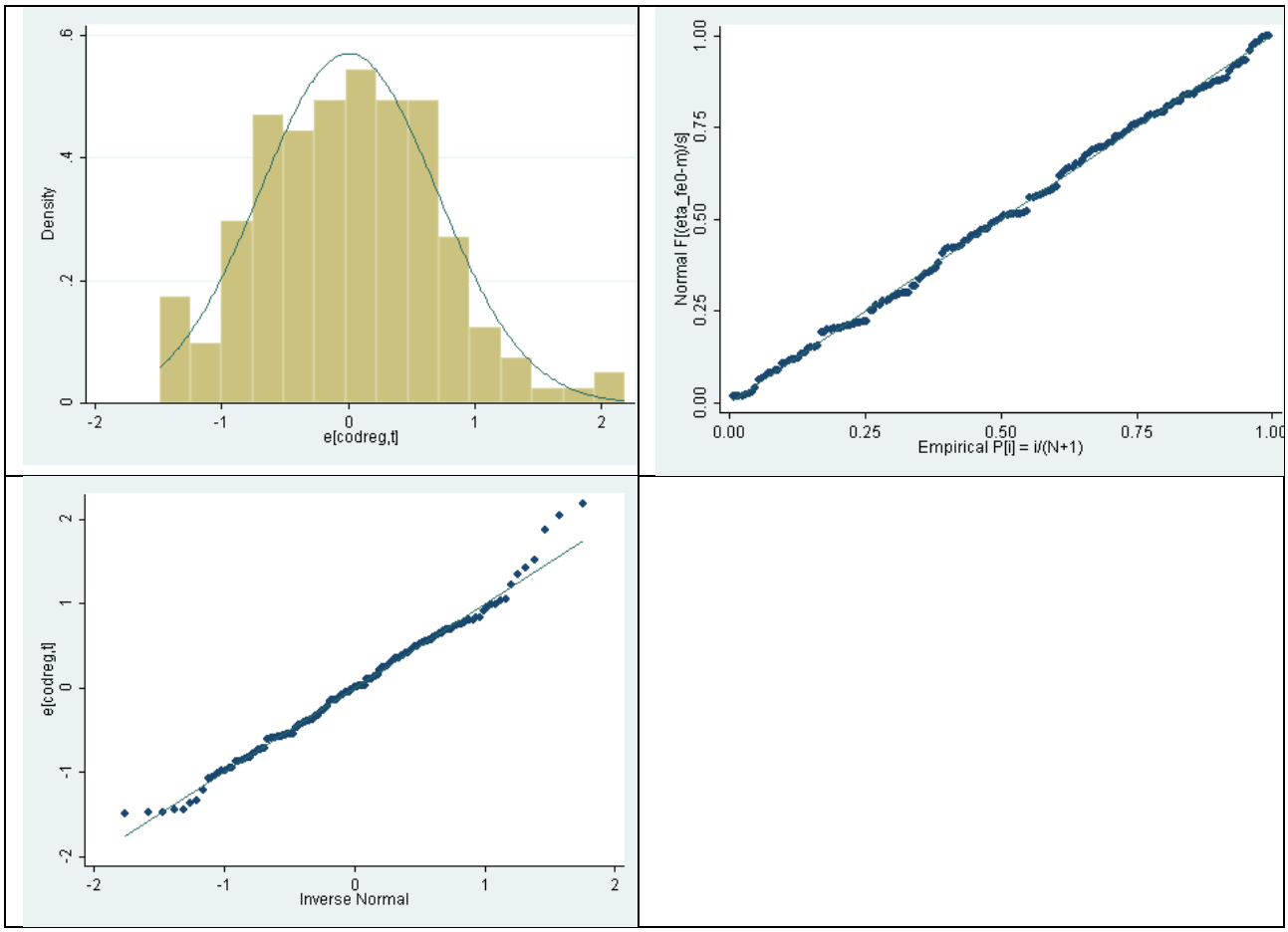
Risultati dell'elaborazione con il software statistico STATA

[85] Tale opzione viene applicata insieme all'opzione cluster, poiché il raggruppamento produce uno stimatore coerente quando i disturbi non sono identicamente distribuiti o vi è una correlazione seriale.

Tabella B7: Valutazione dei modelli Fixed Effect (FE) per i Paesi OLD UE, tramite gli indicatori AIC, BIC e RMSE

Nome Modello	Modello	AIC	BIC	RMSE
FEOLDUE 1	$Y_{it} = \beta_{12}x_{12it} + \beta_1 x_{1it} + \beta_{10} x_{10it} + \beta_7 x_{7it} + \beta_4 x_{4it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_8 x_{8it} + \beta_6 x_{6it} + \alpha_i + \varepsilon_{it}$	394.28	425.35	0.81
FEOLDUE 2	$Y_{it} = \beta_{12}x_{12it} + \beta_7 x_{7it} + \beta_4 x_{4it} + \beta_3 x_{3it} + \beta_8 x_{8it} + \alpha_i + \varepsilon_{it}$	411.31	429.94	0.86
FEOLDUE 3	$Y_{it} = \beta_{10} x_{10it} + \beta_7 x_{7it} + \beta_4 x_{4it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_8 x_{8it} + \beta_6 x_{6it} + D_t + \alpha_i + \varepsilon_{it}$	383.50	436.30	0.77
FEOLDUE 4	$Y_{it} = \beta_{12} x_{12it} + \beta_9 x_{9it} + \beta_{10} x_{10it} + \beta_7 x_{7it} + \beta_8 x_{8it} + \beta_6 x_{6it} + D_t + \alpha_i + \varepsilon_{it}$	432.77	473.15	0.91

Figura B6: Risultati della verifica della normalità dei residui – FE OLDUE3



Risultati dell'elaborazione con il software statistico STATA

Tabella B8: Risultati dei test di verifica della normalità dei residui – FE OLDUE3

Test di normalità					
Test	Osservazioni	W	V	z	Prob>z
Shapiro-Wilk	165	0.98859	1.441	0.832	**0.20263**
Test	Osservazioni	W'	V'	z	Prob>z
Shapiro-Francia	165	0.98922	1.490	0.814	**0.20771**
Test	Osservazioni	Pr(Asimmetria)	Pr(Curtosi)	χ^2	Prob>χ2
Asimmetria/Curtosi	165	0.2294	0.3958	2.20	**0.3335**

Risultati dell'elaborazione con il software statistico STATA

Tabella B9: Modello FE OLDUE3 con procedure robust[86], jackknife e bootstrap

Variabile	Coefficiente di regressione	p-value	p-value robust	p-value jackknife	p-value bootstrap
Disoccupazionegiovanile	0.718857	0.004	0.115	0.291	0.142
Occupazionefemminile	-0.1768448	0.004	0.115	0.241	0.162
Industrializzazione	-0.0367062	0.002	0.168	0.383	0.242
Istruzione di terzolivello	-0.0703668	0.064	0.220	0.499	0.373
Istruzione di secondo livello	-0.2636953	0.000	0.002	0.046	0.003
Pil pro capite	0.0401843	0.000	0.000	0.000	0.000
Tassazioneindiretta	0.2870704	0.005	0.096	0.253	0.145
Dummy2006	-0.4995376	0.056	0.036	0.090	0.039
Dummy 2007	-0.812765	0.003	0.018	0.049	0.020
Dummy2008	-1.10298	0.000	0.001	0.006	0.001
Dummy2010	-0.5338964	0.045	0.039	0.113	0.020
Dummy2011	-0.9485709	0.001	0.005	0.043	0.002
Dummy2012	-0.97377754	0.002	0.023	0.109	0.018
Dummy2013	-1.319977	0.007	0.030	0.158	0.034
Dummy2014	-2.113448	0.001	0.018	0.124	0.022
Dummy2015	-1.407537	0.001	0.033	0.153	0.038
Cons	33.27638	0.000	0.002	0.029	0.001

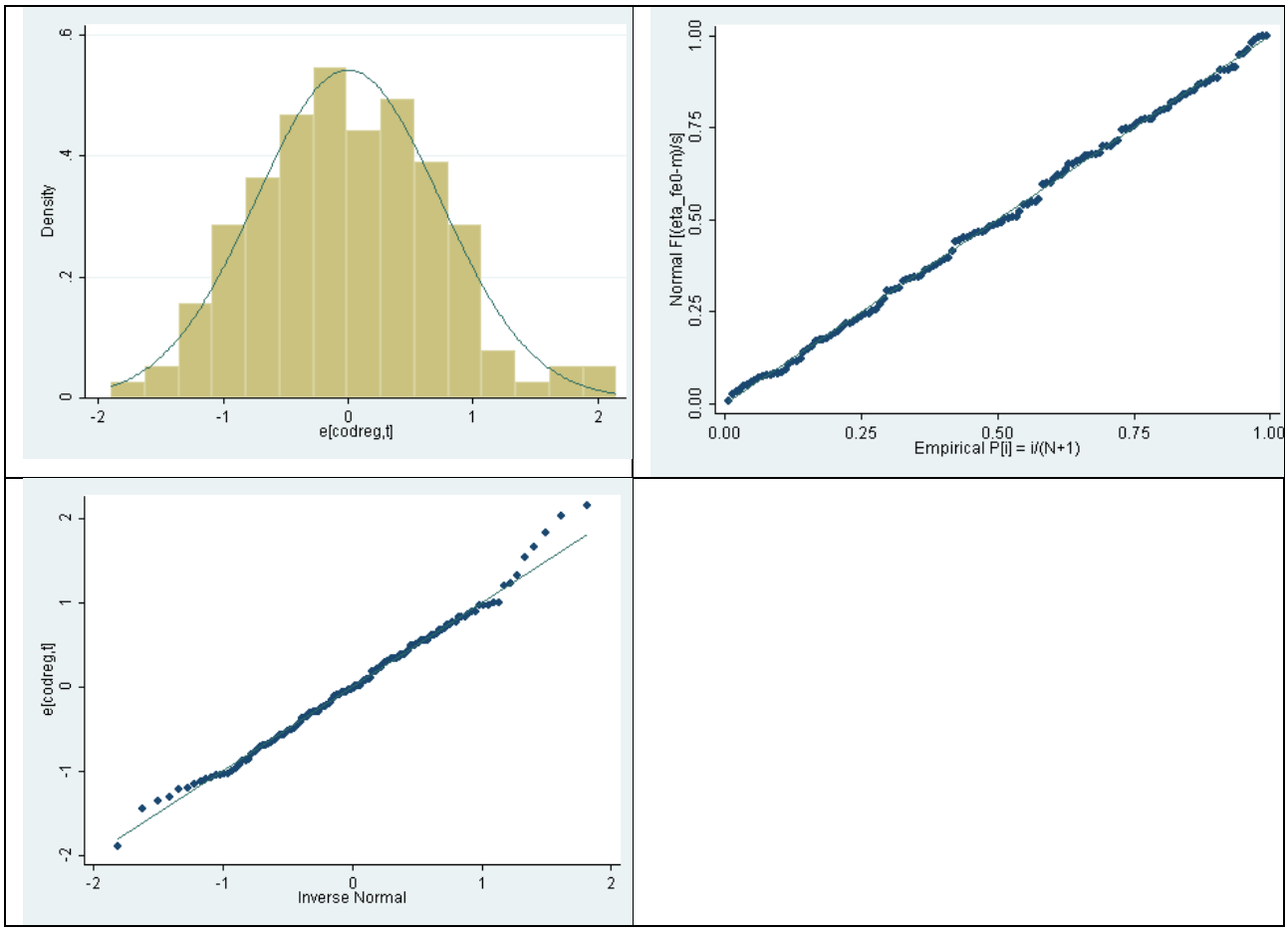
Risultati dell'elaborazione con il software statistico STATA

Tabella B10: Valutazione dei modelli Fixed Effect (FE) per i Paesi NEW UE, tramite gli indicatori AIC, BIC e RMSE

Nome Modello	Modello	AIC	BIC	RMSE
FENEWU E1	$Y_{it} = \beta_{12}x_{12it} + \beta_{10}x_{10it} + \beta_2 x_{2it} + \beta_8 x_{8it} + \beta_6 x_{6it} + \alpha_i + \varepsilon_{it}$	**401.31**	**422.05**	**1.00**
FENEWU E2	$Y_{it} = \beta_{12}x_{12it} + \beta_{11}x_{11it} + \beta_8 x_{8it} + \beta_{10}x_{10it} + \alpha_i + \varepsilon_{it}$	**452.59**	**467.40**	**1.21**
FENEWU E3	$Y_{it} = \beta_{12}x_{12it} + \beta_{10}x_{10it} + \beta_4 x_{4it} + \beta_2 x_{2it} + \beta_5 x_{5it} + \beta_6 x_{6it} + D_t + \alpha_i + \varepsilon_{it}$	**349.26**	**396.67**	**0.82**
FENEWU E 4	$Y_{it} = \beta_{12}x_{12it} + \beta_9 x_{9it} + \beta_{10}x_{10it} + \beta_8 x_{8it} + \beta_6 x_{6it} + D_t + \alpha_i + \varepsilon_{it}$	**395.08**	**436.56**	**0.97**

[86] Tale opzione viene applicata insieme all'opzione cluster poiché il raggruppamento produce uno stimatore coerente quando i disturbi non sono identicamente distribuiti o vi è una correlazione seriale.

Figura B7: Risultati della verifica della normalità dei residui – FE NEWUE3



Risultati dell'elaborazione con il software statistico STATA

Tabella B11: Risultati dei test di verifica della normalità dei residui – FE NEWUE3

Test di normalità					
Test	Osservazioni	W	V	z	Prob>z
Shapiro-Wilk	143	0.99221	0.871	-0.313	**0.62299**
Test	Osservazioni	W'	V'	z	Prob>z
Shapiro-Francia	143	0.99204	0.976	-0.050	**0.51992**
Test	Osservazioni	Pr(Asimmetria)	Pr(Curtosi)	χ^2	Prob>χ2
Asimmetria/Curtosi	143	0.2341	0.6296	1.68	**0.4327**

Risultati dell'elaborazione con il software statistico STATA

Tabella B12: Modello FE NEWUE3 con procedure robust[87], jackknife e bootstrap

Variabile	Coefficiente di regressione	p-value	p-value robust	p-value jackknife	p-value bootstrap
Abitazione	0.103939	0.000	0.000	0.001	0.000
Disoccupazionegiovanile	0.0467055	0.034	0.099	0.167	0.134
Industrializzazione	-0.0665542	0.000	0.001	0.018	0.008
Istruzione di terzolivello	0.1257048	0.049	0.140	0.233	0.205
Tassazione indiretta	0.2917889	0.002	0.103	0.150	0.115
Tassazionediretta	0.2613215	0.009	0.090	0.229	0.120
Dummy2006	-0.962972	0.001	0.000	0.001	0.000
Dummy 2007	-1.63062	0.000	0.000	0.002	0.000
Dummy2008	-1.909833	0.000	0.000	0.001	0.000
Dummy2010	-0.6370982	0.057	0.073	0.135	0.096
Dummy2011	-1.66057	0.000	0.001	0.006	0.001
Dummy2012	-2.261546	0.000	0.001	0.003	0.000
Dummy2013	-2.667136	0.000	0.001	0.002	0.000
Dummy2014	-3.221298	0.000	0.000	0.002	0.000
Dummy2015	-3.16162	0.000	0.002	0.008	0.002
Cons	17.63092	0.000	0.001	0.005	0.001

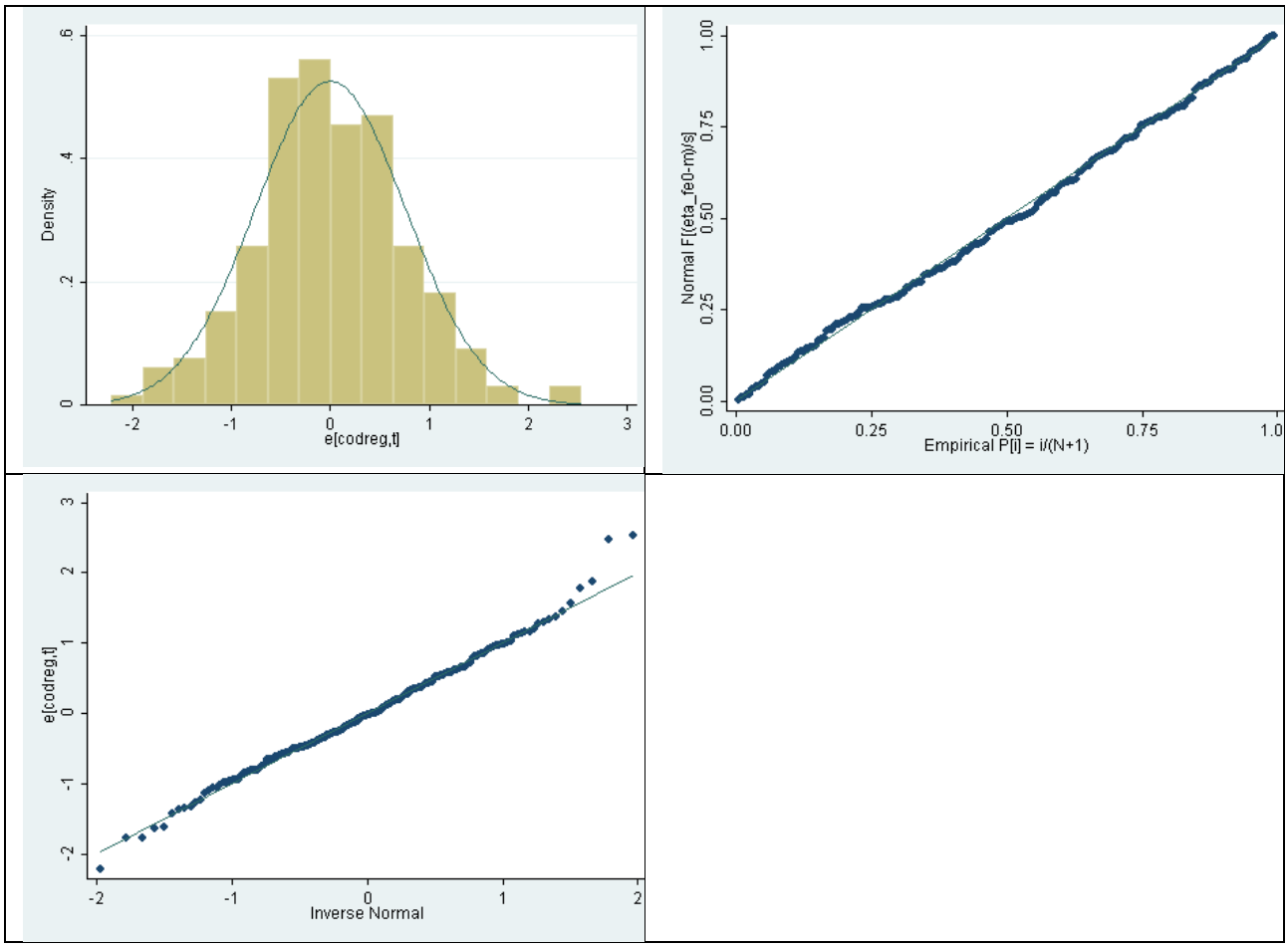
Risultati dell'elaborazione con il software statistico STATA

Tabella B13: Valutazione dei modelli Fixed Effect (FE) per i Paesi AE, tramite gli indicatori AIC, BIC e RMSE

Nome Modello	Modello	AIC	BIC	RMSE
FEAE1	$Y_{it} = \beta_1 x_{1it} + \beta_{10} x_{10it} + \beta_7 x_{7it} + \beta_4 x_{4it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_8 x_{8it} + \beta_6 x_{6it} + \alpha_i + \varepsilon_{it}$	**546.62**	**576.70**	**0.92**
FEAE2	$Y_{it} = \beta_1 x_{1it} + \beta_9 x_{9it} + \beta_{10} x_{10it} + \beta_7 x_{7it} + \beta_4 x_{4it} + \beta_2 x_{2it} + \beta_3 x_{3it} + \beta_8 x_{8it} + \beta_5 x_{5it} + \beta_6 x_{6it} + D_t + \alpha_i + \varepsilon_{it}$	**517.3**	**584.15**	**0.79**
FEAE3	$Y_{it} = \beta_{11} x_{11it} + \beta_9 x_{9it} + \beta_{10} x_{10it} + \beta_5 x_{5it} + \beta_6 x_{6it} + D_t + \alpha_i + \varepsilon_{it}$	**595.47**	**638.91**	**1.02**

[87] Tale opzione viene applicata insieme all'opzione cluster, poiché il raggruppamento produce uno stimatore coerente quando i disturbi non sono identicamente distribuiti o vi è una correlazione seriale.

Figura B8: Risultati della verifica della normalità dei residui – FE AE2



Risultati dell'elaborazione con il software statistico STATA

Tabella B14: Risultati dei test di verifica della normalità dei residui – FE AE2

Test di normalità					
Test	Osservazioni	W	V	z	Prob>z
Shapiro-Wilk	209	0.99353	1.004	0.008	**0.49673**
Test	Osservazioni	W'	V'	z	Prob>z
Shapiro-Francia	209	0.99218	1.322	0.578	**0.28160**
Test	Osservazioni	Pr(Asimmetria)	Pr(Curtosi)	χ^2	Prob>χ2
Asimmetria/Curtosi	209	0.1982	0.1140	4.20	**0.1227**

Risultati dell'elaborazione con il software statistico STATA

Tabella B15: Modello FE AE2 con procedure robust[88], jackknife e bootstrap

Variabile	Coefficiente di regressione	p-value	p-value robust	p-value jackknife	p-value bootstrap
Densità	0.0779926	0.000	0.000	0.002	0.004
DipendentiPubblici	-0.1493823	0.055	0.323	0.783	0.699
Disoccupazionegiovanile	0.075813	0.000	0.002	0.013	0.002
Occupazionefemminile	-0.2361157	0.000	0.003	0.024	0.014
Industrializzazione	-0.0515396	0.000	0.000	0.007	0.003
Istruzione di terzolivello	-0.0922295	0.034	0.018	0.136	0.149
Istruzione di secondo livello	-0.2678007	0.000	0.003	0.045	0.003
PIL pro capite	0.0327438	0.013	0.025	0.282	0.170
Tassazione indiretta	0.252781	0.007	0.097	0.207	0.171
Tassazionediretta	0.189606	0.040	0.196	0.360	0.256
Dummy 2006	-0.5960326	0.018	0.012	0.028	0.007
Dummy 2007	-1.201451	0.000	0.004	0.009	0.002
Dummy2008	-1.293342	0.000	0.001	0.001	0.000
Dummy2010	-0.5296468	0.052	0.024	0.039	0.022
Dummy2011	-1.205034	0.000	0.000	0.001	0.000
Dummy2012	-1.507129	0.000	0.001	0.004	0.001
Dummy2013	-1.647671	0.000	0.001	0.004	0.001
Dummy2014	-2.174703	0.000	0.000	0.000	0.000
Dummy2015	-2.10728	0.000	0.000	0.003	0.001
Cons	28.38126	0.000	0.000	0.009	0.002

Risultati dell'elaborazione con il software statistico STATA

[88] Tale opzione viene applicata insieme all'opzione cluster, poiché il raggruppamento produce uno stimatore coerente quando i disturbi non sono identicamente distribuiti o vi è una correlazione seriale.

Tabella B16: Regressione con AR(1) - Coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FEUE2

Variabile	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza	
Abitazione	0.0313157	0.0159671	1.96	0.051	-0.000142	0.0627734
Densità di popolazione	0.0496903	0.0126165	3.94	0.000	0.0248338	0.0745468
Disoccupazione giovanile	0.0495971	0.0182644	2.72	0.007	0.0136134	0.0855808
Occupazione femminile	-0.0331949	0.062514	-0.53	0.596	-0.156357	0.0899673
Industrializzazione	-0.0661852	0.011089	-5.97	0.000	-0.0880321	-0.0443382
Istruzione di terzo livello	-0.0370302	0.0497323	-0.74	0.457	-0.1350104	0.06095
Istruzione di secondo livello	-0.1069448	0.045289	-2.36	0.019	-0.196171	-0.0177185
Pil pro capite	0.0260437	0.0121583	2.14	0.033	0.00209	0.0499974
Tassazione indiretta	0.1548769	0.0840538	1.84	0.067	-0.010722	0.3204758
Dummy2006	-0.4291993	0.1543615	-2.78	0.006	-0.7333152	-0.1250834
Dummy 2007	-0.8280927	0.2084729	-3.97	0.000	-1.238816	-0.4173692
Dummy2008	-1.195094	0.1850194	-6.46	0.000	-1.559611	-0.8305777
Dummy2010	-0.6424158	0.1592451	-4.03	0.000	-0.9561532	-0.3286785
Dummy2011	-1.356086	0.2147286	-6.32	0.000	-1.779134	-0.9330376
Dummy2012	-1.64587	0.2590327	-6.35	0.000	-2.156205	-1.135536
Dummy2013	-1.854678	0.2954981	-6.28	0.000	-2.436855	-1.272502
Dummy2014	-2.369026	0.3340845	-7.09	0.000	-3.027223	-1.710828
Dummy2015	-2.419241	0.3654711	-6.62	0.000	-3.139275	-1.699206
Cons	19.32266	1.624862	11.89	0.000	16.12143	22.52389

Risultati dell'elaborazione con il software statistico STATA

Tabella B17: Regressione con AR(1) - Coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FEOLDUE3

Variabile	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza	
Disoccupazionegiovanile	0.0309864	0.0296076	1.05	0.297	-0.0276395	0.0896123
Occupazionefemminile	-0.0122996	0.1126034	-0.11	0.913	-0.2352656	0.2106663
Industrializzazione	-0.0469654	0.0182606	-2.57	0.011	-0.0831232	-0.0108076
Istruzione di terzolivello	-0.0369044	0.0565772	-0.65	0.515	-0.1489329	0.0751241
Istruzione di secondo livello	-0.1166933	0.0570983	-2.04	0.043	-0.2297536	-0.0036331
Pil pro capite	0.0174883	0.0135356	1.29	0.199	-0.0093136	0.0442902
Tassazioneindiretta	0.1129068	0.143686	0.79	0.434	-0.1716058	0.3974194
Dummy2006	-0.3531395	0.2330307	-1.52	0.132	-0.8145635	0.1082845
Dummy 2007	-0.6791234	0.3155262	-2.15	0.033	-1.303897	-0.05435
Dummy2008	-1.117069	0.271559	-4.11	0.000	-1.654783	-0.5793547
Dummy2010	-0.5389912	0.2158696	-2.50	0.014	-.9664347	-0.1115478
Dummy2011	-0.9505276	0.2899649	-3.28	0.001	-1.524687	-0.3763682
Dummy2012	-1.030962	0.3564444	-2.89	0.005	-1.736758	-0.3251665
Dummy2013	-1.084741	0.4065579	-2.67	0.009	-1.889766	-0.2797153
Dummy2014	-1.512683	0.4646428	-3.26	0.001	-2.432722	-0.5926433
Dummy2015	-1.746893	0.4909129	-3.56	0.001	-2.718949	-0.7748363
Cons	22.67761	3.483224	6.51	0.000	15.78048	29.57474

Risultati dell'elaborazione con il software statistico STATA

Tabella B18: Regressione con AR(1) - Coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FENEWUE2

Variabile	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza	
Abitazione	0.0435221	0.0183024	2.38	0.019	0.0072193	0.0798248
Disoccupazione giovanile	0.0263217	0.0250162	1.05	0.295	-0.0232979	0.0759412
Industrializzazione	-0.0839109	0.0141452	-5.93	0.000	-0.1119679	-0.0558539
Istruzione di terzo livello	0.0784045	0.1120268	0.70	0.486	-0.1438002	0.3006091
Tassazione indiretta	0.0983699	0.1033842	0.95	0.344	-0.1066923	0.303432
Tassazione diretta	0.3073202	0.1238547	2.48	0.015	0.0616551	0.5529853
Dummy2006	-0.8235141	0.2179553	-3.78	0.000	-1.255827	-0.3912007
Dummy 2007	-1.595753	0.3095918	-5.15	0.000	-2.209827	-0.9816792
Dummy2008	-1.762927	0.2729537	-6.46	0.000	-2.304329	-1.221524
Dummy2010	-0.457348	0.248247	-1.84	0.068	-0.9497448	0.0350488
Dummy2011	-1.399173	0.3543535	-3.95	0.000	-2.102031	-0.6963144
Dummy2012	-1.934935	0.4319344	-4.48	0.000	-2.791675	-1.078195
Dummy2013	-2.301799	0.5107533	-4.51	0.000	-3.314875	-1.288722
Dummy2014	-2.892254	0.5799525	-4.99	0.000	-4.042587	-1.741921
Dummy2015	-2.790072	0.6741419	-4.14	0.000	-4.127229	-1.452915
Cons	24.11203	0.9996496	24.12	0.000	22.12923	26.09483

Risultati dell'elaborazione con il software statistico STATA

Tabella B19: Regressione con AR(1) - Coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FEAE2

Variabile	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Densità	0.0574116	0.0144601	3.97	0.000	0.0288429 0.0859804
DipendentiPubblici	-0.0901557	0.1093699	-0.82	0.411	-0.3062371 0.1259257
Disoccupazionegiovanile	0.0839068	0.0220818	3.80	0.000	0.04028 0.1275336
Occupazionefemminile	-0.0637352	0.0848378	-0.75	0.454	-0.2313487 0.1038783
Industrializzazione	-0.0591735	0.0137813	-4.29	0.000	-0.0864012 -0.0319459
Istruzione di terzolivello	-0.0858641	0.0591646	-1.45	0.149	-0.2027553 0.031027
Istruzione di secondo livello	-0.1680036	0.0556225	-3.02	0.003	-0.2778966 -0.0581107
PIL pro capite	0.0453109	0.0181558	2.50	0.014	0.0094406 0.0811812
Tassazione indiretta	0.1575057	0.1166363	1.35	0.179	-0.0729319 0.3879433
Tassazionediretta	0.2707538	0.1060147	2.55	0.012	0.0613012 0.4802064
Dummy 2006	-0.4045322	0.2156416	-1.88	0.063	-0.8305739 0.0215096
Dummy 2007	-0.9807448	0.289139	-3.39	0.001	-1.551995 -0.4094947
Dummy2008	-1.197579	0.2561684	-4.67	0.000	-1.70369 -0.691469
Dummy2010	-0.5831222	0.2106952	-2.77	0.006	-0.9993913 -0.166853
Dummy2011	-1.261089	0.2720094	-4.64	0.000	-1.798496 -0.7236811
Dummy2012	-1.726211	0.3292727	-5.24	0.000	-2.376753 -1.075669
Dummy2013	-1.986619	0.3789897	-5.24	0.000	-2.735386 -1.237851
Dummy2014	-2.441601	0.4303624	-5.67	0.000	-3.291866 -1.591337
Dummy2015	-2.398768	0.4672075	-5.13	0.000	-3.321827 -1.475709
Cons	16.23424	2.283167	7.11	0.000	11.7234 20.74508

Risultati dell'elaborazione con il software statistico STATA

Tabella B20: Test di Hausmann e Test di Wooldridge

Modello	Test di Hausmann	Test di Wooldridge
UE	χ^2= 186,93; Prob> χ^2 = 0.0000	χ^2= 214,357; Prob> χ^2 = 0.0000
OLDUE	χ^2= 230,57; Prob> χ^2 = 0.0000	χ^2= 56,889; Prob> χ^2 = 0.0000
NEWUE	χ^2= 220,54; Prob> χ^2 = 0.0000	χ^2= 56,659; Prob> χ^2 = 0.0000
AE	χ^2= 78,58; Prob> χ^2 = 0.0000	χ^2= 146,168; Prob> χ^2 = 0.0000

Risultati dell'elaborazione con il software statistico STAT

Tabella B21: Consistenza del modello FE – Coefficiente di determinazione, correlazione residui/variabile esplicativa, test F

Modello	R-square within	R-square between	R-square overall	Corr(ui,Xb)	Test F
FE UE2	0,7262	0,0281	0,0309	0,9366	F(27,262)=117,18; Prob>F=0,0000
FE OLDUE3	0,5896	0,2098	0,2213	0,1646	F(14,134)=87,08; Prob>F=0,0000
FE NEWUE3	0,8340	0,0572	0,1369	0,0010	F(12,115)=268,76; Prob>F=0,0000
FE AE2	0,6971	0,0615	0,0626	-0,9634	F(18,171)=135,08; Prob>F=0,0000

Risultati dell'elaborazione con il software statistico STATA

BIBLIOGRAFIA CAPITOLO 4

Achim, M.V. et al. (2018). Rethinking The Shadow Economy in terms of happines. Evidence for the European Union Members States. TECHNOLOGICAL AND ECONOMIC DEVELOPMENT OF ECONOMY. 2018 Volume 24(1): 199–228.

Amendola, A., e Dell'Anno, R. (2008). Istituzioni, Disuguaglianza ed Economia Sommersa: quale relazione? Quaderno n. 24/2008, Dipartimento di Scienze Economiche, Matematiche e Statistiche; Università degli Studi di Foggia.

Bergheim, S. (2007). The happy variety of capitalism. Deutsche Bank Research, April 25, 1–22.

Busato, F., e Chiarini, B. (2004). Market and Underground Activities in a Two Sector Dynamic Equilibrium Model. Economic Theory, 23(4): 831-861.

Cappariello, R., e Zizza, R. (2009). Dropping the Books and Working Off the Books. Temi di discussione, Working Papers, Ufficio Studi Banca d'Italia, n. 702.

Daniele, V., e Marani, U. (2008). Criminalità e investimenti esteri. Un'analisi per le province italiane. Working Paper, Università Magna Graecia di Catanzaro.

Dell'Anno, R. (2003). Stimare l'economia sommersa con un approccio ad equazioni strutturali. Un'applicazioneall'economiaitaliana (1962-2000). SIEP.

Falkner, G. e Treib, O. (2007). Three Worlds of Compliance or Four?The EU15 Compared to New Member States. Institute for Advanced Studies, Vienna.

Frey, B. - Weck-Hanmeman, H. (1984), "The hidden economy as an unobserved variable". European Economic Review n. 26/1.

Frey, B. S.; Stutzer, A. 2012. The use of happiness research for public policy, Social Choice and Welfare 38(4): 659–674. http://dx.doi.org/10.1007/s00355-011-0629-z

Goetz, K.H. (2001). Making sense of post-communist central administration:modernization, Europeanization or Latinization? Journal of European Public Policy 8:6 December: 1032–1051

Klinglmair, R. e Schneider, F. (2004). Shadow Economies Around the World: What Do We Know? Working Paper No. 1167. Institute for the Study of Labor.IZA Discussion Paper No. 1043.

Lisi, G. (2009). Underground Employment and Unemployment in the Regions of Italy: A panel analysis. University of Cassino (Italy). MPRA Paper No. 18525.

Lisi, G. (2010). Underground Employment and Unemployment in Italy: A panel analysis. University of Cassino (Italy). MPRA Paper No. 22508.

Lucifora, C. (2003). Economia sommersa e lavoro nero. Il Mulino.

Mahadeo, J.D. et al. (2012). Board Composition and Financial Performance: Uncoveringthe Effects of Diversity in an Emerging Economy. Journal of Business Ethics · February 2011.

Medina, L. e Schneider, F. (2017). Shadow Economies Around the World: What Did We Learn Over the Last 20 Years? WP/18/17 – IMF Working Paper

Morvillo, C. (2016). Evoluzione delle determinanti dell'economia sommersa: analisi panel di regioni italiane. Ministero dell'Economia e delle Finanze. Dipartimento del Tesoro. Nota tematica 1/2016.

OECD. 2012. Communication from the Commission to the European Parliament and the council on concrete ways to reinforce the fight against tax fraud and tax evasion including in relation to third countries. COM/2012/0351.

Schmölders, G. 1960. Fiscal psychology: a new branch of public finance, National Tax Journal 12: 340–345.

Schneider, F. (2004). The Size of the Shadow Economies of 145 Countries all over the World: First Results over the Period 1999 to 2003. Papers, No. 1431, Institute for the Study of Labor (IZA), Bonn.

Schneider, F. (2005). Shadow Economies around the World: What do we really know? IAW Diskussionspapiere, No. 16, Institut für Angewandte Wirtschaftsforschung (IAW), Tübingen

Schneider, F. (2013). Size and Development of the Shadow Economy of 31 European and 5 other OECD Countries from 2003 to 2013: A Further Decline.

Schneider, F. (2013). Size and Development of the Shadow Economy of 31 European and 5 other OECD Countries from 2003 to 2012: Some New Facts. ResearchGate

Schneider, F. (2013). The Shadow Economy in Europe. ATKearny.

Schneider, F. e Williams, C. (2013). The Shadow Economy.The Institute of Economic Affairs.
Schneider, F. e Buehn, C. (2016). Estimating the Size of the Shadow Economy: Methods, Problems and Open Questions. Discussion Paper No. 9820.The Institute for the Study of Labor (IZA) in Bonn.
Tanzi, V. (2013). L'economia sotterranea degli Stati Uniti: stime e implicazioni. Moneta e Credito.
Thießen, U. (2010). The shadow economy in international comparison: options for economic policy derived from an OECD panel analysis. Discussion Papers, Deutsches Institut für Wirtschaftsforschung, 1–72.
Voicu, C. (2012). Underground economy nature – conceptual status, Theoretical and Applied Economics XIX 3(568): 109–120.

5 – Considerazioni teoriche e ulteriori sviluppi metodologici

In questo capitolo il fenomeno dell'Economia Sommersa viene analizzato sviluppando alcuni modelli dinamici dei panel dati. Vengono inoltre esplicitate alcune riflessioni di policy implication.

Una strategia utilizzata per rispondere al problema di possibile presenza di endogeneità, che viene applicata soprattutto nello studio dei fenomeni sociali ed economici (Bellemare, Masaki, Pepinsky, 2017), è quella di ritardare le variabili esplicative. Tale strategia è conosciuta come "identificazione del ritardo" ("lag identification"). I ricercatori utilizzano questa metodologia nei seguenti casi: i) teorico, ossia quando ci si aspetta che l'effetto di una variabile esplicativa utilizzi un ritardo di un periodo; ii) statistico, ossia quando la variabile dipendente ritardata viene utilizzata in una funzione statistica, come ad esempio nell'analisi panel dinamica; iii) di identificazione, ossia quando i ricercatori studiano l'effetto di una variabile esplicativa X su una variabile dipendente Y, e viene utilizzato il valore ritardato della variabile esplicativa X per renderla esogena.

Tale strategia viene impiegata molto frequentemente in quanto è necessario avere semplicemente a disposizione il dataset. In questo studio verrà utilizzata a fini statistici attraverso una analisi panel dinamica.

5.1 I modelli panel dinamici

Prendendo come riferimento la singola osservazione e limitando per semplicità la trattazione ai modelli con un solo ritardo, l'equazione generale per un panel dinamico è espressa attraverso la seguente formula:

$$y_{it} = X'_{it}\beta + \phi y_{it-1} + u_{it}$$

dove $u_{it} = \mu_i + \varepsilon_{it}$ e ϕ è il parametro relativo alla componente autoregressiva del modello.

Il problema principale di questo tipo di modelli è dato dal fatto che il termine di errore non è incorrelato con la variabile ritardata, ciò genera stime OLS e GLS inconsistenti. La soluzione a tale inconveniente è quella di considerare un modello in termini di differenze prime e ricorrere allo stimatore a variabili strumentali (Baltagi, 2005; Arellano e Bond, 1991).

La stima con il metodo delle variabili strumentali è utilizzata nell'analisi di regressione lineare. Un'ipotesi standard del modello classico di regressione lineare è che le variabili esplicative non siano correlate con l'errore. In caso contrario sorge un problema di endogeneità dei regressori che rende difficile ottenere stime consistenti con il metodo dei minimi quadrati. Se è disponibile una variabile strumentale, allora è possibile ottenere stime consistenti. Una variabile è definita strumento perché non influenza direttamente la variabile dipendente Y, ma indirettamente attraverso il suo effetto sui regressori. Un insieme di strumenti $Z_1, Z_2, \ldots, Z_n$ deve soddisfare due condizioni per essere valido:

- Rilevanza: lo strumento è correlato con la X $Cor(Zi,Xi)\neq 0$;
- Esogeneità: la parte della variazione di Xi, catturata dalla variabile strumentale è esogena $E(Zi,ui)=0$.

Nello specifico, considerando il seguente modello classico di regressione:

$$y_i = \beta x_i + \varepsilon_i \qquad i=1, \ldots. ,N$$

lo stimatore sarà:

$$\hat{\beta}_{OLS} = \beta + \Sigma_i x_i \varepsilon_i / \Sigma_i x^2_i$$

Poiché x_i e ε_i sono incorrelati, passando al limite per $N\rightarrow\infty$ il secondo termine nell'espressione sopra converge a zero in probabilità, così che la stima $\hat{\beta}$ sia consistente. In caso contrario può essere utile considerare una variabile strumentale, tale da soddisfare la condizione: $\Sigma_i z_i(y_i - \beta x_i) = 0$ ottenendo così lo stimatore delle variabili strumentali consistente

$$\hat{\beta}_{IV}=\beta + \Sigma_i z_i \varepsilon_i / \Sigma_i z_i x_i$$

Gli stimatori più utilizzati e conosciuti sono lo stimatore di Arellano-Bond (Difference GMM estimator) (Arellano e Bond, 1991) e Blundell-Bond (System GMM estimator) (Blundell e Bond, 1998; Roodman,2009). Tali metodi di stima sono basati sul metodo dei momenti generalizzato GMM. L'idea è di rimuovere l'effetto individuale trasformando l'equazione originale in un'equazione in differenze prime e di usare appropriati ritardi delle variabili endogene come strumenti. In particolare, lo stimatore di Arellano e Bond (1991), come primo passo, trasforma tutti i regressori tramite differenziazione e usa il metodo generalizzato dei momenti. Per questo motivo viene chiamato Difference GMM. Lo stimatore di Arellano-Bover (1995) e Blundell-Bond (1998) costruisce invece un sistema di due equazioni, l'equazione originale e quella trasformata, per questa ragione viene denominato System GMM.

Arellano-Bond (Difference GMM estimator) (Arellano e Bond, 1991; Roodman, 2009)

Dato il seguente modello auto regressivo puro, nel quale i regressori esogeni sono omessi (β=0):

$$y_{it} = \phi y_{it-1}+\mu_i+\varepsilon_{it}$$

Le ipotesi alla base di questo metodo di stima sono T è fisso; $N \rightarrow \infty$; $\varepsilon_{it}\sim N(0, \sigma_\varepsilon^2)$.

Il modello in differenze prime sarà:

$$\Delta y_{it}= \phi\Delta y_{it-1}+ \Delta\varepsilon_{it} = \phi(y_{it-1}- y_{it-2})+ \varepsilon_{it}-\varepsilon_{it-1}, \text{ dove } \Delta\varepsilon_{it}\sim MA(1), i=1,..,N \text{ e } t=3,..,T.$$

L'equazione precedente equivale a un sistema di T-2 equazioni con N osservazioni:

$\Delta y_{i3} =\phi \Delta y_{i2} + \Delta\varepsilon_{i3}$ strumenti Δy_{i1}

$\Delta y_{i4} =\phi \Delta y_{i3} + \Delta\varepsilon_{i4}$ strumenti $\Delta y_{i1}\ \Delta y_{i2}$

……..

$\Delta y_{iT} =\phi \Delta y_{iT-1} + \Delta\varepsilon_{iT}$ strumenti $\Delta y_{i1}\ \Delta y_{i2}\ldots\ldots\ \Delta y_{iT-2}$

dove gli strumenti sono selezionati in base alla loro proprietà di essere incorrelati coi termini di errore. In questo modo si ottiene una stima consistente del modello dinamico.

Il modello in differenze prime può essere riscritto nella forma compatta come $\Delta Y_t= \phi\Delta Y_{t-1}+ \Delta\varepsilon_t$, dove ϕ è un parametro scalare. Questo modello è caratterizzato dalla presenza di correlazione, tra l'errore e i regressori, e di eteroschedasticità. Arellano e Bond superano tali inconvenienti inserendo gli strumenti nell'equazione in questo modo:

$$Z'\Delta Y_t= \phi Z'\Delta Y_{t-1}+ Z'\Delta\varepsilon_t$$

dove Z rappresenta gli strumenti[89].

L'espressione appena scritta rappresenta lo stimatore di Arellano e Bond one step, che è uno stimatore GLS del tipo $\widehat{\phi} =(\Delta Y'_{t-1} Z \Omega^{-1} Z' \Delta Y'_{t-1})^{-1} \Delta Y'_{t-1} Z \Omega^{-1} Z' \Delta Y'_{t-1}$ dove $\Omega=Var(Z' \Delta\varepsilon)= \sigma_{\varepsilon}^2 Z'(I_N \otimes Vi)Z$ è la matrice di varianze e covarianze che dipende dalla presenza di N individui[90].

Lo stimatore di Arellano e Bond two step si ottiene sostituendo la matrice dei momenti secondi della popolazione $V_i=E(\Delta\varepsilon\Delta\varepsilon')$ con quella dei momenti secondi campionari data da $W_i=E(\Delta\hat{\varepsilon}\Delta\hat{\varepsilon}')$.

I due stimatori sono asintoticamente equivalenti per $N\rightarrow\infty$.

Blundell-Bond (System GMM estimator) (Blundell e Bond, 1998; Roodman, 2009)

Sebbene lo stimatore GMM (basato sull'equazione in differenze prime) fornisca delle stime consistenti, Arellano e Bover (1995) e Blundell e Bond (1998) mostrano come, sulla base di simulazioni statistiche, questo stimatore presenti una scarsa precisione in campione finiti. In particolare, la trasformazione in differenze con strumenti pari ai ritardi nei livelli, introdotta da Arellano e Bond, viene meno se l'esplicativa del modello è una variabile integrata. Infatti, in generale, se l'esplicativa x_{it} è I(1), il coefficiente di correlazione fra Δx_{it} e x_{it-1} è quasi nullo perché le differenze prime di una variabile integrata, per definizione, non possono essere correlate con i suoi livelli. Invece Δx_{it} e Δx_{it-1} possono essere correlate, a meno che x_{it} sia random walk. Inoltre, il coefficiente di correlazione fra x_{it} e x_{it-1} sarà sempre quasi unitario per qualsiasi x_{it} integrato. L'intuizione sottostante è data dal fatto che quando le variabili esplicative sono persistenti nel tempo, i valori ritardati di queste variabili sono debolmente correlati con le loro differenze nell'equazione di regressione differenziata. Per aumentare la precisione delle stime, Arellano e Bover (1995) e Blundell e Bond (1998) propongono di combinare la regressione differenziata con la regressione originale in livello. Gli strumenti per la regressione in differenze sono quelli descritti per lo stimatore Arellano e Bond (1991), mentre gli strumenti per la regressione in livello sono i valori ritardati delle variabili dipendenti differenziate. Lo svantaggio di quest'ultimo stimatore è che esso riduce la dimensione del campione a causa della sensibilità degli strumenti interni (le variabili esplicative ritardate), inoltre, le sue proprietà in campioni di piccola dimensione sono generalmente sconosciute.

Implicazioni di policy (Castaldo, A., Fiorini, A., Maggi, B., 2018)

È possibile capire come un intervento pubblico possa influire sul fenomeno dell'Economia Sommersa? Proviamo ad affrontare questo quesito esponendo alcune delle possibili implicazioni politiche derivanti dalle stime ottenute. Per fare questo, è necessario introdurre alcuni concetti, che aiutano a studiare gli effetti delle componenti sull'Economia Sommersa nel tempo, quali: i parametri di lungo periodo, la velocità di aggiustamento e il ritardo medio. In questo modo si avranno indicazioni più esplicite sulle variazioni dell'Economia Sommersa nel breve, medio e lungo periodo, a seguito dell'aumento percentuale delle sue determinanti. Poiché tali concetti non hanno uno specifico intervallo di tempo definito, verranno spiegati attraverso il modello con ritardo distribuito.

[89] La matrice degli strumenti Z_i, di dimensione (T-2)xC, dove $C=\sum_{j=1}^{T-2} j$ contiene in ogni riga gli strumenti validi per ciascun istante nel tempo t=3,4,….,T.

[90] La matrice delle varianze e covarianze di $\Delta\varepsilon_{iT}$ nella forma matriciale, per l'individuo iesimo, è una matrice quadrata e simmetrica di dimensione (T-2)(T+2), espressa da $V_i=E(\Delta\varepsilon_i \Delta\varepsilon'_{iT})$. Considerando il modello nella forma generale, la matrice di varianza e covarianza è data da $V=I_N \otimes Vi$.

Un modello con ritardo distribuito è un modello nel quale l'effetto di un regressore x sulla variabile oggetto di studio y si verifica nel tempo. l modello è espresso formalmente dalla seguente equazione:

$$Y_t=\alpha+\beta_0x_t+\beta_1x_{t-1}+\beta_2x_{t-2}+....+\beta_kx_{t-k}+e_t$$

dove e_t è il termine di errore, β_0 è il moltiplicatore di breve periodo, in quanto fornisce la variazione immediata di y al variare di una unità di x nello stesso periodo. Se la variazione di x è la stessa, allora $(\beta_0 + \beta_1)$ fornisce una variazione nel valore medio di Y nel periodo successivo e $(\beta_0 + \beta_1 +\beta_2)$ nel periodo successivo ancora. Le somme parziali dei pesi del ritardo β_t vengono definite moltiplicatori intermedi, mentre le somme totali dei pesi del ritardo $\sum_{i=0}^{k}\beta_i= \beta_0 + \beta_1 +……+\beta_k=B$ viene definito moltiplicatore di lungo periodo. L'espressione standardizzata $\beta_i^*= \beta_i/\sum_{i=0}^{k}\beta_i$ fornisce la proporzione dell'impatto di lungo periodo su un certo periodo di tempo. È possibile aggiungere al modello con ritardo distribuito la variabile ritardata di y attraverso il metodo di Koyck, definito modello con ritardo geometrico.

Si aggiunge al modello appena esposto un parametro $0<\lambda<1$ in modo tale che $\beta_i= \beta_0 \lambda^i$. Formalmente il modello sarà dato da:

$$Y_t=\alpha+\beta_0x_t+\beta_0\lambda x_{t-1}+\beta_0\lambda^2x_{t-2}+....+\beta_k\lambda^kx_{t-k}+e_t$$

oppure:

$$Y_t=\alpha+\beta_0\lambda^0x_t+\beta_0\lambda^1x_{t-1}+\beta_0\lambda^2x_{t-2}+....+\beta_k\lambda^kx_{t-k}+e_t$$

dove:

$$\sum_{k=0}^{\infty}\beta_k = \beta_0(1+ \lambda+ \lambda^2 + \lambda^3 +....)= \beta_0(1/1- \lambda)$$

è il moltiplicatore di lungo periodo.

Si ottiene così:

$$Y_{t-1}=\alpha+\beta_0x_{t-1}+\beta_0\lambda x_{t-2}+\beta_0\lambda^2x_{t-3}+....+\beta_k\lambda^kx_{t-k}+e_t$$

$$\lambda Y_{t-1}=\lambda\alpha+\beta_0\lambda x_{t-1}+\beta_0\lambda^2x_{t-2}+\beta_0\lambda^3x_{t-3}+....+\beta_k\lambda^kx_{t-k}+\lambda e_t$$

Attraverso la trasformazione di Koyck, che trasforma un modello con ritardo distribuito in un modello auto regressivo, si ottiene:

$$(Y_t -\lambda Y_{t-1})= (\alpha-\lambda\alpha)+\beta_0x_t+(e_t -\lambda e_t)$$

quindi:

$$Y_t = \alpha(1-\lambda)+\beta_0x_t+ \lambda Y_{t-1} +\nu_t \text{ dove } \nu_t \sim iid \sim (0,\sigma^2)$$

e

$$Y_t = \theta_0+\theta_1x_t+ \lambda Y_{t-1}+\nu_t$$

dove Y_{t-1} è il termine di breve periodo, costruito su un modello auto regressivo. Il moltiplicatore di lungo periodo può essere ottenuto dal modello auto regressivo attraverso la seguente espressione: $\theta_0[1/(1- \lambda]$.

Uno sviluppo ulteriore del modello di Koyck è il modello di aggiustamento parziale. Sia

$$Y^*_t = \beta_0+\beta_1 x_t+ e_t$$

la funzione di lungo periodo e Y^* il livello non osservato e desiderato. Ponendo

$$Y_t- Y_{t-1}=\delta(Y^*_t- Y_{t-1})$$

dove $0< \delta<1$ è il coefficiente di aggiustamento, ne segue che

$$Y_t=\delta Y^*_t + (1-\delta)Y_{t-1}$$

Sostituendo si avrà

$$Y_t=\delta[\beta_0+\beta_1 x_t+ e_t]+ (1-\delta)Y_{t-1}$$

cioè

$$Y_t=\delta\beta_0+\delta\beta_1 x_t+ (1-\delta)Y_{t-1} +e_t$$

che è la funzione di breve periodo. Pertanto

$$Y_t=\theta_0+\theta_1 x_t+ (1-\delta)Y_{t-1} +\nu_t \quad \text{dove} \quad \nu_t \sim \text{iid} \sim (0,\sigma^2).$$

Ricordando la formulazione del modello panel dinamico

$$Y_{it}= \phi Y_{t-1} +\beta X'_{it}+ \mu_i+\varepsilon_{it}$$

la precedente formula non è altro che il nostro modello panel dinamico espresso in termini di modello con ritardo geometrico. Il parametro ϕ, che rappresenta la persistenza del fenomeno oggetto di studio, è pari a $(1-\delta)$, quindi $\delta= 1- \phi$, espressione che lega il coefficiente di aggiustamento alla persistenza del fenomeno.

Per calcolare il ritardo di medio periodo, seguendo Castaldo, Fiorini, Maggi, (2018), è possibile imporre $\theta=1/\delta$, dove con θ si indica il ritardo di medio periodo.

5.2 I due casi studio: considerazioni e applicazioni

In tutti i modelli studiati nei Capitoli 3 e 4[91], l'analisi dei residui ha fatto emergere qualche criticità in merito all'esistenza di autocorrelazione di primo ordine. L'analisi è stata svolta attraverso lo studio della correlazione dei residui, dell'autocorrelazione a coppie e dei test di Pesaran e Wooldridge. I risultati[92]hanno mostrato l'esistenza di autocorrelazione di primo ordine per tutti i modelli, ed esistenza di correlazione, autocorrelazione a coppie e correlazione seriale per il solo modello FE6(irr1) (caso italiano con N=20, T=15 e variabile dipendente lavoro irregolare in termini di occupati).

I modelli panel dinamici vengono applicati al modello FE6(irr1), nel quale tutte le analisi svolte sui residui hanno confermato la presenza di correlazione seriale. Tale necessità viene avvalorata anche dall'affermazione di Wooldridge (2002), in base alla quale: "*L'eteroschedasticità è sempre un potenziale problema, la correlazione seriale può esserlo in determinate applicazioni, ad esempio quando T è molto grande"*. Infatti, il modello in questione, ha una sequenza temporale molto grande (T = 15).

[91] Caso italiano: modello FE1(irr), FE3(irr1) e FE6(irr1) sviluppati nel Capitolo 3- Caso europeo: modello FEUE2, FEOLDUE3, FENEWUE3 e FEAE2 sviluppati nel Capitolo 4.

[92] Tabella C1-C7 in appendice.

Per i restanti modelli[93] è emersa solo una problematica di autocorrelazione di primo ordine, che è stata affrontata e risolta nei capitoli precedenti[94]. Nell'intento di ottenere maggiori spunti di riflessione, verrà comunque applicata la metodologia GMM al modello FEUE2.

IL CASO ITALIANO: MODELLO FE6(irr1)

Come anticipato nei paragrafi precedenti, l'applicazione di tale metodologia appare interessante anche per una lettura dinamica del fenomeno. Infatti, un simile strumento consente di studiare gli effetti di persistenza nella variabile Economia Sommersa, approfondendo così il suo comportamento nel breve, nel medio e nel lungo periodo. Per enfatizzare tale prospettiva, tutte le variabili vengono trasformate in logaritmo. Questo semplice passaggio matematico consente di introdurre il concetto di elasticità, molto utilizzato in ambito economico. L'elasticità è una misura della sensibilità della variabile dipendente (Y) rispetto a variazioni della variabile esplicativa (X). In sostanza il coefficiente di X viene interpretato in termini di elasticità di Y rispetto ad X: una variazione pari all'1% di X determina una variazione pari a β% di Y.

Tabella 1: Stime panel data dinamico (two step difference GMM): coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE6(irr1)[95]

Variabili espresse in logaritmi	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Economia sommersa con un ritardo	1.060284	0.23673	4.48	0.000	0.5648026 1.555766
Densità di popolazione	0.4069757	0.562214	0.72	0.478	-0.7697517 1.583703
Istruzione terzo livello	0.1935466	0.0863826	2.24	0.037	0.0127458 0.3743473
Tassazione	-0.1947124	0.2044127	-0.95	0.353	-0.6225532 0.2331284
Imprenditorialità femminile	-0.646124	0.5577176	-1.16	0.261	-1.81344 0.5211924
Natalità delle imprese	0.0205309	0.069051	0.30	0.769	-0.1239945 0.1650563
Dipendenti pubblici	-0.083422	0.1553205	-0.54	0.597	-0.4085115 0.2416675
Disoccupazione giovanile	0.0009606	0.0223965	0.04	0.966	-0.0459158 0.0478371
Rapine	-0.0293344	0.03687	-0.80	0.436	-0.1065042 0.0478353
Nro strumenti<Nro gruppi	10<20				
F-test	F(8, 19) = 26.32 Prob > F = 0.000				
Test AR(1)	z = -2.99 Pr > z = 0.003				
Test AR(2)	z = -1.00 Pr > z = 0.318				
Test Sargan	χ^2 = 2.80 Prob > χ^2 = 0.094				
Test Hansen	χ^2 = 2.22 Prob > χ^2 = 0.137				

Risultati dell'elaborazione con il software statistico STATA

[93] Caso italiano: modello FE1(irr) e FE3(irr1) sviluppati nel Capitolo 3- Caso europeo: modello FEUE2, FEOLDUE3, FENEWUE3 e FEAE2 sviluppati nel Capitolo4.

[94] Nei Capitoli 3 e 4 sono state svolte le applicazioni per i modelli fixed effect con la correzione dell'eteroschedasticità, della correlazione e la correzione congiunta eteroschedasticità/correlazione dei residui.

[95] In questa applicazione vengono escluse le dummy in quanto determinano un numero eccessivo di strumenti rispetto alle entità, violando la regola di Roodman (2009), in base alla quale il numero di strumenti deve essere inferiore al numero delle entità oggetto di analisi.

Si è ipotizzata la stretta esogeneità dei regressori e sono stati adottati come strumenti per la variabile dipendente ritardata il tasso di lavoro irregolare[96], nell'accezione irr1, ritardato al tempo 2 ed al tempo 3 ((t-2) e (t-3))[97].

Nella Tabella 1 vengono presentati i risultati del modello di Arellano-Bond (AB) two step. Lo stimatore AB two-step, asintoticamente più efficiente dello stimatore one-step, e robusto all'eteroschedasticità e all'autocorrelazione (Roodman, 2009)[98], conferma i risultati dello stimatore AB one-step. I test di specificazione supportano totalmente il modello prescelto. Sono stati infatti eseguiti i test necessari per poter determinare la consistenza delle stime ottenute. Il test AR(2), necessario per accettare la regressione nel suo insieme, verifica l'ipotesi nulla che non vi sia autocorrelazione di secondo ordine tra i residui di regressione. I test di Hansen e Sargan[99], verificano l'ipotesi che le restrizioni in sovra-identificazione siano tutte valide. L'ipotesi testata è che le variabili strumentali siano specificate correttamente, e quindi siano strumenti accettabili ed efficientemente utilizzabili nel modello. Nello specifico, i test di Hansen e Sargan accettano la validità degli strumenti, e il test di autocorrelazione di secondo ordine degli errori indica che l'errore non è serialmente correlato. L'errore di primo ordine nei livelli non ha componenti auto regressive e si giustifica quindi la scelta degli strumenti.

Osservando i coefficienti delle variabili, risulta significativo e positivo il coefficiente della variabile dipendente ritardata e dell'istruzione di terzo livello. Un valore significativo di rho, pari a 1,06, dimostra la presenza di una persistenza del fenomeno nelle regioni. Inoltre, un aumento dell'1% dell'istruzione di terzo livello determina, ceteris paribus, una variazione positiva dello 0,19% dell'Economia Sommersa nel breve periodo, all'1% di livello di significatività. Le restanti variabili non forniscono una interpretazione statisticamente significativa di breve periodo.

Gli effetti di lungo periodo vengono calcolati solo per l'unica variabile risultata statisticamente significativa, ossia l'istruzione di terzo livello, poiché mantenere anche le variabili non statisticamente significative aumenta la variabilità della stima puntuale, conservando però la correttezza e la consistenza.

[96] Se gli errori non sono autocorrelati, qualsiasi variabile $x_{i,t-s}$ con s >=1, risulterà essere uno strumento valido di $x_{i,t}$. Affinché uno strumento sia valido devono valere le seguenti ipotesi: i) $E(x_{i,t}, x_{i,t-s}) \neq 0$, con s >=1, lo strumento deve essere correlato con la variabile da strumentare; ii) $E(x_{i,t-s}, e_{i,t}) = 0$, lo strumento non deve essere correlato con l'errore. Ovviamente maggiore è il valore di T, maggiore è il numero di strumenti utilizzabili. Inoltre, condizione (necessaria) di identificazione dei parametri è che il numero di strumenti esogeni sia uguale o superiore al numero di variabili endogene da strumentare. Nel caso specifico in esame, modello FE6(irr1), i test AR(1) e AR(2) hanno permesso di verificare l'assenza di autocorrelazione dei residui, pertanto viene considerato come strumento la variabile endogena con numero di ritardi pari ad ¼ della lunghezza della serie (t-2) e (t-3) verificando le condizioni precedenti.

[97] Gli stimatori GMM con troppe restrizioni possono avere scarsi risultati in piccoli campioni (Kiviet, 1995). Per superare tale criticità è possibile impostare un numero massimo di livelli ritardati da includere come strumenti per variabili ritardate o predeterminate. La scelta del numero dei ritardi è una questione squisitamente empirica. Una regola (Roodman, 2009) molto usata è quella di inserire un numero di ritardi pari a circa 1/3 o 1/4 della lunghezza della serie, in modo tale che il numero degli strumenti sia sempre inferiore al numero delle unità. Per raggiungere questo obiettivo è necessario limitare il numero dei lag degli strumenti usati e ridurre la dimensione (collapse) della matrice degli strumenti, ossia creare uno strumento per ogni variabile e lag, anziché uno per ogni periodo di tempo. In piccoli campioni si evita la distorsione che sorge quando il numero di strumenti sale verso il numero di osservazioni. Il modo migliore di operare è quello di partire con un numero "sufficientemente" ampio di ritardi e ridurlo progressivamente fino a trovare il modello in cui i test AR(2) e di Hansen definiscono la consistenza delle stime ottenute.

[98] Lo stimatore two-step corregge la matrice di varianze e covarianze, rilasciando l'assunzione di indipendenza e di omoschedasticità del primo stadio. I risultati dello stimatore one step sono esposti nella Tabella C8 in appendice.

[99] Test di Sargan – H_0: non robustezza e validità degli strumenti; Test di Hansen - H_0: robustezza e non validità degli strumenti.

Tabella 2: Effetti di lungo periodo Modello AB: coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE6(irr1)

Variabile espressa in logaritmi	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Istruzione terzo livello	-3.210562	13.33982	-0.24	0.812	-31.13113 24.71000

Risultati dell'elaborazione con il software statistico STATA

Lo studio degli effetti di lungo periodo non dà una interpretazione statisticamente significativa.

I risultati appena esposti mostrano che la serie non è stazionaria in quanto rho è maggiore di 1. Vista la dimensione contenuta, siamo di fronte ad una stazionarietà del 1° ordine, che salva la significatività delle t-statistiche, ma non quelle del F test. Inoltre, un coefficiente di ritardo di questo tipo, indica che vi è un condizionamento del 100% da un periodo all'altro. Pertanto, se i ritardi fossero infinitesimali, ciò sarebbe istantaneo, cioè y_t=cost, quindi il resto del modello avrebbe poca rilevanza. Per quanto appena evidenziato, si ritiene necessario proseguire lo studio con l'applicazione del metodo Blundell e Bound.

Tabella 3: Stime panel data dinamico (two step system GMM): coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE6(irr1)[100]

Variabili espresse in logaritmi	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Economia sommersa con un ritardo	0.5480834	0.195584	2.80	0.011	0.1387214 0.9574455
Densità di popolazione	0.0411461	0.0362741	1.13	0.271	-0.0347765 0.1170687
Istruzione terzo livello	0.0593764	0.086291	0.69	0.500	-0.1212327 0.2399855
Tassazione	0.3545483	0.3175649	1.12	0.278	-0.3101228 1.019219
Imprenditorialità femminile	-0.1719214	0.1719775	-1.00	0.330	-0.5318745 0.1880316
Natalità delle imprese	0.3548451	0.1632761	2.17	0.043	0.0131042 0.6965859
Dipendenti pubblici	0.198571	0.0878865	2.26	0.036	0.0146226 0.3825195
Disoccupazione giovanile	0.1462476	0.0671363	2.18	0.042	0.0057297 0.2867655
Rapine	0.0046308	0.0294284	0.16	0.877	-0.0569635 0.0662252
Cons	-1.530406	0.9780895	-1.56	0.134	-3.577571 0.5167588
Nro strumenti<Nro gruppi	12<20				
F-test	F(8, 19) = 169.10; Prob > F = 0.000				
Test AR(1)	z = -3.07 Pr > z = 0.002				
Test AR(2)	z = -0.11 Pr > z = 0.912				
Test Sargan	χ^2 = 5.55 Prob > χ^2 = 0.062				
Test Hansen	χ^2 = 5.66 Prob > χ^2 = 0.059				

Risultati dell'elaborazione con il software statistico STATA

Nella Tabella 3 vengono presentati i risultati del modello di Blundell-Bond (BB) two step. Sia i coefficienti che gli standard error, sono più piccoli rispetto alle stime AB. Lo stimatore BB two-step conferma i risultati dello stimatore one-step[101]. I test di specificazione supportano totalmente il modello prescelto. Nello specifico i test di Hansen e Sargan accettano la validità degli strumenti, sebbene debolmente, e il test di autocorrelazione di secondo ordine degli errori indica che l'errore non è serialmente correlato. L'errore di primo ordine nei livelli non ha componenti auto regressive e si giustifica quindi la scelta degli strumenti. Risulta significativo e positivo il coefficiente della variabile dipendente ritardata, della natalità delle imprese,

[100] In questa applicazione vengono escluse le dummy che determinano un numero eccessivo di strumenti rispetto alle entità, violando la regola di Roodman (2009), in base alla quale il numero di strumenti deve essere inferiore al numero delle entità oggetto di analisi.

[101] Tabella C9 in appendice.

dei dipendenti pubblici e della disoccupazione giovanile. Un valore significativo di rho conferma la presenza di una persistenza del fenomeno nelle regioni, già riscontrata nel modello AB. Inoltre, un incremento dell'1% della natalità delle imprese, dei dipendenti pubblici e della disoccupazione giovanile determina, ceteris paribus, una variazione positiva rispettivamente dello 0,35%, 0,20% e 0,15% dell'Economia Sommersa nel breve periodo, al 5% di livello di significatività.

Tabella 4: Effetti di lungo periodo Modello BB: coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE6(irr1)

Variabili espresse in logaritmi	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Natalità delle imprese	0.7852004	0.2289582	3.43	0.003	0.3059855 1.264415
Dipendenti pubblici	0.4393975	0.2242862	1.96	0.065	-0.0300389 0.908834
Disoccupazione giovanile	0.3236164	0.0433667	7.46	0.000	0.2328489 0.4143839

Risultati dell'elaborazione con il software statistico STATA

Lo studio degli effetti di lungo periodo delle determinanti dell'Economia Sommersa fornisce una interpretazione statisticamente significativa dei coefficienti relativi alle variabili natalità delle imprese, dipendenti pubblici e disoccupazione giovanile. Un incremento dell'1% della natalità delle imprese e della disoccupazione giovanile determina, ceteris paribus, un aumento rispettivamente dello 0,78% e dello 0,32% dell'Economia Sommersa nel lungo periodo, all'1% di livello di significatività. Inoltre, una crescita dell'1% dei dipendenti pubblici indica, ceteris paribus, una variazione positiva dello 0,44% dell'Economia Sommersa nel lungo periodo, al 5% di livello di significatività. In tutti i casi si riscontra un effetto positivo più grande nel lungo periodo rispetto al breve periodo.

Con riferimento al modello BB two step, il coefficiente di velocità di aggiustamento (δ) è pari a 0,45 e il ritardo di medio periodo (θ) è pari a 2,22 anni, che corrisponde a ciò che comunemente è considerato medio periodo (Castaldo, A., Fiorini, A., Maggi, B., 2018).

IL CASO EUROPEO: MODELLO FE UE2

Per lo studio europeo si sottopone ad analisi con panel dinamico, il modello FEUE2 nel quale, ricordiamo, si considerano tutti i 28 Paesi dell'Unione europea.

Tabella 5: Stime panel data dinamico (two step difference GMM): coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FEUE2[102]

Variabili espresse in logaritmi	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Economia sommersa con un ritardo	1.016592	0.3535798	2.88	0.008	0.2911058 1.742077
Abitazione	0.022788	0.0460209	0.50	0.624	-0.071639 0.117215
Densità	-0.0642561	0.5314882	-0.12	0.905	-1.15478 1.026268
Disoccupazione giovanile	-0.0950835	0.0604427	-1.57	0.127	-0.2191017 0.0289347
Occupazione femminile	0.9075503	0.4517054	2.01	0.055	-0.0192726 1.834373
Industrializzazione	-1.027197	0.1896481	-5.42	0.000	-1.416322 -0.638071
Istruzione di terzo livello	-0.1312727	0.098859	-1.33	0.195	-0.3341147 0.0715693
Istruzione di secondo livello	-0.3141078	0.1698463	-1.85	0.075	-0.6626036 0.0343879

[102] In questa applicazione vengono escluse le dummy che determinano un numero eccessivo di strumenti rispetto alle entità, violando la regola di Roodman (2009), in base alla quale il numero di strumenti deve essere inferiore al numero delle entità oggetto di analisi.

Pil pro capite	0.5868174	0.2328624	2.52	0.018	0.1090231 1.064612
Tassazione indiretta	-0.2560213	0.155871	-1.64	0.112	-0.5758421 0.0637995
Nro strumenti<Nro gruppi	11<28				
F-test	F(9, 27) = 21.46 Prob > F = 0.000				
Test AR(1)	z = -1.90 Pr > z = 0.057				
Test AR(2)	z = -0.15 Pr > z = 0.880				
Test Sargan	χ^2 = 3.60 Prob > χ^2 = 0.058				
Test Hansen	χ^2 = 4.28 Prob > χ^2 = 0.039				

Risultati dell'elaborazione con il software statistico STATA

Si è ipotizzata la stretta esogeneità dei regressori e sono stati adottati come strumenti per la variabile dipendente ritardata[103], l'Economia Sommersa, con ritardo al tempo 2 ed al tempo 3 ((t-2) e (t-3))[104].

Nella Tabella 5 vengono presentati i risultati del modello di Arellano-Bond (AB) two step. Lo stimatore AB two-step, asintoticamente più efficiente dello stimatore one-step e robusto all'eteroschedasticità e all'autocorrelazione (Roodman, 2009)[105], conferma quanto emerso con lo stimatore AB one-step, sebbene qualche variabile perda di significatività. I test di specificazione supportano totalmente il modello prescelto[106]. Risultano significativi i coefficienti delle seguenti variabili: occupazione femminile, tasso di industrializzazione, istruzione di medio livello, prodotto interno lordo pro capite. Un valore significativo di rho, pari a 1,02, dimostra la presenza di una persistenza del fenomeno nei Paesi UE. Al crescere dell'1% dell'industrializzazione si ha, ceteris paribus, una variazione negativa dell'1,03% dell'Economia Sommersa nel breve periodo, all'1% di livello di significatività; inoltre un aumento dell'1% dell'istruzione di medio livello indica, ceteris paribus, un incremento dello 0,31% dell'Economia Sommersa nel breve periodo, al 5% di livello di significatività; infine una crescita dell'1% del prodotto interno lordo individua, ceteris paribus, una variazione negativa dello 0,59% dell'Economia Sommersa nel breve periodo, al 5% di livello di significatività. La variabile occupazione femminile fornisce un risultato contrastante con la letteratura economica. Le restanti variabili non danno una interpretazione statisticamente significativa nel breve periodo.

Gli effetti di lungo periodo sono analizzati solo per le variabili risultate statisticamente significative[107] nel breve periodo. Nel lungo periodo non si ottiene una interpretazione statisticamente significativa per alcuna variabile.

Come già riscontrato nel caso studio italiano, i risultati appena esposti mettono in evidenza come la serie non sia stazionaria, in quanto rho è maggiore di 1[108]. Pertanto, lo studio prosegue con l'applicazione del metodo Blundell e Bound che conduce a stime più precise.

[103] Si veda la nota 96.

[104] Si veda la nota 97.

[105] Lo stimatore two-step corregge la matrice di varianze e covarianze, rilasciando l'assunzione di indipendenza e di omoschedasticità del primo stadio. I risultati dello stimatore one step sono esposti nella Tabella C10 in appendice.

[106] Sono stati infatti eseguiti i test necessari per poter determinare la consistenza delle stime ottenute. Il test AR(2), necessario per accettare la regressione nel suo insieme, verifica l'ipotesi nulla che non vi sia autocorrelazione di secondo ordine tra i residui di regressione. Il test di Hansen e Sargan, verificano l'ipotesi che le restrizioni in sovra-identificazione siano tutte valide. L'ipotesi testata è che le variabili strumentali siano specificate correttamente, e quindi siano strumenti accettabili ed efficientemente utilizzabili nel modello. Nello specifico i test di Hansen e Sargan accettano la validità degli strumenti, e il test di autocorrelazione di secondo ordine degli errori indica che l'errore di primo e secondo ordine non è serialmente correlato. L'errore di primo ordine nei livelli non ha componenti auto regressive e si giustifica quindi la scelta degli strumenti.

[107] Mantenere anche le variabili non statisticamente significative aumenta la variabilità della stima puntuale, mantenendo però la correttezza e la consistenza.

[108] Vista la dimensione contenuta, siamo di fronte ad una stazionarietà del 1° ordine, che salva la significatività delle t-statistiche, ma non quelle del F test. Inoltre, un coefficiente di ritardo di questo tipo ci informa che vi è un condizionamento del 100% da un periodo all'altro e se i ritardi fossero infinitesimali ciò sarebbe istantaneo, cioè y_t=cost, pertanto il resto del modello avrebbe poca rilevanza.

Tabella 6: Stime panel data dinamico (two step system GMM): coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FEUE2[109]

Variabili espresse in logaritmi	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Economia sommersa con un ritardo	0.8199385	0.0799115	10.26	0.000	0.6559735 0.9839034
Abitazione	0.041057	0.022295	1.84	0.077	-0.0046885 0.0868025
Densità	-0.0189451	0.0132212	-1.43	0.163	-0.0460728 0.0081826
Disoccupazione giovanile	0.0358082	0.0129539	2.76	0.010	0.0092289 0.0623874
Occupazione femminile	-0.3013774	0.1381414	-2.18	0.038	-0.5848202 -0.0179346
Industrializzazione	0.0655638	0.0505127	1.30	0.205	-0.0380797 0.1692074
Istruzione di terzo livello	0.0417071	0.0279236	1.49	0.147	-0.0155873 0.0990015
Istruzione di secondo livello	-0.0493212	0.028699	-1.72	0.097	-0.1082067 0.0095642
Pil pro capite	-0.0416479	0.0314502	-1.32	0.197	-0.1061783 0.0228826
Tassazione indiretta	0.069914	0.0339596	2.06	0.049	0.0002347 0.1395933
Nro strumenti<Nro gruppi	13<28				
F-test	F(9, 27) = 761.79 Prob > F = 0.000				
Test AR(1)	z = -3.95 Pr > z = 0.000				
Test AR(2)	z = -1.84 Pr > z = 0.066				
Test Sargan	χ^2 = 6.29 Prob > χ^2 = 0.043				
Test Hansen	χ^2 = 5.07 Prob > χ^2 = 0.079				

Risultati dell'elaborazione con il software statistico STATA

Nella Tabella 6 sono esposti i risultati del modello di Blundell-Bond (BB) two step[110]. I test di specificazione supportano totalmente il modello prescelto. Nello specifico i test di Hansen e Sargan accettano la validità degli strumenti, sebbene debolmente, e il test di autocorrelazione di secondo ordine degli errori indica che l'errore non è serialmente correlato. L'errore di primo ordine nei livelli non ha componenti auto regressive e si giustifica quindi la scelta degli strumenti.

Una analisi più approfondita mostra che il coefficiente della variabile dipendente ritardata, della condizione abitative, della disoccupazione giovanile e della tassazione indiretta è significativo e positivo. Inoltre, è significativo e negativo il coefficiente della variabile che esprime l'occupazione femminile e il livello di istruzione medio. Un valore significativo di rho conferma la presenza di una persistenza del fenomeno nei Paesi UE, già riscontrata nel modello AB. Al crescere dell'1% delle condizioni abitative e del livello della tassazione indiretta si ha, ceteris paribus, un incremento rispettivamente dello 0,04% e dello 0,07% dell'Economia Sommersa nel breve periodo, al 5% di livello di significatività. Inoltre, un aumento dell'1% della disoccupazione giovanile indica, ceteris paribus, una variazione positiva dello 0,03% dell'Economia Sommersa nel breve periodo, all'1% di livello di significatività. Infine, al crescere dell'1% del livello di istruzione medio si ha una variazione negativa dello 0,05% dell'Economia Sommersa nel breve periodo, al 5% di livello di significatività. La variabile occupazione femminile fornisce un risultato contrastante con la letteratura economica. Le restanti variabili non danno una interpretazione statisticamente significativa di breve periodo.

[109] In questa applicazione vengono escluse le dummy che determinavano un numero eccessivo di strumenti rispetto alle entità, violando la regola di Roodman (2009), in base alla quale il numero di strumenti deve essere inferiore al numero delle entità oggetto di analisi.

[110] I risultati del modello one step sono esposti in appendice Tabella C11.

Tabella 7: Effetti di lungo periodo Modello BB: coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FEUE2

Variabili espresse in logaritmi	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Abitazione	0.2280165	0.093477	2.44	0.022	0.0362175 0.4198154
Disoccupazione giovanile	0.1988663	0.0623382	3.19	0.004	0.0709587 0.3267738
Occupazione femminile	-1.673747	0.3779908	-4.43	0.000	-2.44932 -0.8981737
Istruzione di secondo livello	-0.2739133	0.0729524	-3.75	0.001	-0.4235992 -0.1242273
Tassazione indiretta	0.3882784	0.1891631	2.05	0.050	0.0001478 0.776409

Risultati dell'elaborazione con il software statistico STATA

Lo studio degli effetti di lungo periodo delle determinanti dell'Economia Sommersa dà una interpretazione statisticamente significativa di tutte le variabili risultate significative nel breve periodo. Nello specifico un aumento dell'1% della situazione abitativa e del livello della tassazione indiretta indica, ceteris paribus, una variazione positiva rispettivamente dello 0,23% e dello 0,39% dell'Economia Sommersa nel lungo periodo, al 5% di livello di significatività. Inoltre, al crescere dell'1% della disoccupazione giovanile si ha, ceteris paribus, un aumento dello 0,20% dell'Economia Sommersa nel lungo periodo, all'1% di livello di significatività. Infine, un incremento dell'1% del livello di istruzione medio determina un decremento dello 0,27% dell'Economia Sommersa nel breve periodo, all'1% di livello di significatività. La variabile occupazione femminile mostra un risultato contrastante con la letteratura economica. In tutti i casi si riscontra un effetto positivo più grande nel lungo periodo rispetto al breve periodo.

Con riferimento al modello BB two step, il coefficiente di velocità di aggiustamento (δ) è pari a 0,18 e il ritardo di medio periodo (θ) è pari a 5,55 anni, che corrisponde a ciò che viene comunemente considerato medio periodo (Castaldo, A., Fiorini, A., Maggi, B., 2018).

5.3 Conclusioni

Una strategia utilizzata per rispondere al problema di possibile presenza di endogeneità, che viene applicata soprattutto nello studio dei fenomeni sociali ed economici (Bellemare, Masaki, Pepinsky, 2017), è quella di ritardare le variabili esplicative. Tale strategia è conosciuta come "identificazione del ritardo" ("lag identification"). Tale metodo viene impiegato molto frequentemente in quanto è necessario avere semplicemente a disposizione il dataset. In questo studio è stata adoperata l'analisi panel dinamica.

Gli stimatori più utilizzati e conosciuti sono lo stimatore di Arellano-Bond (Difference GMM estimator) (Arellano e Bond, 1991) e Blundell-Bond (System GMM estimator) (Blundell e Bond, 1998; Roodman,2009). Tali metodi di stima sono basati sul metodo dei momenti generalizzato GMM. L'idea è quella di rimuovere l'effetto individuale trasformando l'equazione in un'equazione in differenze prime e di usare appropriati ritardi delle variabili endogene come strumenti. In particolare, lo stimatore di Arellano e Bond (1991), come primo passo, trasforma tutti i regressori tramite differenziazione e usa il metodo generalizzato dei momenti. Per questo motivo viene chiamato Difference GMM. Lo stimatore di Arellano-Bover (1995) e Blundell-Bond (1998) costruisce invece un sistema di due equazioni, l'equazione originale e quella trasformata, per questa ragione viene denominato System GMM.

Per poter capire come un intervento pubblico possa influire sul fenomeno dell'Economia Sommersa, sono stati introdotti concetti quali i parametri di lungo periodo, la velocità di aggiustamento e il ritardo medio, che aiutano a studiare gli effetti delle componenti sull'Economia Sommersa nel tempo. In questo modo è possibile essere più espliciti sulle variazioni dell'Economia Sommersa nel breve, medio e lungo periodo a

seguito dell'aumento percentuale delle sue determinanti. Per fare questo è stato introdotto il modello con ritardo distribuito attraverso il metodo di Koyck e il modello di aggiustamento parziale.

Tali modelli vengono applicati sia nel contesto italiano che nel contesto europeo. Per enfatizzare la dinamicità del modello, tutte le variabili vengono trasformate in logaritmo. Questo passaggio consente di introdurre il concetto di elasticità, molto utilizzato in ambito economico.

Con riferimento al caso italiano l'istruzione di terzo livello ha avvalorato quanto emerso nei precedenti modelli nella sola analisi di breve periodo; la natalità d'impresa, l'indice di regolamentazione e la disoccupazione giovanile confermano quanto già definito nei precedenti modelli, sia nel breve che nel lungo periodo, con un effetto positivo più grande nel lungo periodo. Con riferimento al modello BB two step, il coefficiente di velocità di aggiustamento (δ) è pari a 0,45 e il ritardo di medio periodo (θ) è pari a 2,22 anni, che corrisponde a ciò che viene comunemente considerato medio periodo.

Nello studio europeo l'istruzione di medio livello, il tasso di industrializzazione e il prodotto interno lordo hanno convalidato quanto emerso nei precedenti modelli nella sola analisi di breve periodo; mentre la condizione abitativa, l'istruzione di medio livello, la tassazione indiretta e la disoccupazione giovanile confermano quanto già definito nei precedenti modelli, sia nel breve che nel lungo periodo, con un effetto positivo più grande nel lungo periodo. Con riferimento al modello BB two step, il coefficiente di velocità di aggiustamento (δ) è pari a 0,18 e il ritardo di medio periodo (θ) è pari a 5,55 anni, che corrisponde a ciò che viene comunemente considerato medio periodo.

APPENDICE CAPITOLO 5

Tabella C1: Studio correlazione dei residui del Modello FE1(irr): correlazione, autocorrelazione a coppie[111], Pesaran test, Wooldgridge test

Correlazione

	u0	u1	u2	u3	u4
u0	1				
u1	0,4718	1			
u2	0,1938	0,4279	1		
u3	0,0074	0,1202	0,4591	1	
u4	-0,146	0,0006	0,141	0,4876	1

Autocorrelazione a coppie

	u0	u1	u2	u3	u4
u0	1				
u1	0,5228*	1			
u2	0,2324*	0,4942*	1		
u3	0,0709	0,1752*	0,4935*	1	
u4	-0,146	0,0006	0,141	0,4876*	1

Pesaran's test di correlazione seriale e cross section
H0: no correlazione seriale z=-2.253, Pr=0.0242

Wooldridge test per l'autocorrelazione nel panel data
H0: no autocorrelazione di primo ordine
F(1,19)=16.774 Prob>0.0006

Fonte: Risultati dell'elaborazione con il software statistico STATA

Tabella C2: Studio correlazione dei residui Modello FE3(irr1): correlazione, autocorrelazione a coppie, Pesaran test, Wooldgridge test

Correlazione

	u0	u1	u2	u3	u4
u0	1				
u1	0,5169	1			
u2	0,0555	0,4530	1		
u3	-0,0958	0,0255	0,4513	1	
u4	-0,0931	-0,0523	0,0693	0,5093	1

Autocorrelazione a coppie

	u0	u1	u2	u3	u4
u0	1				
u1	0,5546*	1			
u2	0,1464*	0,5190*	1		
u3	-0,0037	0,1116	0,5028*	1	
u4	-0,0931	-0,0523	0,0693	0,5093	1

Pesaran's test di correlazione seriale e cross section
H0: no correlazione seriale z=-2.239, Pr=0.0251

Wooldridge test per l'autocorrelazione nel panel data
H0: no autocorrelazione di primo ordine
F(1,19)=69.207 Prob>0.0000

Fonte: Risultati dell'elaborazione con il software statistico STATA

[111] Con * viene indicata la significatività al 5%.

Tabella C3: Studio correlazione dei residui Modello FE6(irr1): correlazione, autocorrelazione a coppie, Pesaran test, Wooldridge test

Correlazione

	u0	u1	u2	u3	u4
u0	1				
u1	0,6381	1			
u2	0,289	0,5977	1		
u3	0,0711	0,2277	0,5794	1	
u4	0,0075	0,0983	0,2685	0,6336	1

Autocorrelazione a coppie

	u0	u1	u2	u3	u4
u0	1				
u1	0,6697*	1			
u2	0,3460*	0,6525*	1		
u3	0,1402*	0,2906*	0,6240*	1	
u4	0,0075	0,0983	0,2685*	0,6336*	1

Pesaran's test di correlazione seriale e cross section
H0: no correlazione seriale z=-2.621, Pr=0.0088

Wooldridge test per l'autocorrelazione nel panel data
H0: no autocorrelazione di primo ordine
F(1,19)=94.708 Prob>0.0000

Risultati dell'elaborazione con il software statistico STATA

Tabella C4: Studio correlazione dei residui Modello FEUE2: correlazione, autocorrelazione a coppie, Pesaran test, Wooldridge test

Correlazione

	u0	u1	u2	u3	u4
u0	1				
u1	0,4621	1			
u2	0,0774	0,4335	1		
u3	-0,1338	0,0618	0,4603	1	
u4	-0,2577	-0,1500	0,0985	0,5031	1

Autocorrelazione a coppie

	u0	u1	u2	u3	u4
u0	1				
u1	0,5201*	1			
u2	0,1395*	0,5226*	1		
u3	0,1020	0,1209	0,5135*	1	
u4	0,0075	-0,1500	0,0985	0,5031*	1

Pesaran's test di correlazione seriale e cross section
H0: no correlazione seriale z=-2.319, Pr=0.0204

Wooldridge test per l'autocorrelazione nel panel data
H0: no autocorrelazione di primo ordine
F(1,27)=42.948 Prob>0.0000

Risultati dell'elaborazione con il software statistico STATA

Tabella C5: Studio correlazione dei residui Modello FEOLDUE3: correlazione, autocorrelazione a coppie, Pesaran test, Wooldridge test

Correlazione

	u0	u1	u2	u3	u4
u0	1				
u1	0,3578	1			
u2	0,0578	0,3612	1		
u3	-0,0861	0,0184	0,4253	1	
u4	-0,2885	-0,1904	0,1002	0,4529	1

Autocorrelazione a coppie

	u0	u1	u2	u3	u4
u0	1				
u1	0,4330*	1			
u2	0,1343	0,4379*	1		
u3	-0,0805	0,0952	0,4448*	1	
u4	-0,2885	-0,1904	0,1002	0,4529*	1

Pesaran's test di correlazione seriale e cross section
H0: no correlazione seriale z=-2.291, Pr=0.0220

Wooldridge test per l'autocorrelazione nel panel data
H0: no autocorrelazione di primo ordine
F(1,14)=12.819 Prob>0.0030

Risultati dell'elaborazione con il software statistico STATA

Tabella C6: Studio correlazione dei residui Modello FENEWUE3: correlazione, autocorrelazione a coppie, Pesaran test, Wooldridge test

Correlazione

	u0	u1	u2	u3	u4
u0	1				
u1	0,5001	1			
u2	0,0126	0,4546	1		
u3	-0,2389	0,0729	0,4472	1	
u4	-0,3379	-0,1957	0,0439	0,4516	1

Autocorrelazione a coppie

	u0	u1	u2	u3	u4
u0	1				
u1	0,4828*	1			
u2	0,0029	0,5054*	1		
u3	-0,2440	0,0768	0,5026*	1	
u4	-0,2885	-0,1957	0,0439	0,4516*	1

Pesaran's test di correlazione seriale e cross section
H0: no correlazione seriale z=-2.129, Pr=0.0333

Wooldridge test per l'autocorrelazione nel panel data
H0: no autocorrelazione di primo ordine
F(1,12)=41.162 Prob>0.0000

Risultati dell'elaborazione con il software statistico STATA

Tabella C7: Studio correlazione dei residui Modello FEAE2: correlazione, autocorrelazione a coppie, Pesaran test, Wooldridge test

Correlazione

	u0	u1	u2	u3	u4
u0	1				
u1	0,4000	1			
u2	0,0251	0,4229	1		
u3	-0,2347	0,0080	0,4598	1	
u4	-0,2900	-0,2509	0,0163	0,4651	1

Autocorrelazione a coppie

	u0	u1	u2	u3	u4
u0	1				
u1	0,4597*	1			
u2	0,0295	0,4775*	1		
u3	0,2005*	0,0280	0,4838*	1	
u4	0,2900*	0,2509*	0,0163	0,4651*	1

Pesaran's test di correlazione seriale e cross section
H0: no correlazione seriale z=-2.225, Pr=0.0261

Wooldridge test per l'autocorrelazione nel panel data
H0: no autocorrelazione di primo ordine
F(1,18)=28.409 Prob>0.0000

Risultati dell'elaborazione con il software statistico STATA

Tabella C8: Stime panel data dinamico (one step difference GMM): coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE6(irr1)

Variabili espresse in logaritmi	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Economia sommersa con un ritardo	0.9296567	0.2114946	4.40	0.000	0.4869935 1.37232
Densità di popolazione	0.3116582	0.4159575	0.75	0.463	-0.5589508 1.182267
Istruzione terzo livello	0.224252	0.0679236	3.30	0.004	0.0820864 0.3664177
Tassazione	-0.1787912	0.185003	-0.97	0.346	-0.566007 0.2084246
Imprenditorialità femminile	-0.3177291	0.5011181	-0.63	0.534	-1.366581 0.7311232
Natalità delle imprese	0.010633	0.0574791	0.18	0.855	-0.1096721 0.1309382
Dipendenti pubblici	-0.1150087	0.1224404	-0.94	0.359	-0.3712795 0.1412621
Disoccupazione giovanile	-0.0123048	0.0204448	-0.60	0.554	-0.0550963 0.0304866
Rapine	-0.048173	0.032365	-1.49	0.153	-0.1159137 0.0195676
Nro strumenti< Nro gruppi	10<20				
F-test	F(8, 19) = 32.78 Prob > F = 0.000				
Test AR(1)	z = -3.08 Pr > z = 0.002				
Test AR(2)	z = -0.92 Pr > z = 0.358				
Test Sargan	χ^2 = 2.80 Prob > χ^2 = 0.094				
Test Hansen	χ^2 = 2.22 Prob > χ^2 = 0.137				

Risultati dell'elaborazione con il software statistico STATA

Tabella C9: Stime panel data dinamico (one step system GMM): coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FE6(irr1)

Variabili espresse in logaritmi	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Economia sommersa con un ritardo	0.6285368	0.1630428	3.86	0.001	0.2872843 0.9697893
Densità di popolazione	0.0149492	0.0315715	0.47	0.641	-0.0511307 0.0810291
Istruzione terzo livello	0.0339115	0.0849183	0.40	0.694	-0.1438246 0.2116476
Tassazione	0.3163983	0.2792254	1.13	0.271	-0.2680273 0.9008238
Imprenditorialità femminile	-0.1077669	0.1287303	-0.84	0.413	-0.3772026 0.1616688
Natalità delle imprese	0.266285	0.1216755	2.19	0.041	0.0116152 0.5209547
Dipendenti pubblici	0.1512649	0.0642863	2.35	0.030	0.0167121 0.2858177
Disoccupazione giovanile	0.1268137	0.0482359	2.63	0.017	0.0258549 0.2277725
Rapine	0.0086585	0.0279031	0.31	0.760	-0.0497433 0.0670603
Cons	-1.243912	0.8118284	-1.53	0.142	-2.943089 0.455264
Nro strumenti< Nro gruppi	12<20				
F-test	F(8, 19) = 260.97; Prob > F = 0.000				
Test AR(1)	z = -3.00 Pr > z = 0.003				
Test AR(2)	z = -0.44 Pr > z = 0.663				
Test Sargan	χ^2 = 5.55 Prob > χ^2 = 0.062				
Test Hansen	χ^2 = 5.66 Prob > χ^2 = 0.059				

Risultati dell'elaborazione con il software statistico STATA

Tabella C10: Stime panel data dinamico (one step difference GMM): coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FEUE2

Variabili espresse in logaritmi	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Economia sommersa con un ritardo	1.157948	0.2956249	3.92	0.001	0.5513763 1.764521
Abitazione	0.0231482	0.0396645	0.58	0.564	-0.0582365 0.104533
Densità	-0.088346	0.4753518	-0.19	0.854	-1.063687 0.8869952
Disoccupazione giovanile	-0.1109435	0.058828	-1.89	0.070	-0.2316485 0.0097616
Occupazione femminile	0.9046868	0.4386116	2.06	0.049	0.0047302 1.804644
Industrializzazione	-1.050929	0.1784385	-5.89	0.000	-1.417055 -0.6848036
Istruzione di terzo livello	-0.1817738	0.1005387	-1.81	0.082	-0.3880623 0.0245147
Istruzione di secondo livello	-0.3193279	0.1590344	-2.01	0.055	-0.6456396 0.0069838
Pil pro capite	0.6005307	0.1973181	3.04	0.005	0.1956673 1.005394
Tassazione indiretta	-0.2600175	0.1455463	-1.79	0.085	-0.558654 0.0386189
Nro strumenti<Nro gruppi	11<28				
F-test	F(9, 27) = 23.13 Prob > F = 0.000				
Test AR(1)	z = -2.20 Pr > z = 0.028				
Test AR(2)	z = -0.39 Pr > z = 0.698				
Test Sargan	χ^2 = 3.60 Prob > χ^2 = 0.058				
Test Hansen	χ^2 = 4.28 Prob > χ^2 = 0.039				

Risultati dell'elaborazione con il software statistico STATA

Tabella C11: Stime panel data dinamico (one step system GMM): coefficienti di regressione, errore standard, t, significatività e intervalli di confidenza (95%) delle variabili esplicative del modello FEUE2

Variabili espresse in logaritmi	Coefficiente di regressione	Errore standard	t	p-value	Intervallo di confidenza
Economia sommersa con un ritardo	0.8646121	0.0725658	11.91	0.000	0.7157195 1.013505
Abitazione	0.0341163	0.0203386	1.68	0.105	-0.007615 0.0758476
Densità	-0.133445	0.0111912	-1.19	0.243	-0.036307 0.0096179
Disoccupazione giovanile	0.0278754	0.0111336	2.50	0.019	0.0050313 0.0507196
Occupazione femminile	-0.2110311	0.1339592	-1.58	0.127	-0.4858928 0.068305
Industrializzazione	0.0196385	0.0492554	0.40	0.693	-0.0814253 0.1207023
Istruzione di terzo livello	0.0291972	0.0255829	1.14	0.264	-0.232946 0.0816891
Istruzione di secondo livello	-0.050173	0.02774	-1.81	0.082	-0.1070908 0.007449
Pil pro capite	-0.0293345	0.027449	-1.07	0.295	-0.0856469 0.0269779
Tassazione indiretta	0.0603265	0.0293133	2.06	0.049	0.0001807 0.1204724
cons	1.108026	0.7558459	1.47	0.154	-0.4428417 2.658894
Nro strumenti<Nro gruppi	11<28				
F-test	F(9, 27) = 972.70 Prob > F = 0.000				
Test AR(1)	z = -4.12 Pr > z = 0.000				
Test AR(2)	z = -1.77 Pr > z = 0.077				
Test Sargan	χ^2 = 6.29 Prob > χ^2 = 0.043				
Test Hansen	χ^2 = 5.07 Prob > χ^2 = 0.079				

Risultati dell'elaborazione con il software statistico STATA

BIBLIOGRAFIA CAPITOLO 5

Bellemare, M.F. Masaki, T. e Pepinsky, T.B. (2017). Lagged Explanatory Variables and the Estimation of Casual Effect. The Journal of Politics (volume 79, n. 3)

Arellano, M. e Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. Review of Economic Studies.

Arellano, M. e Bover, O. (1995). Another look at instrumental variable estimation of error components model. Journal of Econometrics 68 (1995) 29-51

Baltagi, B.H. (2005). Econometric Analysis of Panel data. John Wiley and Son. Third Edition

Blundell, R. Bond, S. (1998). GMM estimation with persistent panel data: an application to production function. Eight International Conference on Panel data – Goteborg University – June 11-12, 1998

Castaldo, A., Fiorini, A., Maggi, B. (2018). Measuring (in time of crisis) the impact of broadban connections on economic growth: an OECD panel analysis. Applied Economics, 2018. Vol. 50 N.8

Kiviet, J.F. (1995). On bias, inconsistency, and efficiency of various estimators in dynamic panel data models. Journal of Econometric. Volume 68.

Palomba, G. (2008). Panel data. Dispense

Pesaran, M.H. (2004). General diagnostic tests for cross section dependence in panel data. IZA Discussion Papers

Roodman, D. (2009). How to do xtabond2: An introduction to difference and system GMM in STATA.

Sekhon, J. S. (2009). Opiates for the Matches: Matching Methods for Casual Inferences. Annual Review of Political Science.

Wooldridge, J. (2002). Econometric Analysis of Cross Section and Panel Data. MIT Press

CONCLUSIONI

Il presente studio di ricerca intende fornire un contributo alla comunità scientifica, economica e statistica, sul tema dell'Economia Sommersa, e supportare l'attività di Governo che già opera sulla materia.

L'Economia Sommersa è un fenomeno estremamente complesso che influisce fortemente sullo sviluppo economico, distorcendo e togliendo efficienza al normale funzionamento della concorrenza e del mercato, poiché assorbe risorse altrimenti destinate al bilancio pubblico. Provoca inoltre iniquità in quanto determina una riduzione del gettito che può pregiudicare la qualità e la quantità dei servizi pubblici offerti a tutti i cittadini, compresi coloro che contribuiscono attraverso il regolare pagamento delle imposte. Essa si esplicita attraverso diverse forme, coglierla non è pertanto affatto semplice e agevole poiché per definizione è non osservabile. Nel corso del tempo l'attenzione degli economisti si è rivolta alla ricerca di una sua idonea definizione in grado di dare un'esatta misurazione della stessa. Più in generale, l'Economia Sommersa riguarda quell'insieme di attività produttive la cui caratteristica principale è quella di sfuggire all'osservazione, alla regolamentazione e alla rilevazione, sia che comporti transazioni monetarie (produzione e distribuzione), sia nel caso di transazioni non monetarie (autoproduzione, scambio e baratto). Sono quindi sommerse tanto le attività produttive legali, ma svolte in modo irregolare, quanto le attività illegali, per le quali si verifica una violazione della legge. La definizione più comune, recepita anche dall'ISTAT, rappresenta il fenomeno dell'Economia Sommersa come l'insieme di tutte le attività economiche non registrate, che sfuggono ad ogni rilevazione statistica e ai controlli fiscali.

La ricerca ha approfondito il concetto di Economia Sommersa, ha esaminato nel dettaglio le diverse metodologie di stima e ha fornito una analisi della letteratura economica nazionale e internazionale. La trattazione ha successivamente illustrato la metodologia applicata: i modelli con dati panel. In particolare, sono stati definiti i dati panel e sono stati mostrati esempi esplicativi, in campo internazionale e italiano, sull'uso degli stessi. Ci si è soffermati inoltre sui vantaggi e sugli svantaggi nell'utilizzo di una tale struttura di dati. È stata infine data una panoramica dei modelli con dati panel attualmente esistenti in letteratura.

Sono stati svolti due studi specifici del fenomeno, in ambito italiano ed europeo.

L'approfondimento svolto in ambito italiano, in accordo con l'ipotesi che il lavoro irregolare è "il principale fattore produttivo su cui si basa il funzionamento dell'economia sommersa" (Lucifora, 2003), ha identificato la variabile in esame con il tasso di irregolarità del lavoro. I dati sono di fonte ISTAT e sono stati costruiti sia come rapporto percentuale tra unità di lavoro non regolare e unità di lavoro totali (irr), attualmente fermi al 2012, che in termini di occupati irregolari (irr1), in quanto più aggiornati. Quest'ultima variabile, quale proxy dell'Economia Sommersa, non viene usualmente utilizzata negli studi econometrici, è sembrato pertanto meritevole di approfondimento non solo in quanto elemento di novità nell'ambito della letteratura dedicata all'argomento, ma anche in considerazione di un facile reperimento di dati aggiornati.

L'analisi è stata applicata su un campione di dati costituito da un panel bilanciato relativo alle 20 regioni d'Italia, con 12 osservazioni annuali comprese tra il 2001 e il 2012, sia per la variabile oggetto di studio nell'accezione irr che nell'accezione irr1. Per entrambe le variabili dipendenti si è analizzato il contributo fornito da 16 variabili esplicative, volte a spiegare la struttura sociodemografica, economica e a definire il fenomeno della criminalità. Avendo a disposizione per la variabile dipendente nell'accezione irr1 e per le variabili esplicative dati annuali per il triennio 2013-2015, si è ritenuto utile definire un modello che considerasse anche questa informazione.

Lo studio effettuato, oltre a confermare alcune relazioni già esistenti, ha fatto emergere diversi risultati importanti. Viene, infatti, dimostrato che un livello di tassazione troppo elevato e la presenza di criminalità, aumentano il livello di Economia Sommersa, mentre la presenza di un elevato volume di industrializzazione lo affievolisce. Avere però un livello di istruzione troppo elevato non sempre è un fenomeno di contrasto all'Economia Sommersa, poiché in mercati di lavoro "saturi" una elevata istruzione garantisce sicuramente delle migliori opportunità di lavoro ma non sempre maggiori possibilità. L'intensità della regolamentazione non sempre aumenta l'Economia Sommersa. Ciò dipende dalla modalità di costruzione dell'indicatore, dal campione di riferimento utilizzato e dalla tecnica di stima applicata. L'interpretazione economica della nuova relazione trovata è perfettamente intuibile, considerando la specifica scelta dell'indicatore. È, infatti, agevole ritenere che nelle zone con una maggiore presenza di dipendenti pubblici il sommerso sia meno radicato, e ciò a dimostrazione della positiva opera dei pubblici dipendenti di tutte le istituzioni centrali e periferiche. Infine, la relazione tra l'Economia Sommersa e la densità di popolazione mostra un segno negativo, poiché laddove la maggior densità è legata a una necessità lavorativa, tale variabile può essere correlata negativamente all'Economia Sommersa.

Per quanto riguarda l'approfondimento in ambito europeo, la variabile in esame è in questo contesto identificata con l'Economia Sommersa, così come la definisce Friedrich Schneider[112] (Medina e Schneider, 2017). La metodologia utilizzata per ottenere tali dati è l'approccio MIMIC (Multiple Indicators, Multiple Causes). I risultati emersi dall'analisi descrittiva e la letteratura economica esaminata hanno consigliato di affiancare l'analisi per tutti i 28 Paesi membri dell'UE, con l'analisi per i 15 Paesi membri appartenenti all'Unione europea dalla sua costituzione al 1995 (OLD UE) e per i 13 Paesi che hanno aderito dal 2004 (NEW UE).

Il database è costituito da un panel bilanciato relativo ai 28 Paesi UE, composto dalla variabile dipendente, ES, che rappresenta l'Economia Sommersa, disponibile per l'arco temporale 1991-2015. Le variabili esplicative a disposizione sono 12, volte a spiegare la struttura socio-demografica, economica e la qualità della vita, tutte fruibili per orizzonti temporali diversi. Pertanto, come periodo di riferimento è stato considerato l'arco temporale 2005-2015, periodo utilizzabile per tutte le variabili considerate.

I risultati hanno confermato che, affiancare l'analisi con tutti e 28 i Paesi europei (UE), con lo studio separato tra i Paesi UE, che hanno aderito all'Unione europea fino al 1995 (OLDUE), e quelli entrati dopo il 2004 (NEWUE), è giustificato. È stata condotta anche una analisi al livello di Paesi appartenenti all'area dell'euro (AE), in quanto interessante a livello istituzionale.

I modelli esaminati hanno confermato quanto già sostenuto nel caso italiano per il PIL, la tassazione, l'industrializzazione, l'occupazione femminile, la disoccupazione giovanile, l'intensità di regolamentazione, il livello di istruzione medio. Solo per il caso europeo è stato considerato il concetto di qualità della vita, che è risultato espresso in modo significativo e coerente nel modello con tutti e 28 i Paesi europei, attraverso l'esplicativa che esprime il tipo di abitazione. Il suo coefficiente mantiene il segno atteso (+), infatti, è logico ritenere che vi sia una correlazione positiva tra la percentuale della popolazione che abita in situazioni disagiate e l'Economia Sommersa. Pertanto, in situazioni in cui la qualità della vita è carente, l'Economia Sommersa aumenta. La relazione tra il livello di istruzione elevato e l'Economia Sommersa, nei modelli FE UE, FE OLDUE, FE AE, ha un segno del coefficiente negativo, in quanto secondo la letteratura economica l'istruzione ha un ruolo di contrasto rispetto al fenomeno in esame. Viceversa, per il modello FE NEWUE la variabile esplicativa ha un segno del coefficiente positivo. Infatti, il livello di istruzione troppo elevato non

[112]Professore di economia presso l'università Johannes Kepler di Linz in Austria e dal 2006 ricercatore presso l'istituto tedesco di ricerche economiche.

sempre è un fenomeno di contrasto dell'Economia Sommersa, poiché in mercati di lavoro "saturi" una elevata istruzione garantisce sicuramente delle migliori opportunità di lavoro, ma non sempre maggiori possibilità. Infine, la relazione tra l'Economia Sommersa e la densità di popolazione mostra un segno positivo, ossia tanto più un Paese è densamente popolato quanto più l'Economia Sommersa aumenta. Tale relazione appare logica, sebbene in contrasto con quanto sostenuto nello studio del caso italiano. Ciò perché la realtà italiana, pur incardinandosi nell'ambito europeo, si comporta in modo diverso.

Inoltre, per poter capire come un intervento pubblico possa influire sul fenomeno dell'Economia Sommersa, sono stati introdotti concetti quali i parametri di lungo periodo, la velocità di aggiustamento e il ritardo medio, che aiutano a studiare gli effetti delle componenti sull'Economia Sommersa nel tempo. In questo modo è possibile essere più espliciti sulle variazioni dell'Economia Sommersa nel breve, medio e lungo periodo, a seguito dell'aumento percentuale delle sue determinanti. Per fare questo, è stato introdotto il modello con ritardo distribuito, implementato con il metodo di Koyck e il modello di aggiustamento parziale.

Tali modelli sono stati applicati sia nel contesto italiano che nel contesto europeo. Per enfatizzare la dinamicità del modello, tutte le variabili vengono trasformate in logaritmo. Questo passaggio consente di introdurre il concetto di elasticità, molto utilizzato in ambito economico.

Con riferimento al caso italiano, il comportamento della variabile istruzione di terzo livello ha ribadito le risultanze dei precedenti modelli nell'analisi di breve periodo; il comportamento delle variabili natalità d'impresa, indice di regolamentazione e disoccupazione giovanile, concorda con quanto già emerso nei precedenti modelli, sia nel breve che nel lungo periodo, con un effetto positivo più grande nel lungo periodo. Il ritardo di medio periodo (θ) è risultato pari a 2,22 anni.

Nello studio europeo l'istruzione di medio livello, il tasso di industrializzazione e il prodotto interno lordo hanno avvalorato le risultanze dei precedenti modelli nell'analisi di breve periodo; la condizione abitativa, l'istruzione di medio livello, la tassazione indiretta e la disoccupazione giovanile convalidano quanto già definito nei precedenti modelli, sia nel breve che nel lungo periodo, con un effetto positivo più grande nel lungo periodo. Il ritardo di medio periodo (θ) è risultato pari a 5,55 anni.

L'approccio modellistico utilizzato nel presente studio ha permesso di descrivere l'Economia Sommersa attraverso le sue cause, non limitandosi solamente all'analisi degli aspetti puramente fiscali, ma individuando anche fattori di carattere sociale ed economico che in misura diversa influenzano il fenomeno. Una simile analisi ha consentito così di fornire gli strumenti per definire in modo più mirato le strategie di policy da adottare in merito all'argomento trattato, sia in campo nazionale che europeo. La questione tempo è sempre decisiva, ma lo è ancora di più in un periodo come questo, in cui vi è una necessità stringente di una ripresa economica.