rivista di statistica ufficiale

n. 2-3
2010

Temi trattati

Alcune considerazioni sulla qualità degli indici dei valori medi unitari del commercio estero
Paola Anitori, M. Serena Causo e Giuseppe de Santis

A Novel Suite of Methods for Mixture Based Record Linkage
Diego Zardetto, Monica Scannapieco

Proposta per una metodologia di stima dell'impermeabilizzazione del suolo in Italia
Michele Munafò, Gianluigi Salvucci, Marco Zitti e Luca Salvati

Direct vs Indirect Forecasts of Foreign Trade Unit Value Indices
Giancarlo Lutero, Marco Marini

Istat
Istituto Nazionale di Statistica

rivista di statistica ufficiale

n. 2-3
2010

Temi trattati

Istat

Per contattare la redazione o per inviare lavori scrivere a:
Segreteria del Comitato di redazione della Rivista di Statistica Ufficiale
All'attenzione di Gilda Sonetti
Istat - Via Cesare Balbo, 16 - 00184 Roma
e-mail: rivista@istat.it

rivista di statistica ufficiale

n. 2-3/2010

Periodico quadrimestrale
ISSN 1828-1982

Registrazione presso il Tribunale di Roma
n. 339 del 19 luglio 2007

Istituto nazionale di statistica
Servizio Editoria
Via Cesare Balbo, 16 - Roma

Stampato nel mese di Settembre 2011
presso il Centro stampa dell'Istat
Via Tuscolana 1788 - Roma
Copie 300

Alcune considerazioni sulla qualità degli indici dei valori medi unitari del commercio estero

Paola Anitori, M. Serena Causo, Giuseppe de Santis[1]

Sommario

Nel 2008 gli indici dei valori medi unitari prodotti dall'Istat sono stati oggetto di una consistente revisione; a distanza di pochi mesi l'Istituto ha diffuso per la prima volta indici dei prezzi dei prodotti industriali sui mercati esteri. L'ampliamento del set di indicatori, sebbene accolto con favore dagli utilizzatori istituzionali, ha dato un nuovo impulso al dibattito sull'adeguatezza delle misurazioni della statistica ufficiale per l'analisi della competitività e dei processi di internazionalizzazione. In questo lavoro vengono approfonditi alcuni aspetti del calcolo degli indici dei valori medi unitari delle esportazioni nel periodo 2005-2009, allo scopo di valutarne la sensibilità all'utilizzo di diverse opzioni metodologiche, in particolare testandone la robustezza in presenza di sostanziali alterazioni del coverage *di riferimento e all'utilizzo di tecniche di imputazione, nonchè l'accuratezza attraverso la definizione di intervalli di confidenza delle stime.*

Abstract

In 2008 a new series of Unit Value Indexes was officially released; after few months the series of Production Price Indexes on non-domestic markets became available for the first time. The new sets of indicators were favourably welcomed by economists and researchers but they also gave a boost to the well-known debate on the adequateness of the official measurement for the analysis of competitiveness and internationalisation. The aim of this paper is both to test the robustness of export UVI to different coverage option and to perform a sensitivity analysis with regard to imputation techniques; moreover a measure of relative accuracy of monthly estimates will be given with regard to year 2005-2009.

Parole chiave: Valori medi unitari, intervalli di confidenza, distribuzioni non parametriche, dati panel.

Introduzione

Successivamente alla diffusione dei nuovi indici del commercio estero, avvenuta all'inizio del 2008, ed all'avvio della pubblicazione degli indici dei prezzi dei prodotti industriali venduti sui mercati esteri, il dibattito sull'adeguatezza delle misurazioni della performance dell'industria italiana prodotte dalla statistica ufficiale ha subito un'evoluzione: in generale, gli utilizzatori istituzionali e gli economisti hanno tenuto conto delle nuove misurazioni, accogliendo positivamente l'arricchimento del set di indicatori

[1] Primo ricercatore (Istat), e-mail: anitori@istat.it; Ricercatore (Istat), e-mail: causo@istat.it; Collaboratore C.T.E.R. (Istat), e-mail: gdesanti@istat.it . L'introduzione, i par. 1,2 e 3 e le conclusioni sono stati curati da P.Anitori, il par. 4 è stato curato da G. de Santis e il par. 5 da M.S. Causo. Qualsiasi errore od omissione è da attribuirsi esclusivamente agli autori.

congiunturali.[2] Critiche hanno riguardato, ancora, le pratiche di deflazione utilizzate dalla Contabilità Nazionale, e l'assenza di indici dei prezzi dei prodotti importati. In questo quadro, il documento si propone di approfondire alcuni aspetti del calcolo degli indici dei valori medi unitari delle esportazioni (VMUX) nel periodo 2005-2009, allo scopo di valutarne la sensibilità all'utilizzo di diverse opzioni metodologiche.

La nota è impostata come segue. Dopo un breve commento all'andamento degli indici VMUX dei principali partner commerciali del nostro paese pubblicati dall'Eurostat, i VMUX pubblicati dall'Istat vengono confrontati con gli indici dei prezzi all'esportazione (PPIX) (par.2). Nel paragrafo tre vengono esaminate le diversità metodologiche fra le due fonti e verificate alcune ipotesi relative alla possibile origine delle divergenze di andamento tra i due indicatori. Di particolare rilievo è l'evidenza del ruolo svolto dalle modifiche del *mix* di prodotti nella determinazione della dinamica dei VMUX, che viene ridimensionata se si considerano i soli prodotti sistematicamente persistenti in ciascun mese riferibili a specifici *panel* di operatori commerciali.

Proseguendo nell'evidenziazione degli aspetti metodologici di costruzione degli indici, nel par. 4 viene proposta una misura dell'efficienza relativa degli indici VMUX pubblicati attraverso la definizione di intervalli di confidenza delle stime mensili, messe a confronto con omologhi indici calcolati utilizzando metodologie di aggregazione alternative. Infine, l'indice ufficiale VMUX viene confrontato con un indice calcolato utilizzando tecniche alternative di imputazione (par.5) sia a livello sia di singole transazioni, sia di indici elementari utilizzati nelle aggregazioni.

1. Il confronto tra gli indici dei valori medi unitari delle esportazioni calcolati da Eurostat e dall'Istat.

Secondo quanto pubblicato da Eurostat,[3] tra il 2005 ed il 2008 gli indici VMUX (base 2000=100) dell'Italia hanno mostrato una dinamica più accentuata di quella dei principali partner comunitari. A fronte di una variazione media annua sull'intero periodo pari al 5,0% per l'Italia (tavola 1), la Spagna mostra un incremento medio annuo del 4,1% e la Francia del 3,9%, mentre Germania e Regno Unito si attestano intorno al 3%.

Il dato riferito all'Italia di fonte Istat registra invece un incremento medio di periodo lievemente superiore, pari al 5,5%,[4] con un differenziale positivo di crescita, rispetto al dato Eurostat, riscontrabile in tutti gli anni considerati.

---

[2] Fa eccezione la posizione assunta dall'Isae che, in un Rapporto pubblicato nel 2009, esprime il "… dubbio che ci si trovi di fonte al manifestarsi (nuovamente) di un problema di andamento abnorme dei VMU, connesso tanto a difficoltà di misurazione statistica della grandezza quanto a modifiche di composizione del paniere dei beni esportati…" Tale diagnosi viene fatta derivare prevalentemente dal confronto diretto tra i livelli dei VMUX e degli indici dei prezzi all'esportazione dei prodotti industriali (PPIX).

[3] http://epp.eurostat.ec.europa.eu/portal/page/portal/external_trade/data/database

[4] Si sottolinea che Eurostat produce indici dei valori medi unitari dei singoli Stati Membri a partire da indici elementari calcolati ad un livello di dettaglio molto fine (prodotto-paese di origine/destinazione). Nonostante ciò, si tratta di un dato che sconta l'aggregazione diretta dei dati di base, mentre l'Istat, utilizza per il calcolo degli indicatori elementari le singole transazioni relative ai movimenti doganali registrati in ciascuno strato (prodotto-paese), aggregandole solo dopo aver eliminato le transazioni affette da errori di misura. In aggiunta, il metodo di controllo utilizzato da Eurostat per il calcolo degli indici a base mobile risulta assai più severo (in termini di perdita di informazioni in valore) di quello adottato dall'Istat (Anitori P, Causo M.S. , 2008) e ciò contribuisce a spiegare la dinamica più contenuta degli indicatori Eurostat.

Tavola 1 - Indici dei valori medi unitari all'esportazioni dei principali partner europei verso il mondo (base 2000=100). Totale prodotti - Anni 2005-2008 (*indici medi annui, variazioni tendenziali percentuali e variazione % media annua sull'intero periodo*)

ANNI	EUROSTAT					ISTAT
	Germania	Spagna	Francia	Regno Unito	Italia	Italia (a)
2005	104,3	108,5	104,7	106,6	110,6	110,3
2006	108,1	113,7	109,5	112,3	115,8	115,9
2007	110,8	117,9	112,4	113,2	121,2	121,8
2008 (b)	113,7	121,8	116,9	115,8	127,2	128,6
var. 2006/2005	*3,6*	*4,8*	*4,6*	*5,3*	*4,7*	*5,1*
var. 2007/2006	*2,5*	*3,7*	*2,6*	*0,8*	*4,7*	*5,1*
var. 2008/2007	*2,6*	*3,3*	*4,0*	*2,3*	*5,0*	*5,6*
Var. media annua sull'intero periodo (c)	*3,0*	*4,1*	*3,9*	*2,9*	*5,0*	*5,5*

Fonte: Eurostat; Comext database; Istat, Staistiche del commercio con l'estero
(a) Per omogeneità con le informazioni in possesso di Eurostat alla data di realizzazione di questo lavoro, anche gli indici Istat riferiti al 2008 sono provvisori
(b) Dati provvisori.
(c) Calcolata sulla variazione cumulata tra 2008 e 2005.

Considerando le due principali aree di sbocco (tavola 2) il differenziale di crescita tra il nostro paese e i principali concorrenti si mostra assai più contenuto per i VMUX verso l'Ue. In generale, i differenziali minori si riscontrano con la Francia per le esportazioni verso l'Ue (0,5%) e con la Spagna per le esportazioni verso i paesi extra-comunitari (+0,2%), mentre le differenze più elevate, e di pari entità (+3%), si rilevano nei confronti di Germania e Regno Unito in entrambe le aree di destinazione.

Nel periodo considerato l'andamento degli indici dell'Italia risulta quindi sempre più intenso di quello dei partner europei, con una divergenza ampia soprattutto con riferimento ai flussi verso l'area extra-Ue (Grafico 1).

In aggiunta, il confronto diretto tra gli indici VMUX dell'Italia calcolati da Eurostat e gli omologhi indici dell'Istat a partire dall'anno base (2000=100) consente di cogliere una notevole omogeneità tra i due indicatori per quanto riguarda la dinamica dei VMUX verso l'area Ue, mentre è evidente una divergenza rilevante tra gli indicatori relativi ai flussi extra-Ue a partire dal 2007.

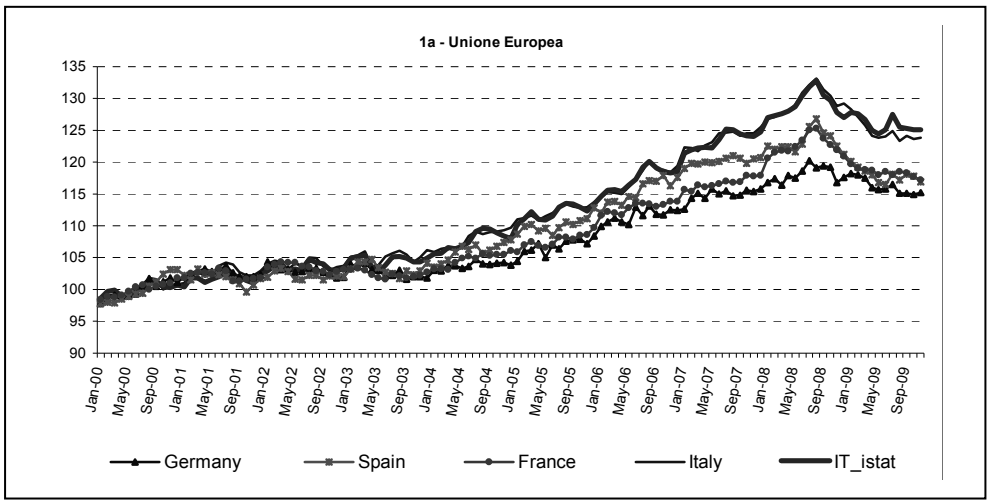
Tavola 2 - Indici dei valori medi unitari all'esportazioni dei principali partner europei per area di destinazione (base 2000=100). Totale prodotti - Anni 2005-2008 *(var. medie annue % sull'intero periodo e differenze assolute tra le var. dell'Italia e quelle dei singoli partner)*

PAESI	Unione europea	Paesi terzi
	VARIAZIONI %	
Germania	3,5	1,8
Spagna	4,0	4,6
Francia	4,6	2,6
Regno Unito	3,6	1,8
Italia (a)	5,1	4,8
	DIFFERENZIALE DI CRESCITA (b)	
Germania	*1,6*	*3,0*
Spagna	*1,2*	*0,2*
Francia	*0,5*	*2,2*
Regno Unito	*1,6*	*3,0*

Fonte: Elaborazioni su dati Eurostat; Comext database
(a) Dati provvisori.
(b) Differenze assolute.

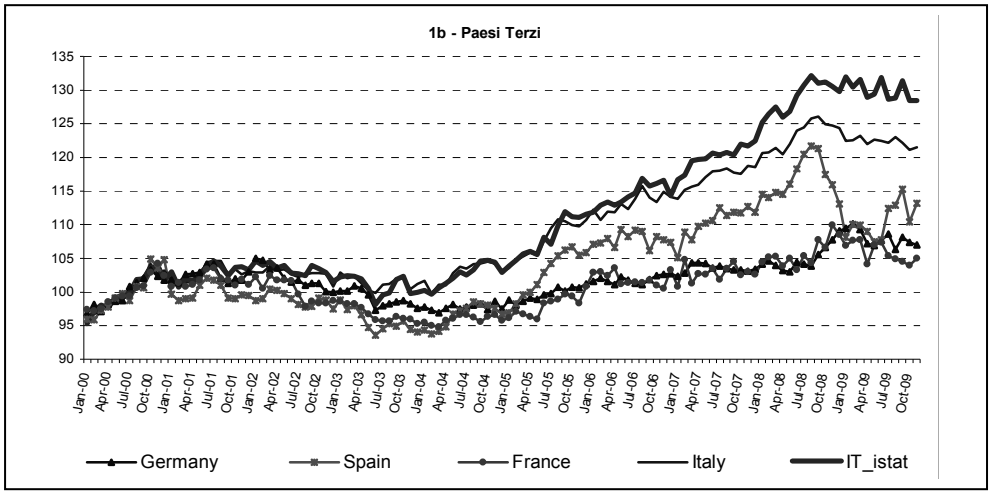
A tale riguardo, una possibile spiegazione potrebbe trovarsi nel fatto che proprio nel 2007 la classificazione delle merci Nomenclatura Combinata (NC) utilizzata per identificare i prodotti ha subito un consistente cambiamento a causa dell'aggiornamento del Sistema Armonizzato da cui deriva.[5]

Grafico 1 - Indici dei valori medi unitari all'esportazioni dei principali partner europei per area di destinazione (base 2000=100). Totale prodotti - Anni 2000-2008 *(indici concatenati)*



[5] La classificazione Sistema Armonizzato è valida a livello mondiale e ha un livello di dettaglio massimo di 6 cifre mentre la NC è la derivazione utilizzata in ambito comunitario e ha un livello di dettaglio di 8 posizioni.

Grafico 1 segue - **Indici dei valori medi unitari all'esportazioni dei principali partner europei per area di destinazione (base 2000=100). Totale prodotti - Anni 2000-2008** *(indici concatenati)*



Fonte: Eurostat, Comext database

La maggiore turbolenza dei flussi all'Extra-Ue,[6] unita ad un diverso trattamento degli indici elementari associati ai nuovi codici di prodotto, potrebbe essere la fonte primaria delle differenze tra i due indicatori.

2. L'andamento dei valori medi unitari all'*export* e dei prezzi alla produzione sui mercati esteri

Senza entrare nel merito delle differenze di tipo metodologico e statistico relative alla costruzione di un indice dei prezzi e di un indice dei valori medi unitari (tema ampiamente trattato in letteratura,[7] nonostante permangano alcune ambiguità sia concettuali sia, soprattutto,

[6] Dato l'elevato numero di paesi di destinazione, vi è di fatto una maggiore polverizzazione degli scambi con i Paesi Terzi. Si noti che l'Eurostat non ha mai fornito spiegazioni sul trattamento dei cambiamenti della classificazione delle merci.

[7] Secondo quanto evidenziato nei manuali internazionali (FMI, 2009), gli elementi metodologici e operativi che diversificano i VMUX dai PPIX sono riassumibili nei seguenti punti:

- diversità del campo di osservazione (universo delle transazioni Vs indagine camponaria);
- diversa identificazione del momento in cui avviene la transazione (*cross-border* Vs. chiusura del contratto di vendita): ciò potrebbe influenzare il valore della transazione e il prezzo di vendita;
- diversa definizione di quantità scambiata (*shipping* Vs. *transaction quantity*) che influenza a sua volta la definizione di prezzo unitario;
- difficoltà di identificare e seguire nel tempo le diverse varietà (*item*) di un singolo prodotto (effetti di composizione);
- diversa gestione delle entrate/uscite di prodotti dal campo di osservazione;
- impossibilità di isolare variazioni dovute ad un puro effetto di prezzo (da cui la diversa gestione dei cambiamenti di qualità dei prodotti);
- diversa incidenza di valori anomali nei microdati;
- possibili *misclassification* dei prodotti;
- diversa copertura del fenomeno (indagini campionarie Vs. informazioni censuarie).

Gli elementi sopra elencati - con riferimento agli indici dei prezzi all'esportazione -sono tutti singolarmente circostanziati nei manuali internazionali e contribuiscono ad identificare le cosiddette "*price determining characteristics*".

di tipo metodologico e tecnico) vengono mostrati di seguito alcuni confronti tra i VMUX e i PPIX, con riferimento al totale generale, alla destinazione d'uso dei beni e alle aree di sbocco.

In via preliminare, giova ricordare che il confronto tra le dinamiche dei due indici risente, tra le altre cose, anche della diversità del campo di osservazione. I PPIX si riferiscono ai soli beni industriali prodotti ed esportati direttamente da imprese industriali appartenenti ai settori B,C,D ed E della classificazione ATECO2007;[8] i VMUX si riferiscono al complesso delle merci (industriali, agricole e altro) vendute all'estero da qualsiasi impresa (industriale, commerciale o dei servizi) attiva sui mercati internazionali.

Una quota non trascurabile dei beni venduti sui mercati esteri (circa il 12% in valore), la cui provenienza non è tutta imputabile alla produzione di origine interna, proviene da imprese commerciali che, a loro volta, rappresentano oltre il 35% del totale delle imprese esportatrici italiane.[9] Inoltre, un aspetto da non trascurare riguarda il tipo di formula con cui vengono calcolati i due indici (Laspeyres a base fissa per i PPIX e Fisher concatenato per i VMUX).[10]

Il confronto tra l'andamento dei due indicatori (tavola 3) per Raggruppamenti Principali di Industrie (RPI), mostra che le variazioni medie annue dei VMUX sono sistematicamente superiori a quelle registrate per i prezzi all'esportazione in ciascun anno del periodo considerato. Dal 2005 al 2008 e con riferimento all'Indice generale, il differenziale tra le variazioni dei due indici è di poco meno di 3 punti percentuale all'anno, anche al netto dell'energia.

Rispetto alle diverse destinazioni dei beni, differenze di oltre 3 punti percentuali si riscontrano in alcuni anni nei tassi di variazione del comparto dei beni intermedi e di quello dei beni di consumo durevoli, mentre negli altri comparti nel 2008 le differenze sembrano ridursi.

Tavola 3 - Indici dei prezzi alla produzione sui mercati esteri e dei valori medi unitari l'esportazione per RPI (base 2005=100) - Anni 2005-2008 (*variazioni tendenziali*)

ANNI	Totale	Totale al netto dell'energia (a)	Beni inter-medi	Beni stru-mentali	Beni di consumo durevoli	Beni di consumo non durevoli	Beni di consumo	Energia
PREZZI ALLA PRODUZIONE SUI MERCATI ESTERI								
var. 2006/2005	2,2	1,5	2,5	0,9	0,2	1,6	1,2	15,8
var. 2007/2006	2,3	2,2	4,0	1,7	1,2	0,8	0,9	1,9
var. 2008/2007	2,8	1,7	1,6	0,6	2,7	3,3	3,1	23,2
var. gen-nov 2009/var. gen-nov 2008	-2,7	-1,3	-4,1	-0,2	1,6	0,3	0,8	-23,1
VALORI MEDI UNITARI ALL'ESPORTAZIONE								
var. 2006/2005	5,1	4,5	6,3	2,8	4,4	4,7	4,7	19,1
var. 2007/2006	5,1	5,1	6,5	4,3	5,2	4,4	4,7	3,6
var. 2008/2007	5,6	4,4	4,1	4,5	3,9	5,2	4,8	27,0
var. gen-nov 2009/var. gen-nov 2008	-1,1	1,5	-3,1	5,1	2,5	1,9	2,0	-35,2

Fonte: Istat, Statistiche del commercio con l'estero e Statistiche dei prezzi alla produzione

(a) Per i prezzi alla produzione il totale fà riferimento al totale prodotti dell'industria comprendente le sezioni B e C al netto delle attività legate alle industrie navali, aerospaziali, ferroviarie, degli armamenti e ai servizi industriali. Per i valori medi unitari invece il totale è riferito alle sezioni A, B, C, D, E, J, M, R ed S.

[8] Fanno eccezione alcune attività economiche secondo quanto stabilito dal Regolamento STS (EUROSTAT,2005).

[9] In questo senso l'affermazione "…i beni esportati dall'industria costituiscono il 99% delle esportazioni italiane di merci…" (ISAE, 2009) risulta inesatta.

[10] Si veda Anitori P, Causo M.S. (2008) op.cit.

Analoghi ordini di grandezza si riscontrano con riferimento alle due maggiori aree di sbocco (tavola 4), anche se con andamenti opposti: mentre il differenziale tra i due indicatori con riferimento all'area Uem si è nettamente ridotto nel corso del tempo, quello riferito all'area dei Paesi terzi si è amplificato.

Tavola 4 - Indici dei prezzi alla produzione sui mercati esteri e dei valori medi unitari all'esportazione per destinazione (base 2005=100) - Anni 2005-2008 (*indici e variazioni % tendenziali*)

ANNI	Prezzi alla produzione sui mercati esteri		Valori medi unitari all'esportazione	
	Uem	Extra-Uem	Uem	Extra-Uem
2005	100,0	100,0	100,0	100,0
2006	101,8	102,6	104,8	105,9
2007	104,7	104,2	109,9	111,2
2008	108,3	106,7	115,1	119,2
gen-nov 2008	108,4	106,9	115,3	119,2
gen-nov 2009	106,1	103,5	111,8	120,4
var. 2006/2005	*1,8*	*2,6*	*4,8*	*5,9*
var. 2007/2006	*2,8*	*1,6*	*4,9*	*5,0*
var. 2008/2007	*3,4*	*2,4*	*4,7*	*7,3*
var. gen-nov 2009/var. gen-nov 2008	*-2,1*	*-3,2*	*-3,0*	*1,0*

Fonte: Istat, Statistiche del commercio con l'estero e Statistiche dei prezzi alla produzione

A scopo illustrativo si riportano alcune informazioni sintetiche sul numero di prodotti, di quotazioni e di imprese, riferite sia alla rilevazione campionaria dei prezzi all'esportazione sia alle rilevazioni del commercio con l'estero (tavola 5) utilizzate per l'elaborazione dei valori medi unitari all'*export*, distintamente per area di destinazione dei prodotti.

Tavola 5 - Prodotti, quotazioni e imprese coinvolte nelle elaborazioni degli indici dei prezzi alla produzione sui mercati esteri e dei valori medi unitari all'esportazione - Anno 2005

AREA	Prezzi alla produzione sui mercati esteri			Valori medi unitari all'esportazione		
	N. prodotti (a)	Quotazioni	Imprese	N. prodotti (b)	Transazioni (c)	Operatori (d)
Uem16	704	3.349	1.312	8.992	458.869	44.847
Extra-Uem	671	3.350	1.251	9.106	785.194	175.300
Totale	**933**	**6.699**	**2.017**	**9.466**	**1.244.063**	**149.664**

Fonte: Istat, Statistiche congiunturali e Statistiche del commercio con l'estero
(a) Classificazione Prodcom a 8 cifre.
(b) Classificazione CN a 8 cifre. Prodotti effettivamente venduti sui mercati esteri nell'anno.
(c) Transazioni effettivamente utilizzate nel calcolo degli indici a base mobile.
(d) Solo operatori con obbligo di dichiarazione mensile. Il totale non corrisponde alla somma dei parziali in quanto gli operatori che sono attivi contemporaneamente all'Uem e all'Extra-Uem sono contati una sola volta.

Per una corretta valutazione delle informazioni riportate si tenga presente che la rilevazione dei PPIX adotta la classificazione Prodcom, la quale individua circa 5.500 prodotti industriali: per ciascun prodotto, le imprese presenti nel campione individuano una specifica varietà di cui seguono nel tempo l'evoluzione del prezzo. Nella stima dei VMUX, invece, nonostante la classificazione NC sia decisamente più dettagliata della Prodcom (si individuano circa 9.500 prodotti, anche non tutti appartenenti ai comparti dell'industria), confluiscono informazioni relative a tutte le possibili varietà dei singoli prodotti: ciò determina il tipico *effetto mix* che, anche a livelli di dettaglio molto fini, può causare un'elevata variabilità interna del valore medio unitario del prodotto.

Trattandosi di due indicatori distinti, ammessi entrambi dai manuali internazionali per la deflazione degli aggregati di commercio estero,[11] l'effetto che essi hanno sulla dinamica "reale" delle esportazioni risulta molto differenziata. A tale riguardo la tavola 6 riporta i risultati della deflazione confrontando sia il valore reale delle esportazioni ottenuto utilizzando i dati di commercio estero, sia il valore reale del fatturato estero delle imprese industriali deflazionato con i PPIX.

Tavola 6 - Fatturato all'esportazione dell'industria a prezzi costanti ed esportazioni in volume - Anno 2005-2008 *(indici e variazioni % tendenziali)*

ANNI	Indice dei volumi delle esportazioni	Indice del fatturato all'export deflazionato
2005	100,0	100,0
2006	105,4	107,9
2007	110,4	117,0
2008	105,9	114,0
gen-nov 2008	107,4	114,9
gen-nov 2009	83,8	89,8
var. 2006/2005	*5,4*	*7,9*
var. 2007/2006	*4,7*	*8,4*
var. 2008/2007	*-4,0*	*-2,5*
var. gen-nov 2009/var. gen-nov 2008	*-22,0*	*-21,8*

Fonte: Istat, Statistiche della produzione industriale e Statistiche del commercio con l'estero

Da questi dati emerge una decisa divergenza fra i risultanti indicatori dei volumi esportati per gli anni 2005-2007 e un progressivo avvicinamento nel corso del biennio 2008-2009.

[11] Il Sistema dei Conti Nazionali SNA95 definisce i PPIX come la scelta ottimale per la deflazione degli aggregati di Contabilità Nazionale, mentre i VMUX rappresentano un *second best* (EUROSTAT; 1996)

3. Alcuni elementi utili per valutare la "robustezza" dei numeri indici dei valori medi unitari all'esportazione

Secondo quanto illustrato nei paragrafi precedenti, la misurazione dell'andamento del commercio estero in termini di dinamica dei valori medi unitari/prezzi e dei volumi mostra elevate differenziazioni a seconda dell'indicatore utilizzato. D'altra parte, le profonde differenze strutturali e metodologiche che caratterizzano le due fonti possono giustificare differenze così ampie nella rappresentazione delle dinamiche del commercio estero.

Come parte della manualistica internazionale sull'argomento ha evidenziato, in assenza di una metodologia ampiamente condivisa e codificata (che invece caratterizza altri domini delle statistiche economiche congiunturali), è possibile che la stima dell'andamento dei VMU possa risentire in misura significativa delle opzioni metodologiche e delle tecniche di trattamento dei dati che vengono utilizzate.

In questo paragrafo verrà valutata la sensibilità degli indici ufficiali del commercio con l'estero alle alterazioni nella composizione del paniere dei beni sulla base del quale vengono calcolati, attraverso un confronto con gli omologhi indicatori ottenuti ricorrendo ad opportuni panel di prodotti persistentemente presenti nelle transazioni mensili. In tal senso, il concetto di "robustezza" degli indicatori che più o meno implicitamente viene richiamato è inteso in senso lato e fa riferimento alla stabilità dei risultati del processo di stima al variare di alcune condizioni legate al trattamento dei dati di base.[12]

3.1 I prodotti utilizzati nel calcolo degli indici a base mobile: frequenze annuali, mensili e pesi in valore.

La struttura dei beni esportati annualmente dal nostro Paese, sulla base della quale vengono calcolati gli indici dei VMU, appare piuttosto differenziata tra le due grandi aree di destinazione, sebbene appaia comunque sostanzialmente stabile nel tempo.

Il numero di prodotti esportati verso l'area comunitaria in ciascun anno del periodo 2005-2008 risulta di circa il 5% più elevato del numero di prodotti venduti sui mercati terzi (Tavola 7), mentre poco più del 90% dei beni che complessivamente si movimentano ogni anno vengono esportati in entrambe le aree.

Anche le percentuali del numero dei prodotti esportati verso le due macro aree restano praticamente immutate nel tempo, evidenziando un leggero incremento solo nel 2008 con riferimento alle destinazioni extra-comunitarie.[13]

Entrando nel merito degli effetti legati alla procedura adottata per la stima dei VMUX riferiti al totale generale, il numero di indici elementari[14] mensilmente utilizzabili ai fini del calcolo degli indici generali a base mobile è, secondo i criteri definiti dalla procedura di *trimming adattativo* adottata dall'Istat, pari a circa 58 mila all'Ue e circa 45 mila all'Extra-Ue.

---

[12] In tal modo esso viene distinto dal concetto di robustezza statistica in senso stretto che sarà richiamata più avanti (cfr. par. 4.1.).

[13] Ciò tenuto conto della riduzione complessiva del numero di prodotti (evidenziata nella tavola 7) che deriva dal processo semplificazione della classificazione delle merci NC attivato dall'Eurostat a partire dal 2005. Nell'ambito delle strategie comunitarie di riduzione della molestia statistica sulle imprese, la semplificazione della NC assume un ruolo preponderante; dal 2005 ad oggi il numero di merci della classificazione NC è passato da un totale di 10.096 a 9.569 ed è destinato ad ulteriori riduzioni. Poiché il processo di semplificazione ha come conseguenza prevalente l'accorpamento dei prodotti che non mostrano alti valori di traffico a livello comunitario, è presumibile che l'effetto *mix* che caratterizza i VMU e che produce un aumento della variabilità degli indicatori anche a livelli di dettaglio molto fini sia destinato ad aumentare.

[14] Gli indici elementari sono gli indici corrispondenti a ciascuno strato prodotto-paese di destinazione-flusso-mese.

Tavola 7 - Indici elementari (a) utilizzati nel calcolo degli indici dei VMU all'esportazione a base mobile per area di destinazione. Indice generale - Anni 2005-2008 (*numero prodotti e quota % sul totale*)

AREA	2005	2006	2007	2008
	TOTALE PRODOTTI			
Ue	9.170	9.009	8.902	8.874
Extra-Ue	8.791	8.629	8.508	8.531
Totale	**9.466**	**9.275**	**9.157**	**9.135**
Prodotti in comune tra le due aree (b)	*89,7*	*90,2*	*90,1*	*90,5*
	QUOTA SUL TOTALE			
Ue	96,9	97,1	97,2	97,1
Extra-Ue	92,9	93,0	92,9	93,4
Totale	**100,0**	**100,0**	**100,0**	**100,0**

Fonte: Istat, Statistiche del commercio con l'estero
(a) Il numero di prodotti è decrescente a causa del processo di semplificazione della Nomenclatura Combinata.
(b) Percentuale di merci in comune tra le due aree di destinazione sul numero totale annuo di prodotti esportati.

Di questi, in media oltre il 92% (corrispondente a circa il 90% in termini di valori esportati) non viene sottoposto ad alcuna procedura di stima[15] o imputazione (Tavola 8); inoltre, la percentuale (in valore) degli indici elementari che annualmente non superano il test della procedura di *trimming* e che pertanto non vengono inclusi nel calcolo degli indici a base mobile varia tra il 10% ed il 20% del totale.

Tavola 8 - Indici elementari utilizzati nel calcolo degli indici a base mobile dei VMU all'esportazione per tipo di trattamento. Indice generale - Anni 2005-2008 (*composizione percentuale in valore*)

TIPO	2005	2006	2007	2008
Originali	94,3	92,5	91,9	94,3
Basi ricostruite	0,8	0,8	0,8	0,7
Sterilizzazioni	4,3	4,3	4,6	4,6
Trasposizioni	0,7	2,4	2,7	0,5
Totale	**100,0**	**100,0**	**100,0**	**100,0**

Fonte: Istat, Statistiche del commercio con l'estero

Poiché una delle caratteristiche principali degli indici dei valori medi unitari è la continua modifica della composizione del "paniere" di beni sulla base del quale essi vengono mensilmente calcolati,[16] un'ulteriore valutazione della portata informativa dell'indicatore generale consiste nel misurare gli eventuali effetti distorsivi associati a queste alterazioni nella composizione dei prodotti, limitando ad esempio la stima ai soli prodotti persistentemente presenti nelle transazioni mensili.

[15] Si fa riferimento alla stima dovuta, ad esempio, alla mancanza di informazioni nell'anno base, alla presenza di cambiamenti nelle classificazioni (trasposizioni), a trattamenti *ad hoc* riservati a merci particolari (navi, aerei, sostanze farmaceutiche ecc.)

[16] La composizione del "paniere" dipende da fattori legati non solo alla possibile stagionalità di alcune merci riscontrabile in molti settori (alimentari, abbigliamento ecc.) ma anche al comportamento delle imprese esportatrici e alla maggiore o minore persistenza della loro attività sui mercati esteri.

Tale persistenza può essere valutata in senso longitudinale o *cross-section,* controllando cioè il numero e il peso in valore degli indici elementari in comune tra i mesi omologhi degli anni sui cui si calcolano le variazioni tendenziali mensili, oppure tra i mesi contigui dello stesso anno per quanto attiene alle variazioni congiunturali.

Facendo riferimento alla prima opzione, la tavola 9a mostra come poco meno di due terzi degli strati utilizzati per calcolare le variazioni tendenziali mensili siano presenti in entrambi gli anni.

Tavola 9a - Indici dei VMU all'esportazione. Indici elementari in comune tra mesi omologhi di due anni contigui, utilizzati nel calcolo delle variazioni tendenziali mensili - Anni 2005-2008 *(numero prodotti, numero indici elementari, peso % e variazioni % del valore)*

MESI DI CALCOLO DELLA VARIAZIONE	2006/2005				2007/2006				2008/2007			
		Indici elementari				Indici elementari				Indici elementari		
	N. prodotti	Numero	Peso (a)	Valore (b)	N. prodotti	Numero	Peso (a)	Valore (b)	N. prodotti	Numero	Peso (a)	Valore (b)
Gennaio	6.347	57.013	58,1	*14,0*	6.165	61.651	58,3	*12,5*	6.643	65.647	62,1	*1,7*
Febbraio	6.467	61.673	58,9	*12,9*	6.285	67.082	59,3	*9,2*	6.782	71.967	64,2	*9,3*
Marzo	6.546	64.712	59,7	*16,4*	6.360	70.890	59,5	*5,1*	6.822	73.902	63,0	*-4,9*
Aprile	6.420	61.138	57,3	*1,5*	6.241	65.189	58,9	*11,2*	6.832	72.190	65,6	*19,2*
Maggio	6.518	63.442	59,0	*15,8*	6.356	69.277	59,1	*8,1*	6.882	73.726	64,1	*2,0*
Giugno	6.466	61.761	58,3	*11,5*	6.297	67.318	59,0	*12,6*	6.750	70.891	62,6	*-6,2*
Luglio	6.410	61.990	58,0	*7,2*	6.304	67.694	59,4	*16,0*	6.833	73.811	64,4	*6,9*
Agosto	5.833	48.870	55,3	*15,2*	5.818	54.391	56,4	*9,0*	6.205	56.572	58,8	*-7,9*
Settembre	6.439	60.733	58,1	*7,8*	6.206	64.903	58,3	*5,3*	6.724	69.447	63,5	*6,1*
Ottobre	6.480	62.415	58,8	*17,4*	6.294	68.346	59,0	*10,3*	6.835	73.232	63,6	*-1,7*
Novembre	6.437	61.371	57,9	*12,5*	6.261	66.457	58,1	*4,6*	6.708	68.323	60,8	*-16,2*
Dicembre	6.291	57.276	56,9	*10,5*	6.083	61.628	56,2	*-3,8*	6.503	62.865	60,4	*-9,2*

Fonte: Istat, Statistiche del commercio con l'estero

(a) Quota % di indici elementari in comune sul totale del mese nell'anno base.

(b) Variazione percentuale del valore mensile delle esportazioni associato agli indici elementari in comune nei due anni a confronto.

Esiste, pertanto un certo numero di prodotti e/o destinazioni (circa un terzo del totale) non persistenti, che potenzialmente influenzano l'ampiezza delle variazioni dell'indice e la cui presenza è legata a fattori strutturali (comportamenti delle imprese, fattori esogeni legati al ciclo, sostituzione di prodotti o destinazioni ecc.) o ad aspetti metodologici (modifica delle classificazioni dei prodotti, che rendono non rintracciabili alcuni codici di prodotto da un anno al successivo). I valori esportati associati ai flussi persistenti evidenziano, comunque, un aumento dell'*export* in ciascun anno ad eccezione del 2008 per effetto del mutato quadro globale.

Prove effettuate con riferimento alle variazioni congiunturali (tavola 9b) hanno dato risultati del tutto analoghi evidenziando tuttavia percentuali di persistenza mensili sensibilmente superiori (circa il 70%). Ciò dipende dal fatto che nel caso dei confronti congiunturali all'interno di ciascun anno la componente legata ai cambi di classificazione è inesistente.

Tavola 9b - Indici dei VMU all'esportazione. Indici elementari in comune tra mesi omologhi di due anni contigui, utilizzati nel calcolo delle variazioni tendenziali mensili - Anni 2005-2008 *(numero prodotti, numero indici elementari, peso % e variazioni % del valore)*

MESI DI CALCOLO DELLA VARIAZIONE	2006				2007				2008			
		Indici elementari				Indici elementari				Indici elementari		
	N. prodotti	Numero	Peso (a)	Valore (b)	N. prodotti	Numero	Peso (a)	Valore (b)	N. prodotti	Numero	Peso (a)	Valore (b)
Feb./Gen.	7.175	75.336	71,2	*11,5*	7.110	73.916	69,9	*7,8*	7.092	75.265	70,6	*16,2*
Mar./Feb.	7.287	81.302	71,9	*17,8*	7.215	79.982	71,4	*14,1*	7.204	79.011	68,9	*-1,3*
Apr./Mar.	7.263	80.168	67,3	*-17,4*	7.247	78.925	67,3	*-15,3*	7.247	79.988	70,2	*4,5*
Mag./Apr.	7.258	79.494	71,8	*17,7*	7.205	78.123	71,0	*16,3*	7.242	80.829	69,1	*-0,8*
Giu./Mag.	7.327	81.092	69,2	*0,0*	7.217	79.447	69,1	*2,5*	7.126	78.003	67,1	*-3,9*
Lug./Giu.	7.255	79.480	69,7	*0,9*	7.224	79.110	69,9	*0,0*	7.180	78.352	70,9	*19,4*
Ago./Lug.	6.909	69.749	61,2	*-24,5*	6.924	69.791	60,9	*-28,0*	6.791	67.728	57,9	*0,0*
Set./Ago.	6.895	69.104	71,7	*24,9*	6.853	67.993	70,7	*18,3*	6.744	65.551	71,2	*36,7*
Ott./Set.	7.246	79.171	71,1	*10,8*	7.193	77.761	71,1	*16,7*	7.194	77.876	70,4	*7,0*
Nov./Ott.	7.270	80.571	69,5	*-0,8*	7.226	78.979	68,6	*-7,7*	7.135	76.090	65,8	*-19,0*
Dic./Nov.	7.145	77.071	67,4	*-8,3*	7.052	73.737	65,6	*-14,7*	6.928	69.772	64,8	*-6,4*

Fonte: Istat, Statistiche del commercio con l'estero
(a) Quota % di indici elementari in comune sul totale del mese nell'anno base.
(b) Variazione percentuale del valore mensile delle esportazioni associato agli indici elementari in comune nei due anni a confronto.

3.2 Indici dei VMU all'export calcolati secondo metodi basati su panel di dati elementari

Quantificata l'influenza potenziale delle modifiche che intervengono nella composizione dell'*export* mensile, ulteriori elementi di valutazione su base empirica della "robustezza" e della capacità informativa degli indici dei valori medi unitari correntemente diffusi possono essere dedotti aggregando gli indici elementari accettati dalla procedura di controllo e correzione secondo logiche tra loro alternative, pur mantenendo lo stesso impianto formale di calcolo che prevede l'uso della formula di Fisher[17] sia per gli indici a base mobile sia per il concatenamento all'anno base.

In particolare, per verificare il possibile disturbo introdotto dalla "vischiosità" del paniere mensile dei beni esportati sono stati estratti due tipi di *panel* di dati sui quali effettuare le aggregazioni:

1) un primo *panel* chiuso costituito solo dai prodotti che sono stati esportati in ciascun mese di ciascun anno dal 2005 al 2008 verso le medesime destinazioni.

Si tratta di 5.831 prodotti esportati nelle 54 aree di destinazione che costituiscono la stratificazione geografica, per un totale annuo di 411.792 indici elementari (34.316 indici mensili). Questa soluzione neutralizza completamente gli effetti indotti dalle variazioni annuali della classificazione delle merci, ma ha il limite che con il passare del tempo i risultati tendono a diventare meno rappresentativi poiché non tiene conto di eventi molto frequenti nella realtà quali l'entrata/uscita di prodotti dal campo di osservazione.

[17] Per approfondimenti si veda Istat (2003).

2) *Panel* "biennali" costruiti in modo che, per ciascun anno successivo al 2005, l'indice a base mobile mensile venga calcolato utilizzando solo i prodotti presenti anche nell'anno precedente, essendo quest'ultimo il corrispondente anno base.

Rispetto al *panel* chiuso sull'intero periodo, questa soluzione rende più solidi gli indici a base mobile ed usa più informazioni (oltre 700 mila indici elementari all'anno), anche se può fornire risultati problematici se tra i due anni confrontati si manifestano modifiche radicali della classificazione (come avviene normalmente ogni 5 anni per effetto dei cambiamenti del Sistema Armonizzato; l'ultima modifica è avvenuta nel 2007 e ha riguardato oltre 1.300 merci).

3) Infine, per facilitare un confronto con gli indici dei prezzi alla produzione sui mercati esteri, i VMUX sono stati calcolati considerando il medesimo campo di osservazione dei PPIX, costituito dalle sezioni B, C e D della classificazione ATECO2007 da cui si escludono le attività industriali legate alla produzione di aeromobili, veicoli aerospaziali, materiale ferroviario, combustibili nucleari ecc., secondo quanto stabilito dal Regolamento comunitario STS.

La tavola 10 riassume le caratteristiche delle tre opzioni.

Tavola 10 - Indici elementari utilizzati nel calcolo degli indici dei VMU all'esportazione a base mobile, secondo il campo di osservazione - Anni 2005-2008 (*numero*)

ANNI	Indici pubblicati	Indici "panel"		Campo di osservazione dei PPIX (b)
		Panel chiuso 2005-2008	Panel "biennale" (a)	
		NUMERO PRODOTTI		
2005	9.466	5.831	…	8.689
2006	9.275	5.831	7.785	8.506
2007	9.157	5.831	7.435	8.396
2008	9.135	5.831	8.083	8.362
		INDICI ELEMENTARI		
2005	1.244.063	411.793	…	1.068.504
2006	1.341.491	411.793	722.394	1.158.804
2007	1.325.367	411.793	784.826	1.145.153
2008	1.324.461	411.793	832.573	1.140.487

Fonte: Istat, Statistiche del commercio con l'estero
(a) I panel biennali considerati sono stati costruiti a partire dal 2006.
(b) Indici VMUX calcolati considerando il medesimo campo di osservazione utilizzato nel calcolo dei PPIX.

I risultati dell'esercizio (tavola 11) mostrano limitate differenze tra gli indici calcolati secondo i quattro approcci considerati. Posto il 2005 uguale a 100, l'indice ufficiale relativo al 2008 risulta pari a 116.7 in media d'anno a fronte di un indice pari a 118.4 nel caso del *panel* chiuso costruito sull'intero periodo, e all'indice pari a 117.7 nel caso del *panel* biennale. In generale, gli indici calcolati sulla base di insiemi persistenti di prodotti mostrano dinamiche più accentuate rispetto agli indici ufficiali e ciò probabilmente sottintende il fatto che i prodotti meno persistenti sono quelli con un VMU mediamente più basso.

Tavola 11 - Indici dei valori medi unitari all'esportazione secondo il campo di osservazione - Anni 2005-2008 (*indici a base 2005=100 e variazioni % tendenziali*)

ANNI	Indici pubblicati	Indici "panel"		Campo di osservazione dei PPIX (b)	
		Panel chiuso 2005-2008	Panel "biennale" (a)	Intero data set	Panel chiuso
2006	105,1	106,1	105,9	105,2	106,1
2007	110,5	111,7	111,2	110,6	111,7
2008	116,7	118,4	117,7	116,7	118,2
Var. 2006/2005	*5,1*	*6,1*	*5,9*	*5,2*	*6,1*
Var. 2007/2006	*5,1*	*5,3*	*5,1*	*5,0*	*5,3*
Var. 2008/2007	*5,7*	*6,0*	*5,8*	*5,5*	*5,9*

Fonte: Istat, Statistiche del commercio con l'estero
(a) Gli indici a base mobile di ciascun anno sono calcolati considerando solo i prodotti in comune con l'anno base.
(b) Indici VMUX calcolati considerando il medesimo campo di osservazione utilizzato nel calcolo dei PPIX.

4. Una misura di accuratezza degli indici dei VMU pubblicati dall'Istat: intervalli di confidenza per distribuzioni non parametriche.

4.1 La misura dell'accuratezza degli indici

Un ulteriore elemento di valutazione della qualità degli indici pubblicati è determinato dalla definizione degli intervalli di confidenza delle stime che, come noto, possono essere visti come una misura dell'accuratezza (in probabilità) delle stesse. L'ampiezza dell'intervallo fornisce indicazioni sull'errore di cui è potenzialmente affetta la stima (ad intervalli più stretti corrispondono stime affette da un errore più piccolo) anche se un giudizio più completo sulla qualità della stessa non può prescindere dalla considerazione di alcuni parametri caratteristici di cui si darà conto nel proseguo.

Nel caso specifico degli indici dei VMU del commercio estero la metodologia di stima di tali intervalli è stata definita nel contesto degli approcci tipici dell'analisi di distribuzioni non parametriche in quanto più adatti alle caratteristiche del fenomeno oggetto di studio; tali approcci individuano soluzioni specifiche per gestire la pressoché ineliminabile "distanza" della distribuzione empirica della variabile dalle ipotesi su cui si basano i metodi di stima classici (ipotesi di normalità e simmetria delle distribuzioni) degli intervalli di confidenza. Poiché l'elemento qualificante del processo di stima degli indici dei VMU è rappresentato dal ricorso a metodi di *trimming* in grado di generare stime più robuste[18] dei metodi che utilizzano tutte le informazioni disponibili, la valutazione qualitativa del risultato ufficiale non può che realizzarsi attraverso un processo comparativo che metta a confronto la stima prodotta con il metodo di *adaptive trimming*[19] utilizzato per gli indici pubblicati con analoghe stime derivanti dall'uso di altri tipi di medie troncate, *in primis* il

[18] In questo ambito il concetto richiamato è quello di robustezza statistica in senso stretto (Wilcox,2006).
[19] Il *trimming adattativo* stabilisce i punti di troncamento sulla base di intervalli asimmetrici che si adattano alla distribuzione empirica sottostante (si veda Anitori, Causo 2008 op.cit.).

metodo utilizzato da Eurostat per il calcolo dei propri indici VMUX basato sul troncamento ad intervalli fissi che sembra essere considerato da alcuni più affidabile.[20]

Nel paragrafo che segue la bontà dei risultati dei metodi messi a confronto verrà valutata affiancando alla misura di accuratezza legata agli intervalli di confidenza riferiti a ciascun indice dei VMUX un *set* di indicatori statistici complementari in grado di fornire una visione più completa sull'accuratezza della stima ufficiale.

4.2 Intervalli di confidenza per distribuzioni non parametriche

Come accennato in precedenza, poiché le distribuzioni empiriche dei VMU violano le ipotesi classiche su cui gli intervalli di confidenza vengono stimati (normalità e simmetria), questi possono essere costruiti solo attraverso tecniche di *bootstrap*.

Il *bootstrap* è un metodo di campionamento ricorsivo effettuato sul medesimo *set* di osservazioni empiriche $X = (X_1, X_2, ..., X_n)$ che consente di ottenere, per ogni campione *i (i=1,2……..B)* di dimensione *n* estratto (con ripetizione) da *X*, una stima $\hat{\theta}_i^*$ che permetta di costruire un intervallo di confidenza per la stima empirica $\hat{\theta}$. A partire dal vettore $\hat{\theta}^* = (\hat{\theta}_1^*, \hat{\theta}_2^*, ..., \hat{\theta}_B^*)$ delle stime *bootstrap* corrispondenti ai *B* campioni estratti da *X* si costruisce un *range* entro il quale cade con probabilità (1-α) il valore calcolato $\hat{\theta}$. In particolare, la stima $\hat{\theta} = \overline{X}_t$ è l'indice VMUX calcolato con la media troncata mentre il vettore delle *B* stime *bootstrap*

$$\overline{X}_t^* = (\overline{X}_{t1}^*, \overline{X}_{t2}^*, ..., \overline{X}_{tB}^*) \quad (1)$$

è quello utilizzato per la costruzione dell'intervallo di confidenza. Ovviamente, per definire l'efficienza della media *trimmed* $\overline{X}_t$ rispetto a indici ottenuti secondo altri metodi di troncamento, l'ampiezza dell'intervallo di confidenza e i valori assunti dai parametri caratteristici della distribuzione delle stime $\hat{\theta}_i^*$ hanno un ruolo fondamentale. Secondo la letteratura corrente[21] il metodo di *bootstrap* più adatto alla stima dell'intervallo per una media troncata è il metodo del percentile secondo cui, deciso il livello di significatività α e ordinate in senso crescente le B stime $\overline{X}_{ti}^*$ in modo che il vettore (1) possa essere riscritto come segue:

$$\overline{X}_t^* = (\overline{X}_{t(1)}^*, \overline{X}_{t(2)}^*, ..., \overline{X}_{t(B)}^*) \quad (2)$$

[20] Nel dibattito corrente che si svolge a livello nazionale parte degli utilizzatori istituzionali (Banca d'Italia, Isae) sembrano privilegiare la metodologia di *trimming* utilizzata dall'Eurostat basata sull'uso di un intervallo fisso pari a (0,5;2); tuttavia, la presunta supremazia di tale metodo non è mai stata supportata da alcuna spiegazione statistica.

[21] Per approfondimenti sul *bootstrap* si veda Wilcox, Rand R. (2006).

l'intervallo di confidenza stimato è il seguente:

$$\left(\overline{X}^{*}_{t(l+1),}\, \overline{X}^{*}_{t(u)}\right) \tag{3}$$

dove $l = \alpha \frac{B}{2}$ (arrotondato alla parte intera del valore corrispondente) indica l'estremo inferiore mentre $u = B - l$ indica l'estremo superiore dell'intervallo stimato entro cui cade con probabilità α la media troncata $\hat{\theta} = \overline{X}_t$.

Nel caso degli indici dei VMU, sono stati estratti per ciascun mese degli anni osservati mille campioni con ripetizione (*B*=1000) dai quali è stato ottenuto il vettore (2) fissando il livello di probabilità pari al 90% (α=0.10) e considerando di volta in volta gli insiemi di riferimento utilizzati nel calcolo dei singoli indicatori messi a confronto; così, ad esempio, per l'intervallo di confidenza dell'indice ufficiale è stato utilizzato il *set* costituito dagli indici elementari non esclusi dalla procedura di *trimming* adattativo, per l'indice troncato con estremi fissi (tipo Eurostat) l'insieme degli indici elementari non esclusi dall'intervallo di troncamento (0,5;2) ecc..

Per ciascun mese dunque è possibile produrre un insieme di indicatori caratteristici della distribuzione delle stime *bootstrap,* con riferimento ai singoli indici messi a confronto, attraverso cui desumere elementi utili alla valutazione dell'accuratezza della stima. Tra i parametri caratteristici della distribuzione delle stime *bootstrap* uno dei più utili alla valutazione dell'accuratezza della stima è l'Errore Relativo medio[22] che per ogni mese *m* e per ciascun tipo di stima *s (s=1….S)* tra quelle messe a confronto, è definito come segue:

$$_{s}EM_{m} = \sum_{i \in B} \left(\frac{\left| {}_{s}\hat{\theta}_i^{*} - {}_{s}\hat{\theta} \right|}{\sum_{s=1}^{S} \left| {}_{s}\theta_i^{*} - {}_{s}\hat{\theta} \right| / S} \right) / B \tag{4}$$

dove *B* è il numero di campioni *bootstrap* e il numeratore rappresenta la *deviazione assoluta* tra la *s*-ma stima *bootstrap* $_{s}\hat{\theta}_i^{*}$ e la stima calcolata sulla base del metodo di calcolo in esame $_{s}\hat{\theta}$. Se l'errore relativo medio è inferiore a 1 ed è inferiore a quello degli indici con cui è confrontato è possibile definire la stima *s*-ma, a parità di altri parametri caratteristici, come relativamente più accurata rispetto a quelle con cui è confrontata. L'indicatore (4) è una misura standardizzata della variabilità delle stime $\hat{\theta}_i^{*}$ ed è quello che fornisce un'informazione immediata sull'efficienza relativa della stima empirica; a parità di ampiezza dell'intervallo di confidenza, un minor errore relativo segnala il fatto che, in probabilità, la stima empirica cui esso corrisponde è più accurata.

La tavola 12 riporta a scopo illustrativo i risultati, riferiti a gennaio 2007, per la divisione ATECO "Articoli in pelle (escluso abbigliamento) e simili". L'indicatore

[22] Si veda S.M. Stigler (1977) per approfondimenti. Si noti che nei casi in cui non si hanno informazioni a priori sul valore del parametro, la formula (4) è valida per campioni molto grandi (in cui si può far valere il criterio di convergenza debole della stima campionaria al valore vero).

pubblicato (trimming adattativo) è messo a confronto con l'indice "tipo Eurostat" (trimming fisso), con l'indice ottenuto utilizzando tutti gli indici elementari (cioè senza applicare metodi di troncamento) e con l'indice ottenuto considerando la mediana della distribuzione originale pesata (che rappresenta il metodo di trimming più radicale). La tavola evidenzia una situazione abbastanza frequente: l'indice ufficiale risulta inferiore all'indice di tipo Eurostat nel livello e vanta una variabilità complessiva delle stime *bootstrap* ad esso riferite più contenuta. Allo stesso tempo l'intervallo di confidenza stimato, a parità di α, è più stretto di quello relativo agli altri metodi anche se la distribuzione delle stime mostra un indice di curtosi leggermente maggiore e una perdita di informazione determinata dal *trimming adattativo* maggiore.

Tavola 12 - Risultati del *bootstrap* per tipo di indice. Esportazioni totali di Articoli in pelle (escluso abbigliamento) e simili - Gennaio 2007 (*bootstrap effettuato sugli indici a base mobile*)

INDICATORI CARATTERISTICI	Indice pubblicato	Media *trimmed* (tipo Eurostat)	Media ponderata (totale indici elementari)	Mediana "pesata" (a)
Media originale	114,651	115,995	155,715	114,406
Parametri caratteristici relativi alle stime bootstrap (θ)*				
Media	114,587	115,944	154,723	114,605
CV corretto (b)	0,0125	0,0129	0,0763	0,0177
Varianza	0,0002	0,0002	0,0140	0,0004
Bias	0,0006	0,0005	0,0099	-0,0020
Errore Quadratico Medio (EQM)	0,0143	0,0149	0,1188	0,0202
Asimmetria	0,0684	0,0530	0,5402	0,1896
Curtosi	3,3604	3,0855	3,3376	2,9903
Errore Relativo	0,3945	0,4238	2,5858	0,5959
Dev. Stand. Errore Relativo	0,3163	0,3367	0,9071	0,4642
Peso % in valore (c)	91,7	94,2	100,0	…
Intervalli di confidenza				
Metodo dei percentili (α=0.10)				
estremo inferiore θ* [5%]	112,267	113,514	136,697	111,425
estremo superiore θ* [95%]	116,954	118,357	175,379	117,761
Metodo standard normal approssimativo (d)				
estremo inferiore θ* [5%]	112,239	113,490	135,243	111,300
estremo superiore θ* [95%]	116,935	118,399	174,204	117,911

Fonte:Istat, Statistiche del commercio con l'estero
(a) Mediana calcolata sulla distribuzione originale pesata.
(b) Rapporto tra EQM e valore originale della stima.
(c) Peso degli indici elementari utilizzati nell'aggregazione.
(d) Gli estremi dell'intervallo sono calcolati ipotizzando l'uso di una normale standardizzata.

L'errore relativo è, tuttavia, più contenuto e dunque vi sono elementi sufficienti per ritenere che la stima pubblicata sia relativamente più accurata. Ovviamente, è possibile che in alcuni mesi il metodo di *trimming adattativo* fornisca risultati meno efficienti di quelli forniti, ad esempio, dal metodo di *trimming* fisso. Per tali ragioni, la valutazione complessiva viene fatta anche sui profili annuali della serie.

La tavola 13, ad esempio, riporta gli intervalli di confidenza in media annua per i soli indici ufficiali Istat e per gli indici di "tipo Eurostat", entrambi concatenati, sempre con riferimento ai prodotti dell'ATECO "Articoli in pelle e simili".

Tavola 13 - Intervalli di confidenza dell'indice dei VMU all'esportazione riferito all'Ateco "Articoli in pelle e simili" - Anni 2005-2008 (*indici concatenati in media annua*)

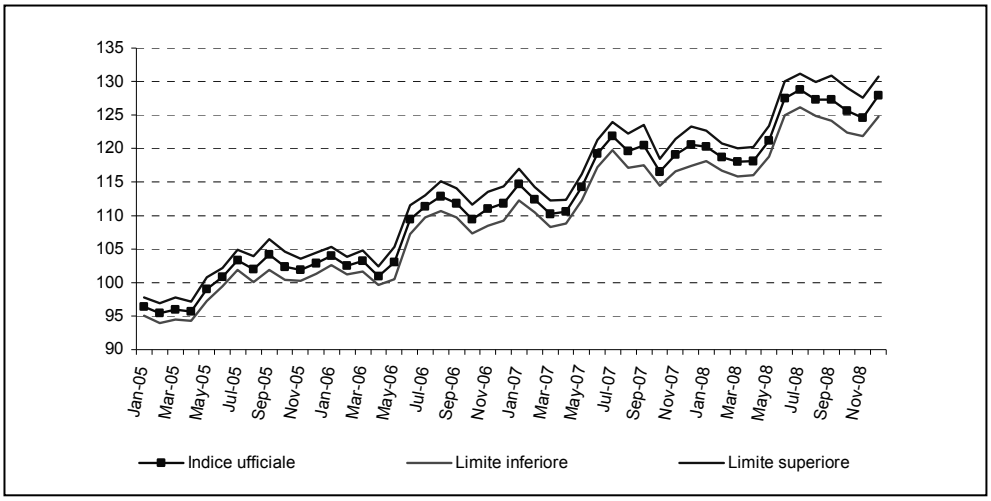
ANNI	Indice pubblicato				Media *trimmed* (tipo Eurostat)			
	Indice	Intervallo di confidenza		Errore relativo (a)	Indice	Intervallo di confidenza		Errore relativo (a)
		Limite inferiore	Limite superiore			Limite inferiore	Limite superiore	
2005	100,0	98,4	101,7	0,524	100,0	98,2	101,8	0,554
2006	107,6	105,7	109,6	0,580	107,8	105,8	109,8	0,582
2007	116,6	114,3	118,8	0,531	117,1	114,7	119,3	0,550
2008	123,8	121,2	126,3	0,416	124,4	121,9	127,0	0,414

Fonte: Istat, Statistiche del commercio con l'estero
(a) Media annua.

Come si nota, l'ampiezza degli intervalli di confidenza della stima *trimmed* pubblicata sono, in media d'anno, quasi sempre più stretti dell'omologa stima effettuata con troncamento fisso e anche l'errore relativo risulta inferiore, ad eccezione dell'anno 2008 quando è la stima basata sul *trimming* fisso a risultare, seppur di poco, più accurata.

Il grafico 2, infine, visualizza l'intervallo di confidenza con riferimento alla serie mensile relativa agli anni 2005-2009 del medesimo gruppo di prodotti.

Grafico 2 - Intervalli di confidenza dell'indice dei VMU all'esportazione riferito all'Ateco "Articoli in pelle e simili" - Anni 2005-2009 (*indici mensili concatenati*)



Fonte: Elaborazione su dati Istat

Quanto riportato nelle tavole 12 e 13 si riscontra nella maggior parte delle sezioni ATECO di cui si pubblicano gli indici VMUX.

Per dare un'idea di quanto si è riscontrato negli altri gruppi di prodotto, si riporta (tavola 14) il valore medio annuo dell'Errore Relativo (4) risultante dal *bootstrap* con riferimento alle divisioni della Manifattura dell'ATECO2007 nell'anno 2007.

Nella maggioranza dei casi alla stima ufficiale corrisponde un errore relativo medio annuo inferiore rispetto alle stime ottenute con gli altri metodi. In particolare, la stima ottenuta con il *trimming* di tipo Eurostat risulta relativamente più accurata in soli cinque gruppi di prodotto anche se lo *spread* con l'errore riferito alla stima ufficiale è molto limitato. In un solo caso, invece, una maggiore accuratezza relativa si rileva per la mediana pesata. Nel caso specifico delle "Apparecchiature elettriche e apparecchiature per uso domestico non elettriche" invece vi è una situazione più incerta in cui la stima ufficiale, la stima di tipo Eurostat e la stima ottenuta con la mediana pesata potrebbero essere equivalenti sotto il profilo dell'accuratezza, a parità di ampiezza dell'intervallo di confidenza.

Tavola 14 - Errore relativo delle stime bootstrap per ATECO e tipo di indice. Esportazioni totali - Anno 2007 (*bootstrap effettuato sugli indici a base mobile*)

GRUPPI DI PRODOTTO	Indice pubblicato	Media *trimmed* (tipo Eurostat)	Media ponderata (totale indici elementari)	Mediana "ponderata" (a)
Prodotti alimentari	0,5282	0,6272	2,2072	0,6374
Bevande	0,7072	0,8176	1,5184	0,9568
Tabacco	0,7402	0,7351	1,3895	1,1352
Prodotti tessili	0,2142	0,2385	3,2760	0,2713
Articoli di abbigliamento (anche in pelle e in pelliccia)	0,7236	0,7125	1,5245	1,0394
Articoli in pelle (escluso abbigliamento) e simili	0,4937	0,5052	2,2813	0,7197
Legno e prodotti in legno e sughero (esclusi i mobili); articoli in paglia e materiali da intreccio	0,6595	0,5340	2,3386	0,4679
Carta e di prodotti di carta	0,3704	0,4676	2,6935	0,4685
Prodotti chimici	0,2521	0,2951	3,1420	0,3108
Articoli in gomma e materie plastiche	0,3412	0,4975	2,7124	0,4489
Altri prodotti della lavorazione di minerali non metalliferi	0,1911	0,2137	3,3335	0,2616
Prodotti della metallurgia	0,3841	0,3815	2,7652	0,4692
Prodotti in metallo, esclusi macchinari e attrezzature	0,4067	0,4630	2,6637	0,4666
Apparecchiature elettriche e apparecchiature per uso domestico non elettriche	0,2626	0,2618	3,2133	0,2623
Macchinari e apparecchiature n.c.a.	0,4801	0,3472	2,7995	0,3732
Autoveicoli, rimorchi e semirimorchi	0,7294	0,8651	1,5469	0,8585
Altri mezzi di trasporto	0,6584	0,7221	1,7444	0,8751
Mobili	0,7732	0,8955	1,2711	1,0602
Altri prodotti manifatturieri	0,3209	0,3089	2,8735	0,4967

Fonte:Elaborazioni su dati Istat
(a) Media calcolata sulla distribuzione originale "pesata".

È rilevante osservare la distanza in termini di accuratezza tra le stime ottenute con metodi di *trimming,* da un lato, e la stima ottenuta semplicemente aggregando gli indici elementari, dall'altro lato. La differenza tra i due metodi di aggregazione, al di là del tipo di media troncata utilizzato, è evidentemente nel diverso contenuto informativo che essi esprimono: il ricorso a medie troncate, infatti, risponde principalmente all'esigenza di individuare un segnale di fondo (*core*) dell'indicatore scevro da elementi spuri e di difficile controllo.

5. Strategie alternative di aggregazione: il calcolo degli indici dei VMU con imputazione delle osservazioni identificate come anomale e l'uso di medie Winsorized

La metodologia di calcolo degli indici dei VMU prevede nella fase iniziale l'identificazione di possibili errori di misura sulle singole transazioni e la loro correzione attraverso l'esclusione dei record riconosciuti come "errati" dal calcolo dei livelli dei VMU negli strati di appartenenza, in modo che i numeri indici elementari, ottenuti dall'aggregazione delle singole transazioni, non ne siano condizionati.

L'identificazione dei valori anomali viene realizzata sulla base di un algoritmo per il trattamento di distribuzioni asimmetriche e non parametriche che, nel caso specifico, utilizza osservazioni non pesate sulla base dell'evidenza empirica secondo cui l'errore di misura non dipende dall'entità della transazione cui esso è associato.

La scelta di escludere completamente le osservazioni "errate" dipende essenzialmente dall'impossibilità di verificare, anche in fase di revisione dei dati grezzi e comunque prima del calcolo del VMU elementare, la reale fonte dell'errore che, quindi, potrebbe essere costituita o da una incorretta dichiarazione del valore, o da errori sulla quantità o da errori su entrambe le variabili nonché da errori legati ad errata attribuzione del codice NC.

Da un punto di vista tecnico la scelta di eliminare le transazioni affette da anomalie equivale alla decisione di non effettuare imputazioni dei VMU "errati", evitando così di alterare la distribuzione originale della variabile.[23]

In fase di aggregazione, invece, l'eliminazione dal calcolo degli indici elementari posti al di fuori dell'intervallo di *trimming* è stata decisa in funzione dell'obiettivo dichiarato di pervenire ad una stima che possa considerarsi un *core index,* cioè un valore il più possibile rappresentativo del comportamento "di fondo" degli operatori al netto di elementi di "disturbo" aleatori e non controllabili. Tale obiettivo coincide, tecnicamente, con l'esigenza di individuare stimatori più robusti della media aritmetica che, notoriamente, non è adeguata a sintetizzare risultati provenienti da distribuzioni affette dalla presenza di outlier.

Per verificare quale sarebbe stato l'impatto di strategie alternative all'esclusione di tali outlier l'indice ufficiale è stato confrontato con:

1) un indice calcolato "imputando" il VMU delle transazioni considerate errate in fase di calcolo degli indici elementari di strato attraverso il metodo del "donatore medio", ma effettuando l'aggregazione secondo il metodo *adaptive trimming* usato per l'indice ufficiale;

2) un indice in cui oltre all'imputazione delle transazioni elementari errate viene utilizzata una media troncata di tipo Winsorized secondo la quale, di fatto, si imputa un valore per gli indici elementari che cadono fuori dall'intervallo di accettazione e se ne conserva il peso originario. Nello specifico, la media Winsorized è un tipo di media troncata che, nell'ottica di preservare almeno parzialmente le informazioni scartate, imputa a queste l'indice elementare riscontrato nei punti di troncamento, mantenendone però il peso originale. In altre parole, alle informazioni scartate appartenenti alla coda sinistra della distribuzione della variabile si attribuisce l'indice elementare corrispondente al punto di troncamento sinistro della distribuzione, mentre alle informazioni scartate appartenenti alla coda destra della distribuzione si attribuisce l'indice elementare corrispondente al punto di troncamento destro della distribuzione.

[23] E' noto che alcuni metodi di imputazione distorcono la distribuzione *ex post* della variabile (ad esempio l'imputazione attraverso un donatore medio). In una situazione caratterizzata da distribuzioni già fortemente asimmetriche come nel caso dei VMU, la scelta di imputare i record affetti da errore di misura è stata giudicata inopportuna.

La tavola 15 riassume la differenza tra la serie degli indici dei VMU all'esportazione e all'importazione calcolati con la media troncata correntemente pubblicata dall'Istat e le analoghe serie calcolate utilizzando i metodi previsti ai precedenti punti 1 (medie *Trimmed* con imputazione) e 2 (medie *Winsorized* con imputazione).

Come atteso, il livello dell'indice "Media *trimmed* con imputazione" è inferiore al livello dell'indice pubblicato, mentre il livello dell'indice *Winsorized* risulta sempre più elevato degli altri due, nonostante sia effettuata l'imputazione delle osservazioni elementari "errate".

Nel primo caso vi è un effetto attribuibile alla concentrazione delle osservazioni della distribuzione *ex post* nel punto corrispondente al valore centrale della distribuzione dei record non "errati", operazione che notoriamente riduce la variabilità complessiva. Nel secondo caso l'effetto dominante è dato dall'imputazione degli indici elementari che cadono oltre i punti di troncamento dell'intervallo di *trimming;* in particolare, poiché le distribuzioni dei VMUX sono generalmente asimmetriche a destra, il peso maggiore è associato ad *outlier* imputati usando come donatore l'estremo superiore dell'intervallo di troncamento e ciò induce un effetto di innalzamento dell'indice aggregato.

Tavola 15 - Indici dei VMU all'esportazione e all'importazione secondo il metodo di calcolo (base 2005=100). Totale prodotti - Anni 2005-2009 (*indici e variazioni % sul corrispondente periodo dell'anno precedente*)

ANNI	Media *trimmed* senza imputazione (a)		Media *trimmed* con imputazione		Media *winsorized* con imputazione	
	Esportazioni	Importazioni	Esportazioni	Importazioni	Esportazioni	Importazioni
2005	100,0	100,0	100,0	100,0	100,0	100,0
2006	105,1	109,5	104,7	108,6	106,7	110,1
2007	110,5	112,8	109,3	111,3	113,4	114,3
2008 (b)	116,7	123,0	114,4	120,3	120,8	123,7
I sem. 2009	116,0	112,8	113,2	110,1	121,3	114,0
Var. 2006/2005	*5,1*	*9,5*	*4,7*	*8,6*	*6,7*	*10,1*
Var. 2007/2006	*5,1*	*3,0*	*4,3*	*2,5*	*6,2*	*3,7*
Var. 2008/2007	*5,7*	*9,0*	*4,7*	*8,0*	*6,6*	*8,3*
Var. I sem09 / I sem08	*0,4*	*-6,6*	*0,0*	*-6,9*	*1,6*	*-6,5*

Fonte: Istat, Statistiche del commercio con l'estero
(a) Indice pubblicato.
(b) Dati provvisori.

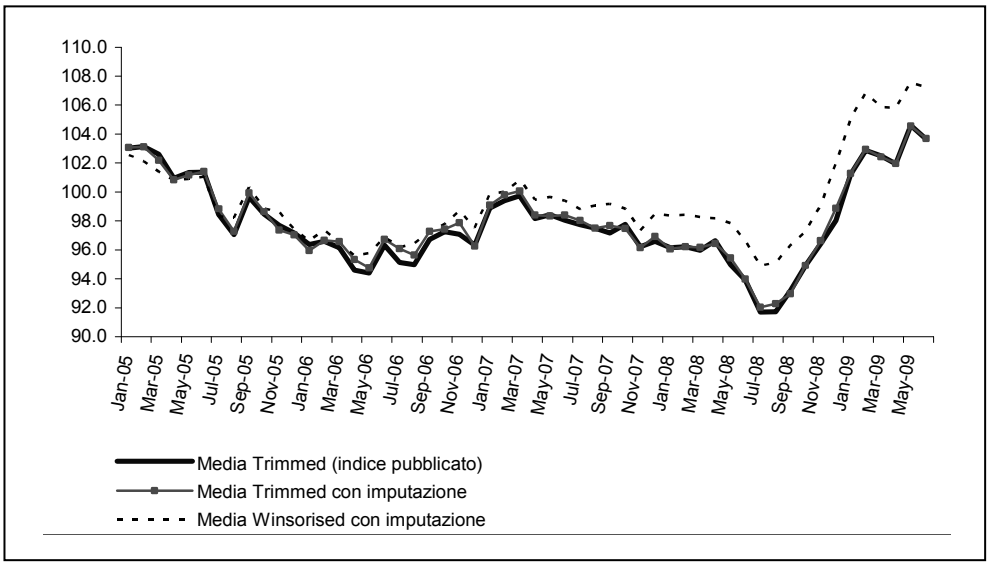
Per una migliore valutazione degli effetti complessivi indotti dai due metodi di imputazione testati vale la pena visualizzare i risultati in termini di ragioni di scambio, determinate dal rapporto tra VMU all'esportazione e VMU all'importazione[24] (grafico 3). Appare evidente che le differenze tra l'indice pubblicato e l'indice *trimmed* con imputazione sono minime e, dove pure sussistono, le ragioni di scambio relative all'indice pubblicato risultano inferiori, denotando la presenza di un minore *spread* tra VMU

[24] La ragione di scambio è un indicatore che misura la competitività delle esportazioni di un paese rispetto alle importazioni. Poiché nel calcolo dei VMU il metodo di imputazione prescelto viene applicato ad entrambi i flussi, se il profilo della ragione di scambio non si modifica al variare del metodo si può ragionevolmente concludere che l'effetto complessivo sulla rappresentatività degli indici finali è scarsamente significativo.

all'esportazione e all'importazione. Ulteriori analisi sugli indici in base mobile (quindi al netto di eventuali effetti indotti dal concatenamento) hanno messo in evidenza come le differenze maggiori tra i due indici siano riscontrabili nella serie delle importazioni.

La media *Winsorize*d, al contrario, evidenzia livelli dei VMU all'esportazione che nel tempo crescono più dei livelli dei VMU all'importazione alterando l'interpretazione economica del fenomeno.

Grafico 3 - Andamento della ragione di scambio secondo differenti metodi di calcolo dei VMU. Totale prodotti - Anni 2005-2009



Fonte: Elaborazione su dati Istat

Conclusioni

La diffusione dei nuovi indici del commercio estero e l'avvio della pubblicazione degli indici dei prezzi dei prodotti industriali sui mercati esteri ha, da un lato, ampliato il *set* di indicatori per l'analisi della competitività del sistema produttivo italiano e dei processi di internazionalizzazione, e dall'altro stimolato ulteriormente il dibattito sull'adeguatezza delle misurazioni prodotte dalla statistica ufficiale.

Quanto riportato nel presente lavoro rappresenta un contributo alla valutazione della sensibilità degli indici dei valori medi unitari ufficiali all'utilizzo di diverse opzioni metodologiche. In particolare, i risultati delle analisi consentono di apprezzare:

- la robustezza degli indici dei valori medi unitari in presenza di opzioni metodologiche caratterizzate da sostanziali alterazioni del *coverage* di riferimento, in particolare dalla limitazione del calcolo degli indici a sottoinsiemi persistenti di operatori e prodotti;
- l'efficienza degli indici pubblicati, misurata attraverso la definizione di intervalli di confidenza delle stime mensili, messe a confronto con omologhi indici calcolati utilizzando metodologie di aggregazione alternative;

- la solidità del metodo di calcolo degli indici pubblicati, attraverso la comparazione con analoghi indici calcolati utilizzando tecniche di imputazione dei record affetti da "errori di misura", applicate a livello sia di singole transazioni sia di indici elementari utilizzati nelle aggregazioni successive.

A conclusione del lavoro vale la pena sottolineare come, a nostro avviso, gran parte del dibattito sulla adeguatezza delle misurazioni legate agli indici dei VMU venga alimentato essenzialmente da due elementi: da un lato, la mancanza di un *frame* metodologico riconosciuto e condiviso al livello internazionale, come quello che caratterizza le rilevazioni campionarie sui prezzi, genera incertezze sulla solidità di qualunque soluzione metodologica adottata. In aggiunta, l'enorme disponibilità di informazioni elementari di fonte amministrativa da trattare mensilmente complica la definizione statistica dell'universo di riferimento entro cui definire le stime.

Dall'altro lato, l'equivoco parzialmente indotto dalla manualistica internazionale - secondo cui gli indici dei VMU vanno intesi come delle *proxy* dei prezzi all'esportazione e all'importazione ai fini della deflazione degli aggregati macroeconomici - ha impedito finora di immaginare un quadro più ampio in cui entrambi gli indicatori (prezzi e VMU) possano convivere in modo integrato, in una prospettiva di costruzione di un "sistema" di indicatori di competitività internazionale.

Con riferimento al primo aspetto, esistono senza dubbio degli ampi margini di affinamento delle tecniche di stima che meritano di essere analizzati e sperimentati; in ciò la mancanza di riferimenti metodologici internazionali pare più uno stimolo per la ricerca che un limite. La statistica ufficiale, a nostro avviso, è sicuramente in grado di fornire il suo contributo.

Con riferimento al secondo aspetto, le potenzialità della base informativa e il crescente orientamento verso un maggiore utilizzo delle fonti amministrative non sembrano affatto in contraddizione con la costruzione di un sistema di indicatori che oltre a cogliere diversi elementi informativi, dall'andamento "lordo" dei VMU ufficiali via via fino ad indici in grado di isolare le componenti più propriamente di prezzo, possa in futuro riferirsi a specifiche partizioni dell'universo di riferimento (indici per dimensione delle imprese esportatrici o importatrici; indici per aggregazioni di prodotti distinti in base all'intensità tecnologica, ecc.) in un contesto prettamente microeconomico, svincolato da finalità legate alla deflazione degli aggregati di Contabilità Nazionale.

Riferimenti bibliografici

ANITORI P., CAUSO M.S. (2008) *"La metodologia di calcolo dei nuovi indici dei valori medi unitari del commercio con l'estero"*, Atti del Convegno "L'informazione statistica ufficiale per l'analisi economica dell'internazionalizzazione delle imprese". Istat, Roma 12 giugno 2008.
http://www.istat.it/istat/eventi/2008/internazionalizzazione_imprese/relazioni/anitori_causo.pdf.

EFRON B. e TIBSHIRANI, ROBERT J. (1993). *An introduction to the bootstrap*. Chapman & Hall , 168-177.

EUROSTAT (2006) *"Methodology of short-term business statistics- Interpretation and guidelines"* Eurostat, Luxemboug.

EUROSTAT (2005) *"EC Regulation n. 1158/2005 modifing EC reg. 1165/98 on* Short-term Statistics", O.J serie L191/1 del 22/7/2005, Luxembourg.

EUROSTAT (1996), *European system of accounts ESA 1995* Council Regulation 2223/96. http://circa.europa.eu/irc/dsis/nfaccount/info/data/esa95/en/titelen.htm

FMI (2009) *"Export and Import Price Index Manual: Theory and Practice"* http://www.imf.org/external/pubs/cat/longres.cfm?sk=19587.0. FMI, Washington.

FULLER W. A. (1991), "Simple Estimators for the Mean of Skewed Populations", Statistica Sinica 1,pp. 137-158,Taiwan.

IACOBACCI T., POLITI M. (2008) "*Gli indici dei prezzi alla produzione dei prodotti industriali venduti sul mercato estero (base 2000=100)"*. Atti del Convegno "L'informazione statistica ufficiale per l'analisi economica dell'internazionalizzazione delle imprese". Istat, Roma 12 giugno 2008.
http://www.istat.it/istat/eventi/2008/internazionalizzazione_imprese/relazioni/iacobacci.pdf

ISAE (2009) *"Rapporto ISAE:le previsioni per l'economia italiana- Luglio 2009"*. ISAE, Roma.

ISTAT (2003) *"I nuovi indici del commercio con l'estero (base 2000=100)"* Nota Informativa del 16 luglio 2003.
http://www.istat.it/salastampa/comunicati/non_calendario/ 20030716_01/

LUZI O. et al. (2007) *"Recommended Practices for Editing and Imputation in Cross-Sectional Business Surveys"*. EDIMBUS-RPM Project
http://epp.eurostat.ec.europa.eu/ portal/page/portal/quality/documents/RPM_EDIMBUS.pdf

RAMSEY PHILIP H., RAMSEY, PATRICIA P. (2007), *Optimal Trimming and Outlier Elimination.* Journal of Modern Applied Statistical Methods Vol.6, No. 2, 355-360

SPRENT P., SEETON N.P. (2001) *"Applied non parametric statistical method"*, 3rd edition Chapman & Hall.

STIGLER STEPHEN M. (1977) "*Do Robust Estimators Work with Real Data?"*. University of Wisconsin at Madison. The Annals of Statistics 1977, vol.5 No.6, 1055-1098

THOMPSON K. J. (1996), *Ratio Edit Tolerance Development using Variations of Exploratory Data Analysis (EDA) Resistans Fences Methods*. United States Bereau of the Census.

TUKEY J. W., MCLAUGHLIN D. H. (1963) *"Less Vulnerable Confidence and Significance Procedures for Location Based on a Single Sample: Trimming/Winsorization*" Indian Journal of Statistics,Sankhya Series A, Vol 25, No 3, p.331-352

UN (1983) *"Price and quantity measurement in external trade"* UN, New York.

VANDERVIERE E., HUBER M. (2004) *"An adjusted boxplot for skewed distribution*" in Compstat Symposium*,* Springer-Verlag, Berlin.

WILCOX RAND R. (2006) *"Robust estimation and hypothesis testing"* , Elsevier Academic Press, UK.

WILCOX RAND R. (2005) *"Trimmed means"*, in Encyclopaedia of Statistics in Behavioural Science. Wiley & Sons. Chichester, UK.

A Novel Suite of Methods for Mixture Based Record Linkage[1]

Diego Zardetto[2], Monica Scannapieco[3]

Abstract

Record Linkage (RL) aims at identifying pairs of records coming from different sources and representing the same real world object. Despite several methods have been proposed to face RL problems, none of them seems to be at the same time fully automated and very effective. In this paper we present a novel suite of methods that instead possesses both these abilities. We adopt a mixture-model based approach, which structures a RL process into two consecutive tasks. First, mixture parameters are estimated by fitting the model to observed distance measures between pairs. Then, a probabilistic clustering of the pairs into Matches and Unmatches is obtained by exploiting the fitted model. In particular, we use a mixture model with component densities belonging to the Beta parametric family and we fit it by means of an original perturbation-like technique. Moreover, we solve the clustering problem according to both Maximum Likelihood and Minimum Cost objectives. To accomplish this task, optimal decision rules fulfilling one-to-one matching constraints are searched by a purposefully designed evolutionary algorithm. We present several experiments on real data that validate our methods and show their excellent effectiveness.

1. Introduction

Record Linkage (RL) (Elmagarmid et al., 2007) is the problem of identifying pairs of records coming from different sources and representing the same real world object. Integration of different data sources and improvement of the quality of single sources are only some of the real application scenarios that need to solve the RL problem. In Official Statistics, to cite just a single example, the need of performing a RL task arises whenever one tries to integrate statistical survey data with data coming from administrative archives, due to lacking or unreliable common record identifiers.

In this paper we present a novel suite of methods for RL, based on mixture models (McLachlan et al., 2000; McLachlan et al., 1988). These are statistical models that allow to represent a probability distribution as a convex combination of other distributions. As RL methods always rely on distance measures between record pairs, the intuition behind the use of mixtures models is that these observed distances arise from a superposition of two distinct probability distributions: the one stemming from the subpopulation of Matches (*M*) and the other from that of Unmatches (*U*). Evidently, the ultimate aim of such a statistical perspective on RL is to exploit the mixture model for classification purposes, i.e., to bring to light the hidden grouping of the pairs in the underlying *M* and *U* classes.

[1] This work is the outcome of a research collaboration and reflects opinions of both authors, however the primary contribution to the work has to be attributed to Diego Zardetto. This work was partially supported by Sapienza Università di Roma, "Progetti di Ricerca di Università" C26A074R53-C26A08L953.

[2] Cter (Istat), e-mail: zardetto@istat.it.

[3] Tecnologo (Istat), e-mail: scannapi@istat.it.

Since we formulate the RL problem as a classification problem driven by a mixture model, our approach requires the execution of two consecutive tasks. First, mixture parameters have to be estimated by fitting the model to the observed distance measures between pairs. Then, a probabilistic clustering of the pairs into Matches and Unmatches must be obtained by exploiting the fitted model.

The fitting step is the crucial one, as it implicitly determines the quality of the subsequent clustering results. However, it represents a very hard task; indeed, the problem of fitting a mixture model is always difficult, but it is even more severe in RL applications. This is due to the huge class-skew inherent in RL problems, where the very few (and unidentified) distance measures stemming from Matches risk to be completely overwhelmed by the bulk of those stemming from Unmatches. To overcome this difficulty we developed an original fitting technique inspired by Perturbation Theory (Bender et al., 1999). This technique allows us to obtain reliable estimates for the mixture parameters without relying on domain knowledge, thus not jeopardizing automation.

In the clustering step we use the fitted mixture model to search an optimal classification rule such that each pair can be assigned, based on its observed distance value, either to the *M* or to the *U* class. This is accomplished in such a way as to optimize a global objective function while satisfying a set of one-to-one matching constraints. These constraints arise when the data sets to be matched do not contain duplicates. In particular, we solve this constrained optimization problem by means of a purposefully designed evolutionary algorithm (Michalewicz, 1996). We use the algorithm to find decision rules that minimize either the probability of classification error (Maximum Likelihood objective) or, alternatively, the expected classification cost (Minimum Cost objective). The resulting rules critically depend on posterior estimates of class membership probabilities, obviously derived from the fitted mixture model.

The RL problem has received considerable attention by the scientific community, see (Elmagarmid et al., 2007) for a survey. Some works, e.g. (Chaudhuri et al., 2005; Guha et al., 2004), focus on solving the problem within a relational DBMS. Our approach is different from the cited ones because it is focused on effectiveness rather than on the ability to manage huge amount of data. The focus on effectiveness allows us to obtain results that are indeed superior to those obtained by (Chaudhuri et al., 2005) (whereas it was possible to compare the obtained results). However, our approach can be usefully inserted as a "decision engine" into a general RL system, even oriented to large databases.[4] Indeed this would just require to undertake a preliminary step dedicated to the reduction of the comparison space, for which several techniques have been proposed (Elmagarmid et al., 2007). This is because we designed our methods to be as general as possible; for instance, we do not rely on any restrictive assumption on the function to be used when comparing records.

Moreover, we remark the completely automated nature of our approach. This makes our work different on the one hand from supervised techniques for RL, e.g. (Tejada et al., 2001). On the other hand, our proposal is also different from the few unsupervised techniques, including (Chaudhuri et al., 2005; Verykios et al., 2000; Christen, 2007), none

[4] The practical feasibility of our methods when dealing with very large amounts of data will be tested in the Experiment section of the paper. There we shall face a big RL problem involving data collected in the Post Enumeration Survey carried out by the Italian National Institute of Statistics to estimate the coverage rate of the 2001 population Census.

of which, to the best of our knowledge, is fully automated. Indeed, though not requiring exactly a clerically prepared training set, such techniques still depend critically on some external inputs: e.g., human intervention is needed to set crucial parameters for the algorithms in (Chaudhuri et al., 2005) and (Christen, 2007), or to provide few labeled data in (Verykios et al., 2000).

Several works on RL rely on a probabilistic approach. A comparison with such works, mostly based on the Fellegi-Sunter formulation of the problem (Fellegi et al., 1969), will be presented later on in the paper (see Section 4), when we shall discuss the details of our proposal.

The paper is organized as follows. In Section 2 basic assumptions underlying statistical approaches to RL are introduced. These assumptions are then distilled in the form of a loose prior knowledge that our method is able to exploit successfully when facing practical RL tasks. Section 3 defines the adopted mixture model, whose component densities belong to the Beta parametric family. Section 4 is devoted to a thorough motivation and description of our original mixture-model fitting technique. Section 5 faces the clustering step, deals with one-to-one matching constraints and describes our clustering evolutionary algorithm. In Section 6 we test the effectiveness of our suite of methods on real-world RL instances. Finally, Section 7 draws some conclusions.

2. A Statistical Perspective on Record Linkage

Let us consider two sets $\mathcal{A}$, $\mathcal{B}$ of real world objects selected from a universe Ω and let us suppose that $\mathcal{A}$ and $\mathcal{B}$ contain some common objects, i.e. $\mathcal{A} \cap \mathcal{B} \neq \varnothing$. We denote by $\mathcal{M}$ the set of matched objects that appear in both $\mathcal{A}$ and $\mathcal{B}$, $\mathcal{M} = \mathcal{A} \cap \mathcal{B}$, and by $\mathcal{U}$ the set of non-matched objects that appear in either $\mathcal{A}$ or $\mathcal{B}$ but not in both, $\mathcal{U} = (\mathcal{A} \cup \mathcal{B}) \setminus \mathcal{M}$. Obviously $\mathcal{A} \cup \mathcal{B} = \mathcal{M} \cup \mathcal{U}$.

We formally represent a *deterministic* data generating process as a mapping g from Ω to a data space S, such that $\Omega \ni \omega \overset{g}{\mapsto} g(\omega) = o \in S$. The image of $\mathcal{A}$ ($\mathcal{B}$) under g is a subset of S: we call it data set A (B). Notice that representing g as a mapping implicitly rules out the possibility that data sets A and B contain duplicated records. For *duplicates* we mean records that *i)* belong to the same data set and *ii)* correspond to the same real world object. We shall denote a real world object with a greek letter, $\alpha \in \mathcal{A}$, and the record to which it is mapped by g with the corresponding latin letter, $g(\alpha) = a \in A$.

The Record Linkage problem is defined as follows. Given two data sets A and B, find a partition of their cartesian product such that:

$$A \times B = M \cup U \tag{1}$$

where we introduced the set of record pairs that are Matches:

$$M = \{(a,b) \in A \times B : \alpha = \beta, \alpha \in \mathcal{A}, \beta \in \mathcal{B}\} \tag{2}$$

and the one of record pairs that are Unmatches:

$$U = \{(a,b) \in A \times B : \alpha \neq \beta, \alpha \in \mathcal{A}, \beta \in \mathcal{B}\} \tag{3}$$

Consider now an *ideal* (i.e. deterministic and error free) data generating process g_0, that is any *injective* mapping of Ω to *S*. If data sets *A* and *B* were generated under g_0, the RL problem would be trivial. Indeed, since $g_0(\alpha) = g_0(\beta) \Rightarrow \alpha = \beta$, the Matches set would be simply:

$$M = \{(a,b) \in A \times B : a = b\} \tag{4}$$

In other words, under an ideal data generating process the information stored into a record is *sufficient* to unambiguously identify the corresponding real world object.

Unfortunately, real world data generating processes are always affected by a wide variety of errors. Being the underlying error mechanism unknown (and hence the generated errors unpredictable), every real data generating process *g* has to be thought as a *stochastic* process. It can be useful to describe such a *g* as the addition of a random noise (representing the errors) to a signal (representing the records that would have been generated under ideal conditions): $g(\alpha) = \mathbf{n}(g_0(\alpha))$. Here the noise **n** is treated as a function of a deterministic argument (i.e. an input record $o \in S$) whose value is a "random variable" (i.e. an output random record $\tilde{o} = \mathbf{n}(o)$ still belonging to the data space *S*). We do not exclude that the real world data sets *A* and *B* have been generated by two distinct data generating processes, rather we only assume that the difference (if any) is entirely due to the error generating mechanism: $A = \mathbf{n}_A(g_0(\mathcal{A}))$, $B = \mathbf{n}_B(g_0(\mathcal{B}))$.

From a RL point of view, the main effect of the stochastic nature of a data generating process is that, with some nonzero unknown probabilities: *i)* the *same* real world object, $\mu \in \mathcal{M}$, can correspond to two *distinct* records, $m_A \neq m_B$ with $m_A \in A$ and $m_B \in B$; *ii)* two *distinct* real world objects, $\upsilon_1 \in \mathcal{U}, \upsilon_2 \in \mathcal{U}$, can correspond to the *same* record, $u \in A, u \in B$. Due to *i)* and *ii)* the set identity (4) is in general not true anymore and the RL problem becomes non-trivial.

Since simply assessing whether two records are *equal* is no longer sufficient to unambiguously classify the pair as a Match or as an Unmatch, let us introduce a "distance" function $d : S \times S \to \Re^+$. We do not require *d* to fulfill the triangle inequality, thus *S* (endowed with *d*) does not need to be a metric space. Instead, we do believe that *d* is able to capture reasonably well the "amount of difference" between two records (whatever their structure). Consider now a record pair (*a*, *b*) belonging to $A \times B$: since *a* and *b* are regarded as (realizations of) random variables, this will also hold true for their distance *d*(*a*, *b*). For the sake of simplicity (but without loss of generality as the distance is a bounded variable) we shall suppose *d* to be normalized so that its values fall inside the [0,1] interval.

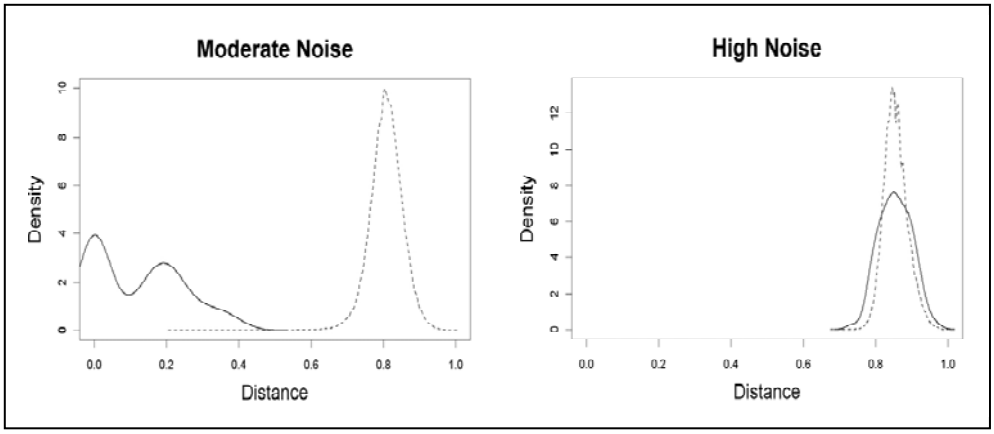
The basic idea behind every statistical approach to the RL problem is simple: we do believe that the observed distances between record pairs (although a priori unpredictable) carry some useful information about whether a given pair belongs to the set of Matches or to the one of Unmatches. Hence the distance *d* is viewed as an observable *auxiliary* random variable that we can use to infer the unknown outcomes of an (artificial) unobservable *interest* random variable *z*, namely the class membership indicator of a pair:

$$z_i = \begin{cases} 1 & \textit{if } i\textit{-th pair} \in M \\ 0 & \textit{if } i\textit{-th pair} \in U \end{cases} \tag{5}$$

Obviously this picture is founded on the hypothesis that the probability distribution of the distance *d* is significantly different inside the *M* and *U* classes. This in turn translates into the assumption that the noise component of the data generating process is only a small perturbation to the ideal signal.

Consider as an example Figure 1 where superimposed distance density plots are showed for Match (solid lines) and Unmatch (dashed lines) pairs. Both panels refer to the same original clean data sets to which we added random errors at moderate (left panel) and very high (right panel) rates.

Figure 1 - Distance density plots at moderate (left) and very high (right) error rates. Each graph shows superimposed density plots for Match (solid blue line) and Unmatch (dashed red line) pairs. [Physics data sets, see Section 6]

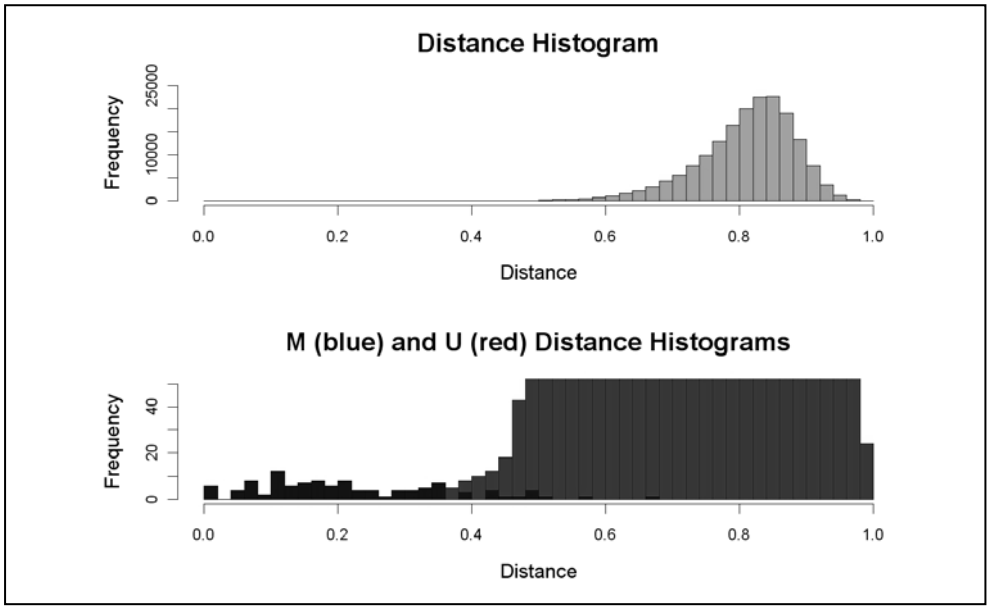


In the "moderate noise" scenario the shapes of the *M* and *U* distance densities are very different: Unmatches tend to be concentrated at higher distances than Matches, which furthermore still exhibit their own distinctive modal peak at zero distance; moreover *M* and *U* densities show only a relatively small overlap. On the contrary in the "high noise" scenario not only both Matches and Unmatches are located in the high-distance region, but in addition their densities almost completely overlap. While we are confident that a statistical approach to the RL problem will reveal itself appropriate to the first scenario, its use seems hopeless in the second. Luckily experience teaches two important lessons: *i)* the "high noise" scenario is almost never met in real-life applications, *ii)* the qualitative features of the *M* and *U* densities we just illustrated for the "moderate noise" scenario are quite general i.e. common to the great majority of practical problems. As a consequence we feel allowed to consider these main features as a *prior knowledge* about the underlying, unknown *M* and *U* distance probability distributions.

Besides this, it seems that another piece of *prior knowledge* is readily available, namely that Matches are rare. Indeed, if data sets *A* and *B* do not contain duplicated records (as we already assumed) the Match rate, i.e. the ratio between the cardinalities of *M* and $A \times B$, cannot exceed the value $1/\max(|A|,|B|)$. This value is very small in almost all the RL problems, even when blocking techniques have been applied.

Therefore, two distinct kinds of prior knowledge – named in the following $\mathcal{PK}1$ and $\mathcal{PK}2$ – are at our disposal: the first one, $\mathcal{PK}1$, concerning the main qualitative features of the *M* and *U* distance probability distributions, the second, $\mathcal{PK}2$, concerning the large class-skew between *M* and *U* pairs, with Matches being rare as compared to Unmatches. Figure 2 exemplifies in a clear-cut way the excellent agreement between the aforementioned basic assumpions $\mathcal{PK}1$ and $\mathcal{PK}2$ and sample data coming from a real-world RL instance, namely "Restaurants" (see Section 6).

Figure 2 - Pairwise-distances coming from a real-world RL instance [Restaurants data sets, see Section 6]. Upper panel: distance histogram of the whole unlabeled data (176,423 pairs). Lower panel: superimposed distance histograms for Match pairs (blue, dark) and Unmatch pairs (red, light); note that a 500 times y-axis zoom was needed to detect the feeble signal arising from the few Matches (112 pairs)



The Restaurants data contain only 112 Matches out of 331×533 = 176,423 pairs, yielding a Match rate of 6.3×10^{-4}, in full compliance with $\mathcal{PK}2$. Therefore, the histogram plotted in the upper panel, which represents the distance distribution of the whole unlabeled data (i.e. pairs belonging to both *M* and *U* classes), turns out to be totally dominated by the overwhelming contribution arising from Unmatches. As a consequence, a 500 times zoom of the low frequency region of the plot was needed in order to detect the feeble *M* distribution, as reported in the lower panel of figure 2. Moreover, the specific features of the *M* and *U* distance probabilities distilled into $\mathcal{PK}2$ are clearly reflected into the observed histograms. Indeed, Matches are dominating at low distances (with a bump at d = 0 arising from *M* pairs that haven't been hit by errors) and exhibit a soft right tail. Unmatches, in turn, dominate the high-distance region and show a soft left tail. Furthermore, there is only a small overlap between *M* and *U* tails.

We shall see in Sections 3 and 4 how our approach successfully exploits both $\mathcal{PK}1$ and $\mathcal{PK}2$ when facing practical RL tasks.

3. A Mixture Model for Record Linkage

Let us denote by n_A and n_B the cardinalities of data sets A and B respectively and by n_M and n_U the (unknown) cardinalities of classes M and U. Furthermore, let us call n_P the cardinality of the set of pairs $A \times B$ and define, for later convenience, $n_{min} = \min(n_A, n_B)$ and $n_{max} = \max(n_A, n_B)$, so that $n_P = n_A \cdot n_B = n_{min} \cdot n_{max}$. We shall denote the observed distances between record pairs by d_i where $i = 1, \ldots, n_P$ and an arbitrary ordering of the pairs is assumed. As we sketched in Section 2, we shall treat values d_i as n_P independent and identically distributed (iid) realizations of a random variable[5] d with values in [0,1].

We represent the probability density function (pdf) of d by the following two-component mixture density (McLachlan et al., 2000; McLachlan et al., 1988):

$$f(d) = \pi_M f_M(d) + \pi_U f_U(d) \tag{6}$$

where the components $f_{M,U}$ are the distance densities for the classes M and U and the mixing weights $\pi_{M,U}$ give the proportions of the classes, $\pi_{M,U} = n_{M,U}/n_P$ (so that $0 \le \pi_{M,U} \le 1$ and $\pi_M + \pi_U = 1$).

In what follows we assume that a suitable description of the mixture (6) can be achieved by supposing that its component densities belong to the *Beta* parametric family, namely $f_{M,U}(d) = \mathcal{Beta}(d; \theta_{M,U})$ where $\theta_{M,U} = (\alpha_{M,U}, \beta_{M,U})$ and:

$$\mathcal{Beta}(d; \alpha, \beta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} d^{\alpha-1}(1-d)^{\beta-1} \tag{7}$$

with Γ denoting the Euler Gamma function and the shape parameters fulfilling $\alpha > 0$ and $\beta > 0$. Hence density (6) is turned into a *parametric* two-component mixture model $f(d) = f(d; \Psi)$ described by five independent parameters $\Psi = (\alpha_M, \beta_M, \alpha_U, \beta_U, \pi_M)$.

The basic reasons that led us to select the Beta family for our mixture model can be summarized as follows:

- The Beta distribution has *support* in (0,1). This is appropriate for our distance random variable d.[6]
- The Beta distribution is *flexible*. By opportunely tuning its α and β parameters the Beta density can take a broad range of shapes, such as: flat, U-shape (as well as inverse U-shape), J-shape (as well as mirrored J-shape), unimodal (both narrowly peaked or smooth), symmetrical or asymmetrical, and so on.
- The Beta distribution can be *skewed*, both positively and negatively. This property is desirable as we expect from the discussion about $\mathcal{PK1}$ in Section 2 that the distance densities for M and U classes are skewed (with longer right and left tail respectively).

[5] For economy of notation random variables will not be typographically distinguished from their realizations: the intended meaning should be clear from the context.

[6] There is a technical subtle point arising from the fact that the support of the Beta distribution does *not* include the boundary values $d = 0$ and $d = 1$, whereas pairwise distances equal to 0 or 1 can be (and in general are) observed in practice. Due to space limitations, we cannot describe in a detailed way how we solved this problem; we just point out that, instead of a mixture of simple Betas, we adopted a mixture of *Inflated Beta* distributions (Ospina et al, 2010).

- The parameters controlling the Beta distribution are *shape parameters*. This is an important point because it allows to easily translate our prior knowledge $\mathcal{PK}1$ into a well defined set of constraints acting on the parameter space (this would not have been the case if, instead, we decided to plug into our mixture model (6) some pdfs described by location and scale parameters).

We stress here that the primary aim of our mixture model is *not* the one of obtaining a *high-quality* description for the observed distance distribution of the d_is. If it were the case, one could be led to model a mixture with more than two components or, otherwise, to worry about the inability of the Beta family to represent multimodal distributions (look e.g. at the shape of the *M* density in the left panel of Figure 1). On the contrary we basically see the mixture as a device to *exploit* the observed d_is distribution in order to bring to light the hidden underlying grouping of the record pairs into the *M* and *U* classes. Since, at least in a mixture approach mainly aimed at *clustering*, it is the fit of the tails of the class distributions that turns out to be crucial (McLachlan et al., 1988) (rather than the fit of the main body of the data, for which a good "average description" seems enough), our choice of the Beta family will prove to be satisfactory.

Once equipped with our two-component Beta mixture model, we have to accomplish two major tasks in order to find a solution to a specific RL problem:

1. We must fit the mixture model to the observed distance values d_is, so as to obtain estimates $\hat{\Psi}$ for the mixture parameters.
2. We must achieve (upon plugging into the model the estimates $\hat{\Psi}$) a probabilistic clustering of the record pairs tied to the observed distances d_is into classes *M* and *U*.

The task of fitting a mixture model can be handled by a variety of methods, including e.g. the method of moments (MOM), Bayesian methods and graphical methods. In the present work we chose the Maximum Likelihood (ML) method which is by far the most popular. Anyway, the technique we developed for obtaining ML estimates for the parameters of our mixture model is indeed original. Therefore, the next section of the paper will be devoted to a detailed motivation and description of this technique.

Section 5 will describe the clustering task, which we shall face in a decision-theoretic framework. An *optimal* classification rule will be searched such that each record pair *i* can be assigned, based on its observed distance value d_i, either to the *M* or to the *U* class, in such a way as to optimize a given *global* objective function.

Before going into further details, we observe – as a general remark – that our methods can be ascribed as a whole to the frequentist inferential scheme. This seems to be the case also for most of the classical papers in the probabilistic RL literature. On the other side, some authors have proposed techniques to face the RL problem in a purely Bayesian framework, see e.g. (Fortini et al., 2001; Larsen, 2005). Although a discussion of such techniques is beyond the scope of the present paper, future works could explore the possibility to embed our strategies (e.g. the way we use $\mathcal{PK}1$ and $\mathcal{PK}2$) in a purely Bayesian inferential scheme.

4. Fitting the Mixture Model: a Perturbation-like Approach

Under our two-component Beta mixture model, the log-Likelihood associated to the observed distance values $\mathbf{d} = (d_1, \ldots, d_{n_P})$ reads:

$$\log \mathcal{L}(\mathbf{d}; \Psi) = \sum_{i=1}^{n_P} \log\left[\pi_M f_M(d_i; \theta_M) + \pi_U f_U(d_i; \theta_U)\right] \tag{8}$$

The maximum Likelihood estimator (MLE) of the unknown mixture parameters Ψ can thus be defined as follows:

$$\hat{\Psi} = \text{argmax}_{\Psi}\left[\, \log \mathcal{L}(\mathbf{d}; \Psi) \,\right] \tag{9}$$

Finding the MLE of a model is, with few exceptions, a hard task. Not only an analytic closed-form solution to the problem is generally not available, what's more maximizing the Likelihood by means of numerical routines turns out to be difficult for most of the general-purpose optimization algorithms. It is not by chance that an ad-hoc class of optimizers, namely the Expectation-Maximization (EM) (Dempster et al., 1977) family, has been specifically tailored to handle ML problems. Unfortunately, the situation is even worse for *mixture* models, since their Likelihood function is often unbounded over the parameter space and typically exhibit many spurious local maxima. Anyway, the EM (or an EM-like) algorithm remains by far the favorite tool to handle ML estimation of mixture model parameters.

Many authors (Jaro, 1989; Armstrong et al., 1992; Winkler, 1993; Winkler, 1994; Belin et al., 1995; Larsen et al., 1997; Larsen et al., 2001) have already proposed the use of mixture models to solve RL problems. Some comments are in order, since the method we developed to fit our mixture is significantly different from the ones implemented in those classical papers:

- They all exploit mixture models adopting a *Fellegi-Sunter* approach (Fellegi et al., 1969). Thus the role of our auxiliary unidimensional distance variable *d* is typically played by a *k*-vector variable whose components represent agreement/disagreement outcomes obtained when comparing record pairs on *k* matching fields.[7]
- With the only exception of (Armstrong et al., 1992), they all use an EM-like algorithm in the mixture model fitting phase.
- A mixture model with more than two components is generally selected. From a clustering point of view, this obviously raises the question of how many and which mixture components have to be associated to each of the *M* and *U* classes.

[7] On the whole, i.e. when taking into account the methods we propose for *both* the *fitting phase* (Section 4) and the *clustering phase* (Section 5), our approach differs very much from Fellegi-Sunter's one (FS). Just to mention some of the differences: *i)* FS has a third class (besides *M* and *U*), namely the *Possible Match* class; *ii)* FS relies on a completely different notion of "optimal decision rule", which is neither based on Maximum-Likelihood nor on Minimum-Cost, but rather involves the *Possible Match* class; *iii)* FS implementations tipically rely on the assumption of conditional-independence for the components of the comparison vector; *iv)* FS applications generally ask the user to set classification thresholds.

- Their methods always encompass some amount of "human work" which is specifically meant to incorporate some kind of "experience" into the mixture fit. Possible ways to achieve this result are the following: *i)* guesses based on previous linkage applications on similar data can be employed to build initial estimates of model parameters; *ii)* selected record pairs (typically the more difficult to classify) can be sent to clerical review and the mixture model can be re-fitted by using also those clerically classified cases; *iii)* a training set of pre-labeled record pairs can be prepared, in order to fit the mixture component(s) describing e.g. the *M* class by means of only cases surely belonging to that class.

With regard to last point, the cited papers agree on the following finding: as far as RL problems are concerned, mixture models tend to cluster the pairs into groups that, despite achieving a good fit, often do *not* correspond to the desired *M* and *U* classes. The reason for this behavior seems more controversial, with some authors ascribing it mainly to model misspecification and some others putting the emphasis on the difficulty in the estimation of mixture model parameters. Anyway, a sharp picture emerges from the literature above: the only way a mixture model can yield high-quality RL results is to incorporate in it some kind of "previous knowledge". We agree on this last conclusion, but we contend that the necessary prior knowledge can be incorporated *without* relying on "clerical work", thus not jeopardizing automation.

In our opinion the poor *clustering* results that a "basic" (i.e. not experience-enriched) mixture-model would generally achieve in RL applications have to be mainly imputed to the huge class-skew inherent in these problems. More specifically, we argue that troubles would usually arise from the previous model *fitting* phase, due to the extreme Match rarity. Indeed, unless some countermeasure is adopted, whatever fitting algorithm would tend to tune *all* the model parameters so as to better describe some peculiar feature of the dominating *U* distance distribution.

The fitting technique we propose tries to exploit our two-fold previous knowledge ($\mathcal{PK}1$ and $\mathcal{PK}2$) in order to prevent the few (and so far unidentified) distance values stemming from the *M* class from being completely overwhelmed by those belonging to the *U* class. This is accomplished by means of a *Two-Step* algorithm. Before going into details, we offer here an intuitive insight into its working mechanism. The *First-Step* concentrates on the *U* component mixture parameters and is specifically aimed at "factorizing" the leading contribution arising from Unmatches. The *Second-Step* strives to increase the Likelihood achieved in the previous step by using the remaining mixture parameters in a "smart way"; that is, *M* density parameters are tuned in such a way as to better fit the behavior of the distance distribution *exactly* in those regions of the [0,1] interval in which values stemming from Matches are more likely to be found.

Our two-step algorithm follows a *perturbation-like* approach to the ML mixture fitting problem (9). *Perturbation Theory* (Bender et al., 1999) is a family of mathematical methods aimed at finding an approximate solution to problems that cannot, in general, be solved exactly *but* would become easy to solve if a parameter, say ε, had a given value, say $\varepsilon = 0$. The key idea is to build an approximation to the unknown solution of the true (i.e. $\varepsilon \neq 0$) problem by *perturbing* the known solution of the easier (i.e. $\varepsilon = 0$) problem, that is by adding to it further terms. These "higher orders" terms can be computed iteratively by some systematic procedure and, if $\varepsilon < 1$, turn out to be suppressed by increasing powers of ε. As a consequence, a

satisfactory solution can very often be obtained by truncating the series at its second term, that is by retaining only the initial solution and the first-order perturbative correction.

Due to prior knowledge $\mathcal{PK}2$, we are aware that the true unknown value of the mixing weight π_M is very small:

$$\pi_M \leq \frac{1}{n_{max}} << 1 \tag{10}$$

Moreover, in the limit $\pi_M = 0$, our hard *mixture* MLE problem (9) is turned into the much simpler problem of finding the MLE for the *U* density *alone*. Therefore we are allowed to look for a perturbative expansion of our mixture model parameters in powers of π_M:

$$\theta_U = \theta_U^{(0)} + \theta_U^{(1)}\pi_M + \theta_U^{(2)}\pi_M^2 + \cdots \tag{11}$$

$$\theta_M = \theta_M^{(0)} + \theta_M^{(1)}\pi_M + \theta_M^{(2)}\pi_M^2 + \cdots \tag{12}$$

where the $^{(j)}$ superscript denotes the *j*-th-order coefficient to be estimated. By inserting expansions (11) and (12) inside (9) we get a hierarchy of sub-problems that can be solved in a chain to yield the desired estimates $\hat{\theta}_{M,U}^{(j)}$. As we are going to see, our First-Step and Second-Step optimizations are respectively in charge of solving the zeroth-order and first-order approximations of problem (9). Moreover, Second-Step optimization has also to incorporate the first piece of prior knowledge we collected, i.e. $\mathcal{PK}1$.

4.1 First-Step Optimization

To zeroth-order in perturbation theory our MLE problem reads:

$$\log \mathcal{L}_I(\mathbf{d}; \theta_U^{(0)}) = \sum_{i=1}^{n_P} \log\left[f_U(d_i; \theta_U^{(0)}) \right] \tag{13}$$

$$\hat{\theta}_U^{(0)} = \underset{\theta_U^{(0)}}{\operatorname{argmax}} \left[\log \mathcal{L}_I(\mathbf{d}; \theta_U^{(0)}) \right] \tag{14}$$

where our First-Step *effective Likelihood* $\mathcal{L}_I$ differs from the real Likelihood $\mathcal{L}$ by terms that are at most of order π_M.

Since the mixture structure has disappeared from (13), an EM-like optimizer is no longer a mandatory choice. Indeed, when implementing the method, we chose to maximize $\mathcal{L}_I$ by means of a *faster* quasi-Newton[8] multivariate optimization algorithm. The starting guess needed to initialize the optimizer was computed as the MOM estimator of parameters $\theta_U^{(0)} = (\alpha_U^{(0)}, \beta_U^{(0)})$ given the observed distance distribution, namely:

[8] This required to derive the analytical expression of the gradient of the effective log-Likelihood (13), which we cannot report here due to space limitations.

$$\alpha_U^{(0)}\Big|_{start} = \left[\frac{\bar{d}(1-\bar{d})}{v} - 1\right]\bar{d} \tag{15}$$

$$\beta_U^{(0)}\Big|_{start} = \left[\frac{\bar{d}(1-\bar{d})}{v} - 1\right](1-\bar{d}) \tag{16}$$

where the sample mean $\bar{d} = (1/n_P)\sum_{i=1}^{n_P} d_i$ and the sample variance $v = (1/n_P)\sum_{i=1}^{n_P}(d_i - \bar{d})^2$ of the d_is have been used.

Coming back to the method, once a solution for the First-Step problem (14) is at hand, it becomes possible to take a step further in the perturbative approximation of the original MLE problem. This is accomplished by the subsequent Second-Step optimization.

4.2 Second-Step Optimization

If, after having inserted expansions (11) and (12) inside (9), we keep terms up to first-order in π_M and, in addition, we use the achieved zeroth-order solution (14), we get:

$$\log \mathcal{L}_{II}(\mathbf{d}; \theta_M^{(0)}, \pi_M) = \sum_{i=1}^{n_P} \log\left[\pi_M f_M(d_i; \theta_M^{(0)}) + (1-\pi_M) f_U(d_i; \hat{\theta}_U^{(0)})\right] \tag{17}$$

where our Second-Step *effective Likelihood* $\mathcal{L}_{II}$ differs from the real Likelihood $\mathcal{L}$ by terms that are at most of order π_M^2.

It is worth noting that, despite contributions of order π_M have been retained, equation (17) does *not* contain the first-order coefficient $\theta_U^{(1)}$ of (11), that is $\mathcal{L}_{II}$ does *not* depend on U parameters. This is a direct consequence of $\hat{\theta}_U^{(0)}$ being a (local) maximum of $\log \mathcal{L}_I$. Indeed, as the gradient of $\log \mathcal{L}_I$ is zero in $\hat{\theta}_U^{(0)}$, a deviation of order π_M from $\hat{\theta}_U^{(0)}$ can have at most an effect of order π_M^2 on the log-Likelihood. Therefore, everything goes as if we were now switching on the parameters describing the M component of the mixture, while those from the U component have been "frozen" to the estimated values found in the previous optimization step.

We must now incorporate our prior knowledge $\mathcal{PK}1$ and $\mathcal{PK}2$ inside the Second-Step optimization problem. Since equation (10) already summarizes $\mathcal{PK}2$, we have only to translate $\mathcal{PK}1$ into a set of constraints acting on M parameters $\theta_M^{(0)} = (\alpha_M^{(0)}, \beta_M^{(0)})$. This task can be completed by studying the dependence of the Beta distribution (7) on the shape parameters and by exploiting known formulae for its first moments; the following chain of translations results:

1. Unmatches are mainly located in the high distance region and the U density is negatively skewed:
$$\beta_U < \alpha_U$$
2. Matches are mainly located in the low distance region and the M density is positively skewed:
$$\beta_M > \alpha_M$$

3. The U density dominates the M density in the limit $d \to 1$:

$$\beta_M > \beta_U$$

4. The M density dominates the U density in the limit $d \to 0$:

$$\alpha_M < \alpha_U$$

5. The M and U densities have a small overlap. The easiest way to express this somewhat fuzzy requirement, while fulfilling constraints 1) – 4), is as follows:

$$\alpha_M < \beta_U \quad and \quad \beta_M > \alpha_U$$

Replacing inside 1) – 5) U parameters with estimates and neglecting redundant constraints, we eventually obtain the complete formulation of the Second-Step MLE problem:

$$\{\hat{\theta}_M^{(0)}, \hat{\pi}_M\} = \underset{\{\theta_M^{(0)}, \pi_M\}}{\text{argmax}} \left[\log \mathcal{L}_{II}(\mathbf{d}; \theta_M^{(0)}, \pi_M) \right] \tag{18}$$

$$subject\ to:$$

$$\alpha_M^{(0)} < \hat{\beta}_U^{(0)} \tag{19}$$

$$\beta_M^{(0)} > \hat{\alpha}_U^{(0)} \tag{20}$$

$$\pi_M \leq 1/n_{max} \tag{21}$$

It should be noted that, since $\pi_M = 0$ is a feasible point for the constrained ML problem (18)-(21), and as $\mathcal{L}_{II}(\mathbf{d}; \theta_M^{(0)}, \pi_M = 0) \equiv \mathcal{L}_I(\mathbf{d}; \hat{\theta}_U^{(0)}) \quad \forall \theta_M^{(0)}$, the following inequality will be satisfied: $\max(\log \mathcal{L}_{II}) \geq \max(\log \mathcal{L}_I)$. Consequently, Second-Step Optimization cannot decrease the Likelihood achieved in the First-Step, but rather will in general *increase* it. This is coherent with the quick outline we gave in Section 4 on our Two-Step perturbative fitting technique.

Furthermore, thanks to the decoupling of U and M parameters, once again the need of an EM-like optimizer has been overcome. Indeed, the software we developed solves the *constrained* ML problem (18)-(21) by means of a box-constrained quasi-Newton[9] multivariate algorithm. Besides the usual by-product of saving computation time, this choice freed us from the tricky machinery required to manage a *constrained* EM-like algorithm (Winkler, 1993). When testing our application, random starting values drawn from the feasible region (19)-(21) were used to initialize the optimizer. Because fairly stable results were found, we eventually fixed the starting guess as follows:

[9] Again, this required to derive the analytical expression of the gradient of the effective log-Likelihood (17), which we cannot report here due to space limitations.

$$\alpha_M^{(0)}\Big|_{start} = \hat{\beta}_U^{(0)}/10 \tag{22}$$

$$\beta_M^{(0)}\Big|_{start} = 10\hat{\alpha}_U^{(0)} \tag{23}$$

$$\pi_M\Big|_{start} = (1/2)(1/n_{max}) \tag{24}$$

Computing a solution of the Second-Step optimization problem (18)-(21) completes the task of fitting our two-component Beta mixture model. By plugging into the model the achieved estimates:

$$\hat{\Psi} = (\hat{\alpha}_M^{(0)},\ \hat{\beta}_M^{(0)},\ \hat{\alpha}_U^{(0)},\ \hat{\beta}_U^{(0)},\ \hat{\pi}_M) \tag{25}$$

we can now switch to the problem of *clustering* the record pairs into classes *M* and *U*.

5. Clustering Pairs using the Mixture Model

As we already mentioned in Section 3, our mixture model has been specifically designed to be used for clustering purposes. The goal is, indeed, to exploit the model – along with its ML estimated parameters (25) – to assign each record pair *i* either to the *M* or to the *U* class. Clearly the obtained classification for the *i*-th pair will depend on its observed distance value d_i.

5.1 Optimal Classification Rules from Decision Theory

Let us attach to each record pair *i* a class membership indicator z_i with value 1 if the pair is a Match and 0 otherwise. The *true* value of z_i is obviously *unknown*: it will precisely represent the target of our inferences. We shall, consequently, treat variables z_i as *iid*[10] realizations of a *latent* random variable *z*. Variable *z* can be incorporated inside our mixture model (6), which describes the distance pdf, by assuming that:

1. Variable *z* is distributed according to a single draw from a Binomial distribution with success probability given by π_M.
2. The conditional densities of *d*, given *z* = 1 and *z* = 0, are $f_M(d;\theta_M)$ and $f_U(d;\theta_U)$ respectively.

Under conditions 1) and 2), the *complete* mixture density can be expressed as follows:

$$g(d,z;\Psi) = \left[\pi_M f_M(d;\theta_M)\right]^z \left[\pi_U f_U(d;\theta_U)\right]^{1-z} \tag{26}$$

Correspondingly, the *complete* (and *unobservable*) log-Likelihood associated to the observed distance values $\mathbf{d} = (d_1,\ldots,d_{n_P})$ and to the hidden class labels $\mathbf{z} = (z_1,\ldots,z_{n_P})$ reads:

[10] Notice that the independence assumption on variables z_i cannot hold true for 1:1 RL problems (i.e. when the data sets to be matched do not contain duplicates). We shall come back to this issue in Section 5.2.

$$\log \mathcal{L}^c(\mathbf{d},\mathbf{z};\Psi) = \sum_{i=1}^{n_P} \log\left[g(d_i,z_i;\Psi)\right] =$$
$$= \sum_{i=1}^{n_P} \log\left[\pi_U f_U(d_i;\theta_U)\right] + \sum_{i=1}^{n_P} z_i \log\left[\frac{\pi_M f_M(d_i;\theta_M)}{\pi_U f_U(d_i;\theta_U)}\right] \quad (27)$$

Thanks to (26), the mixing weights π_M and π_U can now be understood as the *prior probabilities* that the *i*-th pair belongs to class *M* and *U* respectively, while the corresponding *posterior probabilities*, given the distance value d_i observed for the pair, are:

$$\tau_i^C(d_i;\Psi) = \pi_C f_C(d_i;\theta_C)/f(d_i;\Psi) \qquad C = \{M,U\} \quad (28)$$

Estimates of posterior probabilities $\hat{\tau}_i^C$ can be built by simply plugging into (28) the previously computed estimates (25) for the model parameters $\hat{\Psi}$. As we are going to see, these values $\hat{\tau}_i^C$ play a central role in the clustering task.

A "classification rule" that assigns each record pair *i* to a class is nothing but a rule to *infer* a value $\hat{z}_i$ for the hidden variable z_i. An *optimal* rule has moreover to work in such a way as to optimize some global objective function.

A first, very natural choice is to select as objective function the complete log-Likelihood itself. This means that we look for an allocation vector $\hat{\mathbf{z}}$ that maximizes the complete data Likelihood under the model (26), namely:

$$\hat{\mathbf{z}} = \text{argmax}_{\mathbf{z}}\left[\log \mathcal{L}^c(\mathbf{d},\mathbf{z};\hat{\Psi})\right] \quad (29)$$

The solution of (29) follows easily from the structure of (27):

$$\hat{z}_i = \begin{cases} 1 & if \quad \hat{\tau}_i^M \geq \hat{\tau}_i^U \\ 0 & otherwise \end{cases} \quad (30)$$

Formula (30) is a classical Decision Theory result (Duda et al., 2000), known as the "Bayes decision rule" (or as "Maximum a Posteriori (MAP) rule"): it assigns each pair to the class to which the pair has the highest estimated posterior probability of belonging.

A sometimes useful alternative is to find the classification rule that minimizes the expected value of a *Loss Function*. For instance, different costs can be assigned to the possible outcomes of a decision (Verykios et al., 2003). Indeed, the cost C_{PU} for declaring a pair to be a Match (we call this a *positive* decision) when it is actually an Unmatch can differ from the cost C_{NM} for declaring a pair to be an Unmatch (we call this a *negative* decision) when it is actually a Match. In this situation, an appropriate Loss Function would be the expected *Total Cost* associated to a classification $\hat{\mathbf{z}}$:

$$C_{TOT}(\mathbf{d},\hat{\mathbf{z}};\hat{\Psi}) = \sum_{i=1}^{n_P}\left[C_{PM}\hat{\tau}_i^M + C_{PU}\hat{\tau}_i^U\right]\hat{z}_i + \sum_{i=1}^{n_P}\left[C_{NM}\hat{\tau}_i^M + C_{NU}\hat{\tau}_i^U\right](1-\hat{z}_i) \quad (31)$$

Minimizing the expected Total Cost (31) with respect to $\hat{\mathbf{z}}$ is straightforward and leads to the following optimal decision rule:

$$\hat{z}_i = \begin{cases} 1 & if \quad (C_{NM} - C_{PM})\,\hat{\tau}_i^M \geq (C_{PU} - C_{NU})\,\hat{\tau}_i^U \\ 0 & otherwise \end{cases} \tag{32}$$

Once again, formula (32) is a classical Decision Theory result (Duda et al., 2000). If a zero cost is assigned to a correct decision (i.e. $C_{PM} = C_{NU} = 0$) and the same cost, say one, is assumed for both kinds of wrong decisions (i.e. $C_{PU} = C_{NM} = 1$), the expected Total Cost (31) simply measures the expected number of classification errors; accordingly, the optimal decision rule (32) is turned into the Bayes rule (30).

The software we developed is able to cluster the pairs according to both (30) and (32) rules. Anyway, these clustering results have to be understood only as *"provisional solutions"* to the RL problem at hand. Indeed, while building our mixture fitting method in Section 4, we assumed – recall e.g. equation (21) – that the data sets to be matched did *not* contain duplicates. This obviously translates into a set of one-to-one matching constraints that, in general, are *not* fulfilled by the aforementioned optimal decision rules. Next section is devoted to describe how our method overcomes this difficulty.

5.2 Dealing with One-to-One Matching Constraints

As data sets *A* and *B* do not contain duplicates, each record belonging to either data set can match at most a single record selected from the other data set. Therefore, the number of true Matches cannot exceed the cardinality of the smaller data set, i.e. $n_M \leq n_{min}$, whence equations (10) and (21) follow.

In order to express these one-to-one (1:1) matching constraints in a formal way, let us switch to a more convenient matrix notation. We start by arranging the n_P observed distance values d_i into a $n_{min} \times n_{max}$ matrix D, in such a way that element D_{ij} represents the distance between the *i*-th record of the smaller data set and the *j*-th record of the bigger data set. Next we introduce matrices Z and $\hat{Z}$ to store, in the same way, the unknown class membership indicators of the pairs and their inferred values, respectively. Accordingly, quantities depending on features of the generic *i*-th pair (like d_i, z_i, $\hat{z}_i$, $\hat{\tau}_i^C$, and so on) have to be replaced, inside all previous sections formulae, by the corresponding two-index quantities (like D_{ij}, Z_{ij}, $\hat{Z}_{ij}$, $\hat{\tau}_{ij}^C$, and so on). At the end, 1:1 matching constraints can be readily incorporated into the problem of finding an optimal decision rule. For instance, the ML problem (29) now reads:

$$\hat{Z} = \operatorname{argmax}_Z \left[\log \mathcal{L}^c(D, Z; \hat{\Psi}) \right] \tag{33}$$

subject to:

$$\sum_{j=1}^{n_{max}} Z_{ij} \leq 1 \quad \forall i \tag{34}$$

$$\sum_{i=1}^{n_{min}} Z_{ij} \leq 1 \quad \forall j \tag{35}$$

with (34) and (35) obviously implying $\sum_{ij} Z_{ij} \leq n_{min}$.

Problem (33)-(35) deserves some comments. First of all, it is apparent that, due to the 1:1 restrictions (34) and (35), variables Z_{ij} cannot be *stochastically independent* (e.g. if $Z_{km} = 1$ then $Z_{im} = 0 \ \forall i \neq k$ and $Z_{kj} = 0 \ \forall j \neq m$). As a consequence, being the underlying *iid* property violated, one should not use formula (27) to express the complete data Likelihood. Anyway, the task of deriving its correct expression without relying on the *iid* assumption turns out to be too difficult to be handled. Therefore, one is led to the practical compromise of solving the constrained problem (33)-(35) while keeping the old formula (27) for the complete data Likelihood.

In any case, 1:1 matching constraints heavily affect the complexity of both ML and Minimum Cost optimization problems: now, indeed, a decision taken on a pair *influences* decisions to be taken on other pairs. Moreover, clustering results based on classical decision rules (30) and (32) will not, in general, fulfill 1:1 constraints. Consequently, our software treats these quickly computed results as *"provisional solutions"*. This means that they are checked against (34) and (35), and, only if 1:1 constraints happen to be already satisfied, they are retained as *"definitive results"*. When, on the contrary, 1:1 constraints turn out to be broken, *"definitive results"* are searched by facing directly the constrained optimization problem (33)-(35) (or its Minimum Cost counterpart). This new – and *harder* – clustering task is accomplished by means of a purposefully designed Evolutionary Algorithm (EA) (Michalewicz, 1996). Even in this case the work carried out to compute the *"provisional solutions"* will not get wasted, as useful pieces of information stemming from these solutions will be exploited by the clustering EA.

Before going into further details, we briefly argue why we chose an EA. A few papers from the RL literature (Jaro, 1989; Winkler, 1994) tackled the 1:1 problem[11] by using Simplex-based algorithms. Indeed, since both the objective function and the constraints are linear in Z_{ij}, equations (33)-(35) can be formulated as a Binary Linear Programming (BLP) problem. The main concern with this approach is tied to memory usage. As a matter of fact, if one denotes with n the size of the data sets to be matched (i.e. $n_A \approx n_B \approx n$), one sees that the number of unknowns and inequalities for the BLP problem grow like n^2 and n, respectively. The net result is that, due the heavy memory overhead of a Simplex-based solver, the BLP approach cannot be applied to real-world data sets *unless* a very efficient previous blocking step has been performed. On the contrary, as we shall see in Section 5.3, the size of the biggest data structure stored by our EA grows almost linearly with n. As a consequence, our software was able to handle all the RL problems listed in Section 6 (with the obvious exception of the huge $\mathbf{PES_{full}}$ instance, see the discussion therein) by running in an ordinary PC environment and without relying on blocking. Finally, we observe that our EA could be readily applied to clustering tasks that involve more complex Loss Functions than (32), e.g. nonlinear Loss Functions.

[11] Notice, however, that those authors simply force 1:1 clustering restrictions on a statistical model that does not encompass them. On the contrary, we already took into account 1:1 matching constraints while fitting our mixture model parameters (see equation (21)).

5.3 An Evolutionary Algorithm for 1:1 Clustering

The Evolutionary optimization metaheuristic is so versatile that EAs can often be employed to find a satisfactory solution even for problems for which no other solution strategy is known. This extreme flexibility has two major prices: i) no standard design rules for EAs are available; ii) EAs performance critically depend on how smartly problem-specific pieces of information are incorporated into the algorithm. Considerations i) and ii) should push us to provide a thorough justification for each aspect of our clustering EA. Nevertheless, due to space limitations, we shall restrict ourselves to a very concise outline.[12] In what follows we list the pseudocode of the algorithm and quickly describe basic choices, parameters and operators.

EA PSEUDOCODE

```
EA[n_ind, n_gen, p_muta, g_stall]
g ← 0
generate Initial Population[n_ind]
compute Fitness
while (¬ Termination Criterion[n_gen, g_stall]) do
  g ← g + 1
  apply Selection
  apply Reproduction + Repair
  apply Mutation[p_muta]
  compute Fitness
end while
return Best Fit individual found
```

Search Space. The EA search space is the set of all the $n_{min} \times n_{max}$ matrices Z with $\{0,1\}$ elements that fulfill 1:1 constraints (34) and (35). It is a *huge* search space whose cardinality is given by:

$$\mathcal{N}_Z = \sum_{k=0}^{n_{min}} \binom{n_{max}}{k} \binom{n_{min}}{k} k!$$

(just to get an impression: if $n_{max} = n_{min} = 150$ then $\mathcal{N}_Z = 1.28 \times 10^{272}$).

Representation. We encode a generic candidate solution Z (*phenotype*) by means of a vector ζ of length n_{min} (*genotype*). Elements of ζ (*alleles*) can be 0 or integers between 1 and n_{max}, namely $\zeta_k \in \{0,1,\ldots,n_{max}\}$ with $k = 1, 2, \ldots, n_{min}$. The meaning of the alleles is easily understood. If $\zeta_k = 0$, then the candidate solution states that the k-th record of the smaller data set *does not match any record* of the bigger data set. If, on the contrary, $\zeta_k = j > 0$, then the candidate solution states that the k-th record of the smaller data set

[12] We assume a basic knowledge of EAs and refer to (Michalewicz, 1996) (and references therein) for more advanced topics.

does match the *j*-th record of the bigger data set. Obviously a *legal* genotype, that is a genotype encoding a *feasible* candidate solution *Z*, is not allowed to contain *duplicated* alleles other than the 0 allele.

Fitness. The `Fitness` functions for ML and Minimum Cost clustering are obviously modeled on the corresponding objectives (27) and (31). For instance, in the ML case we have:

$$\text{Fitness}(\zeta) = \sum_{k:\zeta_k>0} \log\left(\hat{\tau}^{M}_{k\zeta_k} \Big/ \hat{\tau}^{U}_{k\zeta_k} \right) \tag{36}$$

where uninfluential constant terms appearing in (27) have been dropped.

Constraints. Even though only legal individuals are generated in the initial population, some *illegal* genotype may arise during evolution, due to `Reproduction`. In order to maintain a population of feasible candidate solution, these illegal genotypes are *repaired* by means of a purposefully designed operator. The `Repair` operator, $\texttt{Repair}:\zeta \to \zeta^*$, acts as a *stochastic* function mapping a genotype, ζ, into a *randomly repaired* version of it, ζ^*. If the ζ individual is legal, `Repair` leaves it unchanged. If, instead, ζ is illegal, `Repair` works as follows. Suppose ζ has ρ groups of duplicated non-zero alleles, with multiplicities n_r where $r = 1,\ldots,\rho$. For each group *r*, `Repair` first randomly selects inside the group just a single allele to be left unchanged, then it substitutes all the remaining $n_r - 1$ duplicates with the 0 allele.

Initial Population. As the search space of our EA is so huge, generating a good initial population is crucial. It is apparent that creating n_{ind} random individuals by *uniformly* sampling the search space would be a very poor choice. On the contrary, our algorithm samples more heavily those regions of the search space that are believed to be "more promising" on the basis of the already computed posterior probabilities (28). This is accomplished by the following Monte Carlo (MC) technique. First, the following posterior probabilities are computed for each record *k* in the smaller data set:

$$p^0_k = \Pr(Z_{ki} = 0 \;\forall i) = \prod_i \hat{\tau}^{U}_{ki} \tag{37}$$

$$p^j_k = \Pr((Z_{kj}=1) \;\; AND \;\; (Z_{ki}=0 \;\; \forall\, i \neq j)) = \hat{\tau}^{M}_{kj} \prod_{i \neq j} \hat{\tau}^{U}_{ki} \tag{38}$$

where $j = 1, 2, \ldots, n_{max}$. Value p^0_k is the posterior probability that the *k*-th record of the smaller data set does not match any record of the bigger, while p^j_k gives the posterior probability that the *k*-th record matches *only*[13] the *j*-th. The MC procedure generates *each* element ζ_k (for $k = 1, 2, \ldots, n_{min}$) of *each* genotype ζ of the initial population by sampling an allele value from $\{0, 1, \ldots, n_{max}\}$ with probability proportional to (37) and (38), i.e. $\Pr(\zeta_k = 0) \propto p^0_k$ and $\Pr(\zeta_k = j > 0) \propto p^j_k$. These MC generated genotypes are eventually processed by the The `Repair` operator, in order to warranty that the whole initial population is *legal*.

[13] Recall that the model in Section 5.1 does not encompass 1:1 constraints for the latent variable.

Selection. `Selection` is performed by means of a *rank-2 tournament*. Random pairs of individuals are formed. For each pair, the fitness of the individuals are compared. The fitter individual survives whereas the weaker dies and is dropped from the population, so as to make room for new individuals to be generated in the Reproduction phase. Notice that this `Selection` method is intrinsically *elitist*: the fittest individual of a generation surely survives and passes to the next generation.

Reproduction. Reproduction is performed as follows. Individuals that survived to `Selection` are randomly paired. Each pair generates two children. These children take the place of individuals that have been eliminated in the previous `Selection` phase. As a consequence the size of the population n_{ind} is kept *fixed* during the evolution. Children genotypes are obtained by merging those of the parents by means of *one-point crossover*. Call ζ^{p1} and ζ^{p2} the parents and ζ^{c1} and ζ^{c2} the children. A random cut point $cut \in \{1,\ldots,n_{min}-1\}$ is selected for the parents genotypes. Hence both ζ^{p1} and ζ^{p2} are cut into a *left* portion and a *right* portion. The first child receives the left portion from the first parent and the right from the second, i.e. $\zeta^{c1} = (\zeta_1^{p1},\ldots,\zeta_{cut}^{p1},\zeta_{cut+1}^{p2},\ldots,\zeta_{n_{min}}^{p2})$. The second child receives the left portion from the second parent and the right from the first, i.e. $\zeta^{c2} = (\zeta_1^{p2},\ldots,\zeta_{cut}^{p2},\zeta_{cut+1}^{p1},\ldots,\zeta_{n_{min}}^{p1})$. As there is no warranty that the generated children ζ^{c1} and ζ^{c2} are legal, they eventually undergo the `Repair` treatment before being plugged into the population.

Mutation. Each individual of the population has the same probability p_{muta} of undergoing `Mutation`. `Mutation` acts on a genotype ζ by affecting only a single allele. The outcome can be either that a nonzero allele is replaced by 0 (i.e. a declared Match is deleted from *Z*), or that a 0 allele is turned into a nonzero allele (i.e. a new declared Match is inserted into *Z*). The stochastic algorithm implementing `Mutation` exploits the *posterior estimate* of the number of Matches $\hat{n}_M = \sum_i \hat{z}_i$ obtained from (30) (or (32) for Minimum Cost). A random integer $p \in \{0,1,\ldots,n_{min}\}$ is drawn from a Binomial distribution with size n_{min} and success probability $\hat{n}_M / n_{min}$. If the genotype to mutate has more than *p* nonzero alleles, $\sum_k \text{sgn}(\zeta_k) > p$, then a random nonzero allele is replaced by 0. Otherwise, i.e. when $\sum_k \text{sgn}(\zeta_k) \leq p$, a random 0 allele is replaced by a nonzero one randomly selected from the set $\{1,\ldots,n_{max}\} \setminus \zeta$ (namely, by a new *legal* nonzero allele that did not already appear in ζ). Notice that, since the expected value of *p* is exactly $\hat{n}_M$, `Mutation` *tends on average* to delete declared Matches from candidate solution that contain *"too many"* of them, and conversely to add declared Matches to candidate solution that contain *"too few"* of them.

Termination Criterion. The Termination Criterion for the EA is two-fold. A first parameter, n_{gen}, controls the maximum number of generations that can be spent during evolution. If g denotes a generations counter, then the EA would stop as soon as $g > n_{\text{gen}}$. A second parameter, g_{stall}, gives the maximum number of generations that the EA is allowed to process without achieving a fitness improvement. If g' denotes the number of generations elapsed from the *last* fitness improvement, then the EA would stop as soon as $g' > g_{\text{stall}}$. The EA effectively stops as soon as either of the two conditions is verified.

Return Value. The return value of the EA is the genotype ζ^{Best} of the `Best Fit` individual found during evolution. This genotype is readily decoded into the corresponding phenotype matrix Z^{Best}, which in turn yields the *final* clustering result for the RL problem.

Memory Usage and Parameters Values. Storing a whole population of candidate solutions determines the EA memory overhead. As population size is kept fixed during evolution, if n denotes the size of the data sets to be matched then memory usage grows like $n_{\text{ind}} \times n$, i.e. almost linearly with n. Indeed, only a weak (less than linear) dependence of n_{ind} on n is expected. As a matter of fact, all the case studies listed in Section 6, despite their n values span over nearly an order of magnitude, have been carried out with the following default values for the EA parameters: $n_{\text{ind}} = 300$, $n_{\text{gen}} = 200$, $p_{\text{muta}} = 0.1$, $g_{\text{stall}} = 50$.

6. Experiments

Here we present an experimental evaluation of our mixture based suite of methods. Our focus will be on effectiveness; however, with respect to time complexity, we point out that our methods perform quadratically in the input data sets size (see equations (13), (17) for the fitting phase and (37), (38) for the clustering phase). Experiments have been carried out by using a comprehensive software system that implements all the methods proposed in the previous sections. We developed the system in the R programming language (R Development Core Team, 2009). All experiments have been run in an ordinary PC environment, equipped with: Windows XP 64 Operating System, 4 GB RAM, 2 GHz CPU.

We shall describe 9 RL instances involving 5 *very different* data sources. Indeed, a major aim of this section is to verify the *robustness* of our system against variations of the main characteristics of the RL problem. These include: data set size, Match rate (i.e. fraction of pairs that are Matches), type of records to be matched, number and discrimination power of variables used to compute distance measures, error rates affecting such variables, tendency of Unmatches to be similar even for clean data.

6.1 Experimental Setup

We first introduce the quality measures that we are going to use to assess the effectiveness of our RL system. We completely agree with (Christen et al., 2007) and hence avoid the Accuracy measure that, due to the huge class-skew inherent in RL problems, always gives a misleading impression of high effectiveness. On the contrary we choose to rely on traditional Precision (Prec) and Recall (Rec) measures. Moreover, whenever a single quality measure will be needed, we shall select the F-measure, $\mathrm{F} = 2/(\mathrm{Prec}^{-1} + \mathrm{Rec}^{-1})$. Notice that the F-measure is a *conservative* quality measure, as it can reach an high value only when *both* Precision and Recall are high.

A fundamental issue influencing our testing strategy concerns the *distance function* to be used in the RL process. As it should be clear from the previous sections, our mixture based approach does *not* rely on any restrictive assumption on the distance function (other than supposing it has unidimensional values). Therefore, our RL system can cope with *every* distance function (vectorial measures can easily be handled by averaging their components in a suitable way). Anyway, it is obvious that the choice of adopting a distance function rather than another for a specific RL task, can (and in general will) affect the

quality of the results. This is simply because, as a very rich literature in the RL field confirms (Elmagarmid et al., 2007), some distance functions are abler than others to capture some *specific aspects* of a given RL task. Nevertheless, since the comparison of alternative distance functions is completely beyond the scope of the present research, we shall use just a single distance for each RL instance. As a consequence, when reporting the quality of our results, the problem would arise of understanding how much of that quality actually depends on our methods, and how much, instead, on the adopted distance function. In other words, as the choice of the distance is just a free input for our system, we would like to build some kind of performance measure that is able to *factorize* the influence of the distance function. We developed such a "distance-independent" quality measure by exploiting what we call the 'Optimal Threshold Fully Supervised' classifier (OTFS).

The OTFS is a *theoretical* device. It is a 'Threshold'-based classifier in the sense that it classifies as Matches all the pairs with distance below a given threshold, and as Unmatches all the pairs above it. It is 'Fully Supervised' because it has full access to the true class labels of all pairs. It is an 'Optimal' classifier as, by knowing in advance the true results of the RL problem, it determines its classification threshold in such a way as to maximize the F-measure.[14] How to exploit the OTFS is easily understood. Indeed, given a quality metric Q (where Q $\in$ {Prec, Rec, F}), an approximately distance-independent measure of Q for our system can be computed as:

$$\Lambda_{\mathrm{Q}} = \mathrm{Q}_{\mathrm{SYS}}(dist)/\mathrm{Q}_{\mathrm{OTFS}}(dist) \tag{39}$$

where both classifiers, our system (SYS) and the OTFS, rely on the *same* distance function *dist*. We believe, in fact, that, even though both the numerator and the denominator in (39) depend on the choice of the distance, these dependencies will almost completely cancel out in the ratio.

6.2 Data Sources and RL Instances

Now we briefly describe the 9 proposed RL instances and the underlying 5 data sources, to which we shall refer as **Restaurants**, **Parks**, **Cens**, **Physics** and **PES**. Table 1 reports some basic information concerning these RL instances. With the only exception of **Cens**, all RL instances involve real-world data. All these problems are very hard, as indicated by (though not exclusively due to) their very low Match rates.

For all problems we choose the Levenshtein distance. When more than one matching variable is used, the following averaging procedure is adopted to obtain a scalar distance value. Call $\mathbf{d}^j = (d^j{}_1, \ldots, d^j{}_{n_P})$ the distance values measured with respect to the j-th matching variable on the n_P pairs,[15] and denote their mean and standard deviation with μ^j and σ^j. First, standardize these values and sum the standardized scores: $\mathbf{d}' = \sum_j [(\mathbf{d}^j - \mu^j)/\sigma^j]$. Then, simply normalize the obtained values in such a way that they fall inside the interval [0, 1], namely $\mathbf{d} = [\mathbf{d}' - \min(\mathbf{d}')]/[\max(\mathbf{d}') - \min(\mathbf{d}')]$.

[14] Notice that knowing in advance the true results of the OM problem is in general not sufficient for the OTFS to find a perfect classification threshold such that F = 1. Indeed, this is possible only if the histograms of the M and U distance distributions do not overlap at all.

[15] Whenever a variable had a missing value in one (or both) the records of a pair, we set the corresponding distance contribution to the blind average value 0.5.

Table 1 - Relevant Features of RL instances

RL Instance	Data Origin	Matching Variables	Data/Error Nature	Pairs (n_{min} x n_{max})	Matches	Match Rate
$Rest_1$	Riddle	name address city type	Real/Real	176,423 (331 x 533)	112	6.3E-4
$Rest_2$	Riddle	name	Real/Real	176,423 (331 x 533)	112	6.3E-4
Parks	SecondString	name	Real/Real	101,394 (258 x 393)	247	2.4E-3
Cens	SecondString	surname name midinit number street	Artificial/Artificial	176,008 (392 x 449)	327	1.9E-3
$Phys_1$	lanl.arXiv.org	title	Real/No	388,080 (588 x 660)	88	2.3E-4
$Phys_2$	lanl.arXiv.org	title	Real/Artificial	388,080 (588 x 660)	88	2.3E-4
PES_1	Istat	surname name sex birth.dd birth.mm birth.yyyy	Real/Real	1,033,272 (1,016 x 1,017)	984	9.5E-4
PES_2	Istat	surname name sex birth.dd birth.mm birth.yyyy	Real/Real	4,044,040 (2,002 x 2,020)	1,954	4.8E-4
PES_{full}	Istat	surname name sex birth.dd birth.mm birth.yyyy	Real/Real	32,876,434,096 (180,133 x 182,512)	172,621	5.3E-6

The **Restaurants** source, available at the RIDDLE[16] repository, contains restaurant records affected by real-world errors. It is used for two different RL tasks, **Rest$_1$** and **Rest$_2$**. They differ in the number of matching variables: **Rest$_1$** uses 4 variables, `name`, `address`, `city` and `type`, while **Rest$_2$** relies only on the `name` variable.

Both **Parks** and **Cens** sources are provided by the SECONDSTRING package.[17] Records in the **Parks** RL instance represent U.S. National Parks and the `name` of the park is used as the only matching variable. The **Cens** source, originally provided by William Winkler, contains synthetic census-like records; the corresponding RL instance relies on 5 matching variables: `surname`, `name`, `midinit`, `number` and `street`.

The **Physics** source refers to two partially overlapping selections of scientific papers in the field of high-energy physics.[18] These selections stem from two queries that have been intentionally designed to retrieve papers with very similar titles (even when papers are different). Accordingly, the **Physics** source is used for two different RL tasks, **Phys$_1$** and **Phys$_2$**, both constrained to adopt the `title` field as the only matching variable. The **Phys$_1$** task involves the original clean data, whereas artificially generated random errors have been introduced in **Phys$_2$**. This has been accomplished as follows. Each record from both data sets had a probability of 1/3 to be perturbed; for each selected record, first a number ν, ranging from 0 to the length of its `title` string λ, has been drawn from a Binomial distribution with size λ and success probability 1/6; then a random sample of ν characters drawn from the original `title` has been replaced by ν new random character values. The overall proportion of perturbed characters is about 6%.

PES$_1$, **PES$_2$** and **PES$_{full}$** involve data coming from the Post Enumeration Survey (PES) carried out by the Italian National Institute of Statistics to estimate the coverage rate of the 2001 population Census. Therefore, all the three RL instances deal with real-world data affected by real-world errors, including missing values. Each one of these RL tasks entails the matching of two lists of people, the first collected by the Census and the second by the PES; moreover for all the three RL tasks the same 6 matching variables are used: `surname`, `name`, `sex`, `birth.dd`, `birth.mm` and `birth.yyyy`. Both **PES$_1$** and **PES$_2$** tasks deal with a sample of enumeration areas belonging to the province of Rome, while **PES$_{full}$** faces the RL problem for the *whole* PES data.

As Table 1 shows clearly, **PES$_{full}$** represents a severe test-bed for assessing the practical feasibility of our methods when very large data sets are involved. Being the comparison-space so huge (about 33 *billions* of pairwise distances), a preliminary *blocking* step has been performed. The enumeration area code was selected as blocking variable. Since this variable was believed to be accurate (i.e. almost not affected by errors), the blocking step was expected to quickly filter-out pairs belonging, with high probability, to the U class. From a computational complexity point of view, the net result of the blocking phase was to transform the original, global RL task (which was not affordable) into a sequence of

[16] http://www.cs.utexas.edu/users/ml/riddle/index.html
[17] http://www.cs.utexas.edu/users/ml/riddle/data/secondstring.tar.gz
[18] http://xxx.lanl.gov/find/hep-ph, queries issued on 03/19/2009:
Keyword query 1 = abstract:(QCD and infrared)
Keyword query 2 = abstract:(QCD and confinement)

smaller, independent RL subtasks, one for each block. The overall number of processed blocks was 1,098. Correspondingly, the size of the comparison-space decreased from about 33 *billions* to about 86 *millions* pairs.

6.3 Results

For all the 9 instances described above, we ran our system choosing to solve the RL problem with the Maximum Likelihood objective. Though our system is able to deal also with the Minimum Cost objective, we did not consider this possibility in the experiments because: *i)* such a choice would have negatively affected the understanding and the comparability of our results; *ii)* the specification of misclassification costs is an application specific task.

The results are collectively shown in Table 2. It has to be stressed that, since $\mathbf{PES_{full}}$ is the only instance for which we used our RL system after a preliminary blocking step, and as every comparison-space reduction technique is a possible source of bias in the RL results, we excluded $\mathbf{PES_{full}}$ from the computation of the average performances reported in Table 2.

Table 2 - Precision, Recall and F-measure Results for the Proposed RL System (SYS)

	Precision			Recall			F-measure		
RL Instance	**OTFS**	**SYS**	Λ_{Prec}	**OTFS**	**SYS**	Λ_{Rec}	**OTFS**	**SYS**	Λ_F
$Rest_1$	0.941	0.933	**99.1%**	0.857	0.866	**101.0%**	0.897	0.898	**100.1%**
$Rest_2$	0.988	0.793	**80.2%**	0.759	0.786	**103.5%**	0.859	0.789	**91.9%**
Parks	0.934	0.971	**103.9%**	0.923	0.960	**103.9%**	0.929	0.965	**103.9%**
Cens	0.859	1.000	**116.4%**	0.911	0.982	**107.8%**	0.884	0.995	**112.6%**
$Phys_1$	1.000	0.854	**85.4%**	1.000	1.000	**100.0%**	1.000	0.921	**92.1%**
$Phys_2$	0.964	0.907	**94.1%**	0.909	1.000	**110.0%**	0.936	0.951	**101.7%**
PES_1	0.969	0.998	**102.9%**	0.997	0.996	**99.9%**	0.983	0.997	**101.4%**
PES_2	0.993	0.997	**100.4%**	0.984	0.996	**101.2%**	0.988	0.997	**100.8%**
PES_{full}	-	0.999	-	-	0.992	-	-	0.996	-
Average* Performance	0.956	0.932	**97.8%**	0.918	0.948	**103.4%**	0.934	0.939	**100.6%**

(*): Average values have been computed excluding the $\mathbf{PES_{full}}$ instance (see text).

A first look to the average Precision (0.932), Recall (0.948), and F-measure (0.939) performance immediately reveals the remarkable effectiveness of our system. Moreover, our systems exhibits also a very good robustness: Precision, Recall, and F-measures scores *never* significantly fall below 0.8, despite the addressed RL instances where deliberately selected to be very different.

Turning the attention to the "distance-independent" version of the three quality metrics (Λ columns, bold figures in the table), we observe that *i)* their average values are impressively high, Λ_{Prec} = 97.8%, Λ_{Rec} = 103.4%, Λ_F = 100.6%, and *ii)* all of them *never* fall below 80%. Again, these results strongly support the robustness of our methods. It is also interesting to note that the F-measure scores achieved by our system even outperform the OTFS classifier in 6 cases out of 8.

The **Rest$_2$** instance gives us the opportunity to compare our system with at least one previous proposal. Indeed, **Rest$_2$** exactly corresponds to one of the several RL instances considered in (Chaudhuri et al., 2005): same data sets (**Restaurants**), same matching variables (just `name`) and same distance function (Levenshtein distance). For this RL instance, authors of (Chaudhuri et al., 2005) present a Precision vs. Recall graph obtained when varying the parameters of their algorithms. Though the corresponding quality scores are not explicitly provided, it is possible to deduce from the aforementioned graph that, by fine-tuning parameters, their methods reach a maximum F-measure of about 0.54. We point out that our parameter-free system achieves an F-measure score of 0.789 for **Rest$_2$**, which means a relative gain in effectiveness of nearly 50%.

Moreover, our system also exhibits a very satisfactory behavior for the **PES$_{full}$** instance, from both the points of view of *effectiveness*[19] and *computational efficiency*. Indeed, on the one hand, the obtained Precision (0.999), Recall (0.992), and F-measure (0.996) scores are excellent. On the other hand, the run time performance of our system turned out to be very good. The overall execution time for processing about 86 millions of pairs, partitioned into 1,098 blocks, was 293 minutes (i.e. less than 5 hours) corresponding to an average processing time of about 2×10^{-4} seconds per pair.

7 Conclusions

In this paper, we presented a novel approach to the RL problem based on mixture models. Several original contributions enable our methods to be at the same time effective and fully automated. We validated our suite of methods by testing its Precision, Recall and F-measure scores on real data sets, obtaining excellent results. Our extensive experimental study, which deliberately involved very different RL instances, also showed the remarkable robustness of our methods.

[19] For the **PES$_{full}$** instance, running the OTFS was computationally unfeasible (whence the lacking scores in Table 2). This is because the OTFS cannot be used after a blocking step: indeed, by definition, it has to process the distance distribution of all the pairs as a whole.

References

Armstrong, J. and Mayda, J., "Estimation of Record Linkage Models Using Dependent Data", *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1992.

Belin, T. and Rubin, D., "A Method for Calibrating False-Match Rates in Record Linkage", *Journal of the American Statistical Association*, 90, 1995.

Bender, C. and Orszag, S., *Advanced Mathematical Methods for Scientists and Engineers: Asymptotic methods and perturbation theory*, Springer, New York, 1999.

Chaudhuri, S. and Ganti, V. and Motwani, R., "Robust Identification of Fuzzy Duplicates", *Proceedings of the International Conference on Data Engineering*, 2005.

Christen, P., "A two-step classification approach to unsupervised record linkage", *Proceedings of the Australasian Data Mining Conference*, 2007.

Christen, P. and Goiser, K., "Quality and complexity measures for data linkage and deduplication", In F. Guillet and H. Hamilton, editors, *Quality Measures in Data Mining*, Springer Studies in Computational Intelligence, 2007.

Dempster, A. and Laird, N. and Rubin, D., "Maximum-likelihood from Incomplete Data via the EM Algorithm", *Journal of the Royal Statistical Society*, *SERIES B*, 39(1), 1977.

Duda, R. and Hart, P. and Stork, D., *Pattern Classification*, John Wiley & Sons, 2000.

Elmagarmid, A. and Ipeirotis, P. and Verykios, V., "Duplicate Record Detection: A Survey", *IEEE Transactions on Knowledge and Data Engineering*, 19(1), 2007.

Fellegi, I. and Sunter, A., "A Theory for Record Linkage", *Journal of the American Statistical Association*, 64, 1969.

Fortini, M. and Liseo, B. and Nuccitelli, A. and Scanu, M., "On Bayesian Record Linkage", *Research in Official Statistics*, 4(1), 2001.

Guha, S. and Koudas, N. and Marathe, A. and Srivastava, D., "Merging the Results of Approximate Match Operations", *Proceedings of the International Conference on Very Large Databases*, 2004.

Jaro, M., "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida", *Journal of the American Statistical Association*, 84, 1989.

Larsen, M., "Modeling issues and the use of experience in record linkage", *Proceedings of the Federal Committee on Statistical Methodology*, Record Linkage Workshop, 1997.

Larsen, M. and Rubin, D., "Iterative Automated Record Linkage Using Mixture Models", *Journal of the American Statistical Association*, 96(453), 2001.

Larsen, M., "Hierarchical Bayesian Record Linkage Theory", *Proceedings of the Survey Research Methods Section*, American Statistical Association, 2005.

McLachlan, G. and Basford, K., *Mixture Models: Inference and Applications to Clustering*, M. Dekker, New York, 1988.

McLachlan, G. and Peel, D., *Finite Mixture Models*, John Wiley & Sons, 2000.

Michalewicz, Z., *Genetic Algorithms + Data Structures = Evolution Programs*, Springer, Berlin, 1996.

Ospina, R. and Ferrari, S. L. P., "Inflated Beta Distributions". *Statistical Papers*, 51-1, Springer, Berlin, 2010.

R Development Core Team (2009), "R: A Language and Environment for Statistical Computing". *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Tejada, S. and Knoblock, C. and Minton, S., "Learning object identification rules for information integration", *Information Systems*, 26(8), 2001.

Verykios, V. and Elmagarmid, A. and Houstis, E., "Automating the approximate record-matching process", *Journal of Information Sciences*, 126(1-4), 2000.

Verykios, V. and Moustakides, G. and Elfeky, M., "A Bayesian Decision Model for Cost Optimal Record Matching", *VLDB Journal*, 12(1), 2003.

Winkler, W., "Improved decision rules in the Fellegi-Sunter model of record linkage", *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1993.

Winkler, W., "Advanced Methods for Record Linkage", *Proceedings of the Survey Research Methods Section*, American Statistical Association, 1994.

Proposta per una metodologia di stima dell'impermeabilizzazione del suolo in Italia[1]

Michele Munafò,[2] Gianluigi Salvucci,[3] Marco Zitti[4] e Luca Salvati[5]

Sommario

L'impermeabilizzazione del suolo (meglio conosciuta con il termine inglese di soil sealing) deve essere intesa come un costo ambientale, risultato di una diffusione indiscriminata delle tipologie artificiali di uso del suolo che porta alla perdita dell'equilibrio ecosistemico. Questo lavoro propone una procedura di monitoraggio permanente della cementificazione attraverso una metodologia campionaria basata sulla foto-interpretazione di ortofoto e carte topografiche storiche. La stima diacronica della percentuale di suolo impermeabilizzato in Italia evidenzia una dinamica complessa, legata sia alla crescita urbana compatta propria del secondo dopoguerra, sia alla diffusione pervasiva tipica degli anni più recenti.

Parole chiave: Impermeabilizzazione del suolo, consumo di suolo, crescita urbana, sprawl, indicatori ambientali, Italia.

Abstract

Soil sealing may be considered as a negative externality of the economic growth, because the sealed soil loses several of its ecological functions. The analysis of land cover changes carried out by traditional data sources, allows only a relatively rough estimation of soil sealing, since it is a process associated, although with different rates, to both urban and semi-natural land use categories. The aim of this paper is to illustrate a sampling procedure quantifying the soil sealing rate over time in Italy. The procedure was based on the visual interpretation of 12,000 plots from aerial photographs and land cover maps at different times (1956, 1994, 1999, 2006). Results indicate that soil sealing was continuously increasing in Italy throughout the investigated period. The highest rate was found in northern Italy. The implications such dynamics have on planning strategies aimed at containing urban sprawl are finally discussed.

Parole chiave: Soil sealing, land consumption, urban sprawl, indicators, Italy.

---

[1] Gli autori ringraziano C. Abbate, e P. Napolitano (Istat) per la revisione critica del manoscritto. Il supporto tecnico di A. Strollo, C. Norero, C. Santilli (Università di Roma 'La Sapienza') si è rivelato prezioso nella redazione del lavoro. Grazie anche a S. Tersigni (Istat), R. Gemmiti (Università di Roma 'La Sapienza') e L. Perini (Cra-Cma) per aver discusso con gli autori sugli aspetti di ricerca più innovativi del *soil sealing* e dell' *urban sprawl*.
Sebbene il lavoro sia frutto dell'opera di tutti gli autori, sono da attribuire: i paragrafi 1, 2 e 4 a Michele Munafò e Luca Salvati, mentre il paragrafo 3 a Gianluigi Salvucci. L'apparato tabellare e grafico è opera di Marco Zitti.

[2] Ricercatore (Ispra), e-mail: michele.munafo@isprambiente.it

[3] Collaboratore Tecnico (Istat), e-mail: gianluigi.salvucci@istat.it

[4] Assegnista di ricerca (Cra-Cma), e-mail: mzitti@ucea.it

[5] Ricercatore (Cra-Rps), e-mail: lsalvati@entecra.it

1. Introduzione

Il suolo è una preziosa risorsa naturale. Seppure apparentemente inerte, esso deve essere considerato un elemento da preservare, al pari degli esseri viventi, come componente essenziale per l'equilibrio della biosfera. La componente edafica dell'ecosistema, infatti, regola i cicli nutrizionali indispensabili per la vegetazione, che è posta alla base della catena alimentare. La formazione del suolo è legata al lento processo di decomposizione fisica ed organica delle rocce superficiali della crosta terrestre. Ciò suggerisce che tale supporto deve essere considerato come una risorsa naturale finita: il tempo di distruzione di questa delicata componente è, infatti, brevissimo (Barberis, 2005).

Il suolo è visto sovente come spazio astratto, da occupare con processi di urbanizzazione, privi troppo spesso di una visione ecosistemica. Sebbene la distruzione o la degradazione del suolo avvenga sovente tramite altri processi, quali varie forme di inquinamento, dispersione di rifiuti tossici, concentrazione di liquami, salinizzazione, erosione e compattazione, l'impermeabilizzazione rappresenta una forma di degrado silente, spesso non percepita come tale. Un terreno impermeabilizzato influenza negativamente il clima urbano, aumenta la quantità e la velocità delle acque di scorrimento superficiale e i conseguenti fenomeni erosivi (Johnson, 2001; Hough, 2004). Diventa perciò necessario prendere atto dell'incremento del suolo impermeabilizzato quale misura di un processo di potenziale degrado ambientale.

Il suolo è storicamente inteso dalle discipline economico-territoriali come un tradizionale fattore produttivo. Al pari degli altri fattori, quando il suo valore, in questo caso i vantaggi derivanti dalle economie di urbanizzazione, diventa rilevante, ne deriva un aumento di domanda. Inoltre, la ricerca di una maggior qualità abitativa in termini di tipologie edilizie a bassa densità, la necessità di spazi da destinare a nuove infrastrutture di trasporto e la forte crescita dei valori immobiliari hanno contribuito allo sviluppo delle città riducendo la concentrazione e incrementando le aree con livelli di urbanizzazione a minore densità abitativa a distanze progressivamente maggiori dai centri urbani. Si può sostenere, quindi, che una parte consistente dell'attività edilizia e della crescita suburbana sia fortemente legata alla valorizzazione della rendita fondiaria e alla necessità di capitalizzare risorse economiche attraverso l'attività immobiliare (Insolera, 1993; Gibelli e Salzano, 2006; Berdini, 2010). In questo processo il suolo diventa fattore produttivo di nuove economie di urbanizzazione, viene occupato (consumato) e distrutto (impermeabilizzato) nel processo dello *sprawl* (Bruegmann, 2005).

Parallelamente alle prassi di pianificazione consolidata, si riproducono tutt'ora meccanismi di consumo non completamente controllabile del suolo attraverso abusi edilizi e, soprattutto, la deregolazione della pianificazione stessa. La situazione appare potenzialmente grave in Italia, dove un importante quesito da sciogliere è relativo al tasso di consumo sostenibile di suolo. Poiché appare inevitabile che ogni attività umana occupi spazio, quale soluzione può essere adottata nel medio e lungo termine per conciliare la protezione dell'ambiente con lo sviluppo economico e sociale del territorio? Legambiente, in collaborazione con il Politecnico di Milano, ha avanzato una proposta di legge con due obiettivi essenziali (Pileri, Lanzani, 2007): (i) limitare l'uso edificatorio del suolo evitando che esso diventi un deposito (confuso) di manufatti spesso sottoutilizzati e abbandonati e che i livelli di urbanizzazione in alcune porzioni del territorio raggiungano livelli insostenibili; (ii) legare ogni attività di urbanizzazione ad una contestuale attività di valorizzazione dell'ambiente negli spazi aperti limitrofi. L'ottica è, dunque, quella di bilanciare le attività umane che portano alla distruzione del suolo con attività esogene di riqualificazione, che accelerino e/o fortifichino i processi di riequilibrio naturale anche attraverso meccanismi di compensazione preventiva (Pileri, 2007).

L'impermeabilizzazione del suolo (meglio conosciuta con il termine inglese di *soil sealing*) deve essere quindi intesa come un costo ambientale, risultato di una diffusione indiscriminata delle tipologie artificiali di uso del suolo che porta al degrado delle funzioni ecosistemiche e, in definitiva, all'alterazione dell'equilibrio ecologico. In tale contesto, e nell'ottica di una valutazione del fenomeno finalizzata al riequilibrio delle attività di sviluppo a livello territoriale, si ritiene necessaria una misurazione attendibile del consumo di suolo e della contemporanea sua progressiva impermeabilizzazione.

L'analisi delle tipologie di copertura del suolo, attraverso mezzi divenuti ormai tradizionali, non consente tuttavia di cogliere appieno il fenomeno dell'impermeabilizzazione che risulta invece trasversale alle diverse classi. Infatti, se la probabilità di trovare suoli impermeabilizzati in un tessuto urbano compatto è elevata, non si può non considerare l'impatto generato da piccole porzioni di suolo impermeabilizzato nelle altre classi di copertura la cui estensione è spesso inferiore alla minima unità rilevabile dalla cartografia. Da questa esigenza nasce l'interesse verso l'applicazione di una metodologia campionaria attraverso cui stimare, diacronicamente e per aree vaste, la percentuale di suolo impermeabilizzato, svincolandosi dalla scarsa disponibilità di carte di copertura del suolo sufficientemente dettagliate, frequentemente aggiornate e con caratteristiche omogenee sull'intero territorio nazionale.

Questo lavoro propone, a livello esplorativo, una procedura di monitoraggio permanente dell'impermeabilizzazione attraverso una metodologia campionaria basata sulla foto-interpretazione di ortofoto e carte topografiche storiche. La stima dell'impermeabilizzazione del suolo e la sua evoluzione nel tempo supporta la presa di coscienza del problema e consente di tracciare linee di intervento nelle politiche territoriali di sviluppo: essa è stata condotta in questo studio per ripartizione geografica e consente di evidenziare dinamiche complesse di occupazione del suolo. Completa l'analisi una valutazione dei principali determinanti della distribuzione geografica del *soil sealing* in Italia, quali la distanza dal mare, l'altimetria, e la distanza dai centri urbani più importanti (e.g. Hasse e Lathrop, 2003, Frenkel e Ashkenazi, 2007, Schneider e Woodcock, 2008).

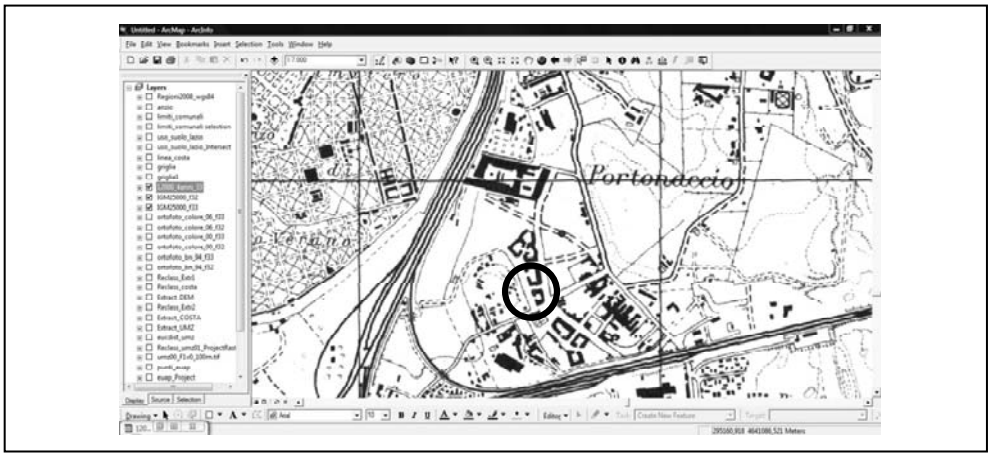
2. Materiali e metodi

La metodologia proposta prevede la verifica diacronica dello status di sigillamento attraverso una variabile dicotomica (0: non sigillato; 1: impermeabilizzato) ottenuta mediante foto-interpretazione di un campione particolarmente ampio di siti distribuiti omogeneamente sul territorio Italiano.[6] A questo scopo si è resa necessaria la foto-interpretazione a diverse epoche (1956, 1994, 1999, 2006) di numerose ortofoto e altro materiale cartografico raffigurante, secondo formati omogenei, la situazione dei singoli punti di rilievo nelle varie epoche di indagine (Figure 1 e 2). Il materiale cartografico ed orto-fotografico a disposizione è dato dalla cartografia dell'Istituto Geografico Militare (Firenze), a scala 1:25.000 (serie 25/V), con copertura nazionale e databile fra il 1949 ed il 1962 (a riguardo, nell'analisi si è assegnata a tale cartografia una datazione media corrispondente all'anno 1956), nonché da

[6] I codici ai punti sono stati assegnati sulla base del seguente criterio: (0) suolo permeabile per boschi, prati e altre aree naturali, aree agricole, aree aperte, giardini privati, parchi, aiuole cittadine, corpi idrici (escluso il mare), etc.; (1) suolo impermeabile per edifici, capannoni, cortili e aree pavimentate, piazzali, parcheggi, strade, ferrovie, campi da calcio, cave, cantieri, discariche, serre, etc; (n.v.) non valutabile per i punti non fotointerpretabili a causa, ad esempio, della obliterazione delle ortofoto in alcune aree militari o sensibili, o per i punti ricadenti in mare e in aree lagunari.

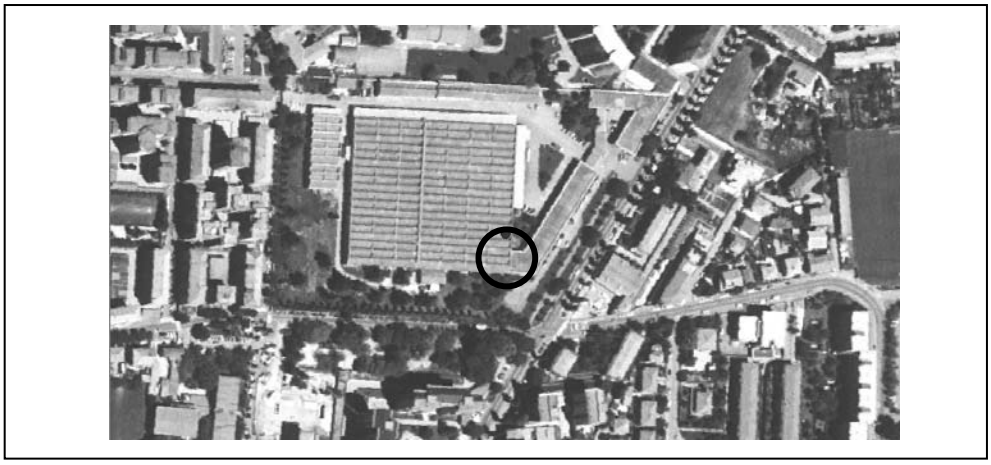
fotografie aeree policromatiche ad alta risoluzione spaziale (1994: ortofoto AIMA; 1999: volo IT2000 - Compagnia Generale Ripreseaeree, acquisito negli anni 1998 e 1999; 2006: Ministero dell'Ambiente, foto acquisite nel periodo 2005-2007).

Figura 1 - Esempio di foto-interpretazione della cartografia IGM (1956), il punto evidenziato viene classificato come impermeabile



Fonte: Elaborazione su mappe IGM

Figura 2 - Esempio di foto-interpretazione dell'ortofoto del 1994, il punto evidenziato viene classificato come impermeabile



Fonte: Elaborazione su ortofoto AIMA

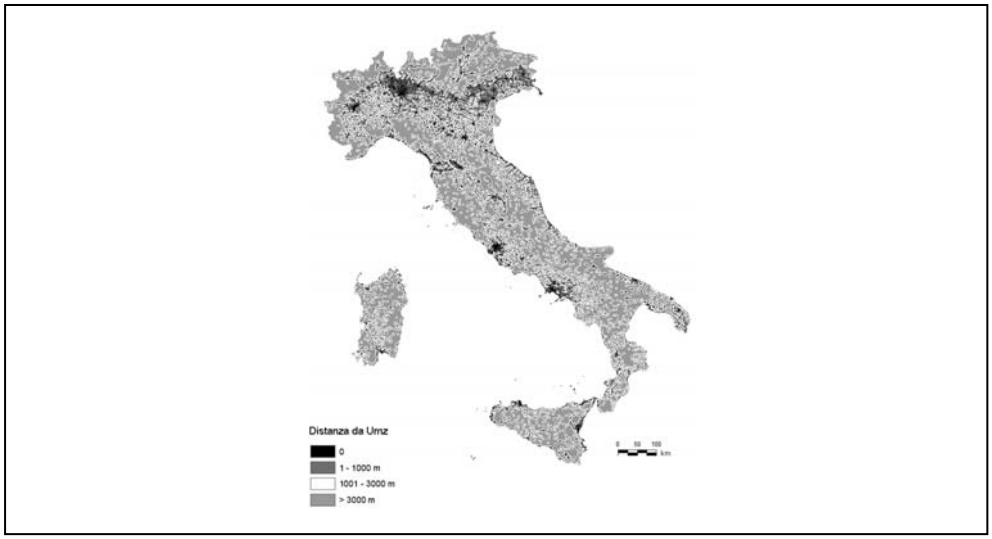
Il campione, formato da 12.000 punti scelti in modo casuale all'interno di celle generate con un reticolo sistematico di passo di 5 km, garantisce una copertura omogenea e rappresentativa dell'intero territorio nazionale e coincide con la rete di punti di validazione a terra predisposta da ISPRA per il monitoraggio dell'uso del suolo nazionale e per la verifica dei dati del progetto Corine Land Cover (Maricchiolo et al., 2005).

Tale scelta, congrua con gli obiettivi di parsimonia e flessibilità dell'applicazione esplorativa qui presentata, è finalizzata a garantire la massima integrabilità spaziale e temporale con altre rilevazioni ed informazioni raccolte nell'ambito delle iniziative Corine consentendo, inoltre, di disporre di informazioni e di immagini rilevate direttamente *in situ*. La dimensione campionaria è idonea a produrre risultati a scala nazionale e ripartizionale.

Per evitare l'errore di stima dovuto alla scelta di un'unità minima cartografata, particolarmente evidente nell'analisi delle dinamiche di cambiamento di coperture del suolo molto frammentate come quelle in questione, si è optato per una foto-interpretazione multitemporale su base puntuale e non areale, distinguendo tra terreni permeabili ed impermeabili in corrispondenza del punto e utilizzando scale di lavoro comprese tra 1:1.000 e 1:4.000. Il controllo di qualità finale è consistito in sessioni di foto-interpretazione separate, a scale di ulteriore dettaglio (1:1.000 o superiori) sui punti che risultavano con codifica cambiata, oltre che su un sottoinsieme casuale dei punti stabili, per un totale di circa il 10% dei punti. I limiti fiduciali del campione sono stati calcolati e presentati con un livello di probabilità del 99%. Gli indicatori stimati dai dati elementari includono la percentuale di sigillamento, la stima intervallare di tale variabile (intesa come ampiezza percentuale dell'intervallo di confidenza al 99%), i tassi di incremento annuo dei punti impermeabilizzati nonché dell'impermeabilizzazione pro-capite, in base alla popolazione residente rilevata ad ogni tempo *t* di rilevazione.

La distribuzione spaziale dei punti impermeabili è stata, inoltre, studiata rispetto ad alcune variabili fisiografiche, quali l'altimetria, la distanza dal mare e dalle aree urbane (valutate attraverso il confine degli Agglomerati morfologici urbani (Amu), cfr. Istat 2009, indipendentemente dalla loro dimensione demografica) mediante specifica applicazione GIS. La suddivisione in Amu ha l'obiettivo di individuare porzioni di territorio con caratteristiche urbane così come sono state definite dalle direttive Unece/Eurostat (1998). Un esempio di tale analisi, relativo alla distanza dalle aree urbane, è riportato nella figura 3.

Figura 3 - Classificazione del territorio italiano in base alla distanza dagli Amu.



Fonte: elaborazione su dati European Environment Agency (2007)

3. Risultati e discussione

L'analisi dei dati ottenuti dalla foto-interpretazione evidenzia un aumento dell'impermeabilizzazione in Italia durante tutto il periodo di studio, sintomo della presenza di fenomeni di consumo del suolo piuttosto costanti nel tempo (Tabella 1). Il tasso di impermeabilizzazione, pari al 2,4% dei punti campionati nel 1956, si attesta al 6,3% nel 2006. Il primo periodo (1956-1994) è legato alla crescita urbana propria del secondo dopoguerra. Il periodo successivo (1994-2006) è caratterizzato, invece, da uno sviluppo urbano diffuso, tipico delle aree peri-urbane costiere e pianeggianti. Le variazioni annue maggiori si registrano proprio nell'ultimo intervallo temporale considerato, quello compreso fra il 1999 ed il 2006 (Tabella 2).

Tabella 1 - Statistiche sui punti-campione per la rilevazione dell'impermeabilizzazione dei suoli in Italia per ripartizione geografica – Anni 1956-2006

Ripartizioni geografiche	Superficie [ha]	Punti valutati	Punti impermeabilizzati			
			1956	1994	1999	2006
Italia nord-occidentale	5.792.023	2.291	73	147	154	168
Italia nord-orientale	6.200.906	2.469	63	138	150	165
Italia centrale	5.840.863	2.320	51	121	122	145
Italia meridionale	7.379.536	2.899	59	139	141	174
Italia insulare	4.991.924	1.936	37	86	88	104
Italia nel complesso	**30.205.252**	**11.915**[7]	**283**	**631**	**655**	**756**

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

L'incremento di suolo impermeabilizzato risulta, inoltre, sproporzionato a quello della popolazione: ciò suggerisce che, nei periodi più recenti, lo sviluppo peri-urbano si è realizzato attraverso insediamenti a bassa densità abitativa e ad alta intensità infrastrutturale, con un effetto netto sul consumo di suolo ancora più marcato rispetto a quello esercitato dalla crescita urbana compatta e ad alta densità tipica dei primi decenni del secondo dopoguerra.

Tabella 2 - Indicatori di impermeabilizzazione dei suoli in Italia – Anni 1956-2006

Indicatore	1956	1994	1999	2006
Percentuale di punti-campione impermeabili	2,37	5,29	5,49	6,34
Incremento percentuale annuo di punti impermeabili	-	0,08	0,04	0,11
Incremento % annuo di punti impermeabili *pro-capite*	-	0,11	0,43	0,89

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Il trend rilevato a livello nazionale può essere approfondito attraverso un'analisi condotta per ripartizione geografica (Tabelle 3 e 4). I dati a livello ripartizionale rilevano, come atteso, una maggiore intensità dell'uso del suolo impermeabilizzato nel nord Italia (Tabella 5). Tuttavia, normalizzando i tassi rispetto alla variazione della popolazione residente, i tassi di crescita risultano più elevati nelle aree meridionali del paese, soprattutto nell'ultimo periodo considerato (Tabella 6). Tale fenomeno appare significativo perché

[7] Dai 12.000 punti del campione sono stati eliminati quelli ricadenti in mare.

osservato in aree caratterizzate da insediamenti rurali a bassa densità e da insediamenti urbani compatti, in grado di esercitare, almeno in origine, un limitato consumo di suolo.

Tabella 3 - Percentuale di punti impermeabili in Italia per ripartizione geografica – Anni 1956-2006

Ripartizioni geografiche	1956	1994	1999	2006
Italia nord-occidentale	3,2	6,4	6,7	7,3
Italia nord-orientale	2,6	5,6	6,1	6,7
Italia centrale	2,2	5,2	5,3	6,3
Italia meridionale	2,0	4,8	4,9	6,0
Italia insulare	1,9	4,4	4,5	5,4
Italia nel complesso	**2,4**	**5,3**	**5,5**	**6,3**

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Tabella 4 - Ampiezza percentuale dell'intervallo di confidenza al 99% associato alla stima percentuale dei punti impermeabilizzati in Italia per ripartizione geografica – Anni 1956-2006

Ripartizioni geografiche	1956	1994	1999	2006
Italia nord-occidentale	2,2	3,1	3,1	3,3
Italia nord-orientale	1,9	2,8	2,9	3,0
Italia centrale	1,8	2,8	2,8	3,0
Italia meridionale	1,6	2,4	2,4	2,6
Italia insulare	1,9	2,8	2,8	3,1
Italia nel complesso	**0,8**	**1,2**	**1,3**	**1,3**

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Tabella 5 - Incremento percentuale annuo dei punti impermeabilizzati in Italia per ripartizione geografica – Anni 1956-2006

Ripartizioni geografiche	1956 - 1994	1994 - 1999	1999 - 2006	1956 – 2006
Italia nord-occidentale	0,08	0,06	0,09	0,08
Italia nord-orientale	0,08	0,10	0,09	0,08
Italia centrale	0,08	0,02	0,14	0,08
Italia meridionale	0,07	0,02	0,16	0,08
Italia insulare	0,07	0,02	0,13	0,07
Italia nel complesso	**0,08**	**0,04**	**0,11**	**0,08**

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Tabella 6 - Incremento percentuale annuo dei punti impermeabili *pro-capite* in Italia per ripartizione geografica – Anni 1956-2006

Ripartizioni geografiche	1956 - 1994	1994 - 1999	1999 - 2006	1956 – 2006
Italia nord-occidentale	0,09	0,48	0,47	0,35
Italia nord-orientale	0,13	1,14	0,69	0,65
Italia centrale	0,11	0,10	1,04	0,42
Italia meridionale	0,10	0,14	1,21	0,48
Italia insulare	0,13	0,32	1,26	0,57
Italia nel complesso	**0,11**	**0,42**	**0,89**	**0,47**

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

La distribuzione del suolo impermeabilizzato è stata successivamente qualificata rispetto a tre gradienti: la distanza dalle aree urbane, dalla linea di costa e l'altimetria. L'analisi dei punti ricadenti negli Amu (cfr. Figura 3) ha permesso di individuare l'andamento diacronico dell'impermeabilizzazione del suolo in funzione della distanza dai centri urbani. Come atteso, tale processo assume nel tempo un andamento decrescente basato sulla distanza dalle città (Tabella 7).

Tabella 7 - Statistiche sui punti-campione per la rilevazione dell'impermeabilizzazione dei suoli in Italia in base alla distanza dagli Amu – Anni 1956-2006

Distanza (km)	Superficie (ha)	Punti valutati	Punti impermeabilizzati			
			1956	1994	1999	2006
< 1	7.847.793	3.026	175	431	449	503
1 – 3	10.866.324	4.351	72	138	144	176
> 3	11.474.758	4.538	36	62	62	77

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Questo andamento viene pienamente confermato dall'analisi diacronica (Tabelle 8 e 9). Si debbono rilevare due andamenti, come già rilevato precedentemente. Da una parte l'intensificazione urbana, che porta ad oltre il 16% di suolo impermeabilizzato nel 2006 (a partire dal 6% osservato nel 1956). Dall'altra, un massiccio fenomeno di sigillamento nella fascia limitrofa alle aree urbane (fascia 1 – 3 km), con un aumento dei punti impermeabilizzati dal 1% del 1956 al 4% nel 2006.

Tabella 8 - Percentuale dei punti impermeabili in Italia per distanza dagli Amu – Anni 1956-2006

Distanza (km)	1956	1994	1999	2006
< 1	5,8	14,2	14,8	16.6
1 – 3	1,7	3,2	3,3	4,0
> 3	0,8	1,4	1,4	1,7

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Tabella 9 - Ampiezza percentuale dell'intervallo di confidenza al 99% per la stima della percentuale dell'impermeabilizzazione in Italia per distanza dagli Amu – Anni 1956-2006

Distanza (km)	1956	1994	1999	2006
< 1	2,5	3,8	3,9	4,1
1 – 3	1,2	1,6	1,6	1,8
> 3	0,8	1,0	1,0	1,1

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Tali risultati evidenziano la dinamica di edificazione compatta tipica del primo dopoguerra soprattutto nell'Italia settentrionale e centrale nonché l'espansione urbana diffusa propria degli ultimi due decenni, con un tasso annuo di crescita tornato a livelli elevati dopo una fase di relativo rallentamento (Tabella 10).

Pur non essendo identificabile in maniera univoca una zona costiera in Italia (Munafò, 2008) la distanza dalla linea di costa è un indicatore della prossimità ad un ecosistema caratterizzato da particolari esigenze di conservazione (Tabella 11). Si è delimitata, pertanto, la zona costiera considerandola ricadente in una distanza massima di 10 Km dalla linea di costa, come suggerito dalle linee guida del Progetto Europeo Lacoast (Perdigao e Christensen 2000), ottenendo così una ripartizione dell'intero territorio in due fasce: da 0 a 10 chilometri di distanza dalla linea di costa e a distanze superiori a 10 chilometri dalla riva (Tabelle 12 e 13).

Tabella 10 - Incremento percentuale annuo dei punti impermeabili in Italia per distanza dagli Amu – Anni 1956-2006

Distanza (km)	1956 - 1994	1994 - 1999	1999 - 2006	1956 - 2006
< 1	0,22	0,12	0,26	0,22
1 – 3	0,04	0,02	0,10	0,05
> 3	0,02	0,00	0,04	0,02

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Tabella 11 - Statistiche sui punti-campione per la rilevazione dell'impermeabilizzazione dei suoli in Italia per distanza dalla linea di costa – Anni 1956-2006

Distanza (km)	Superficie (ha)	Punti valutati	Punti impermeabilizzati			
			1956	1994	1999	2006
< 10	4.980.537	1.856	58	140	144	170
> 10	25.208.338	10.059	225	491	511	586

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Tabella 12 - Percentuale di punti impermeabili in Italia per distanza dalla linea di costa – Anni 1956-2006

Distanza (km)	1956	1994	1999	2006
< 10	3,1	7,5	7,8	9,2
> 10	2,2	4,9	5,1	5,8

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Tabella 13 - Ampiezza percentuale dell'intervallo di confidenza per la stima della percentuale dell'impermeabilizzazione in Italia per distanza dalla linea di costa – Anni 1956-2006

Distanza (km)	1956	1994	1999	2006
< 10	1,6	2,5	2,5	2,7
> 10	0,6	0,9	0,9	0,9

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

La maggiore superficie impermeabilizzata, come atteso, si registra lungo la fascia costiera, tradizionalmente sottoposta in Italia a fenomeni massivi di littoralizzazione già dal primo dopoguerra, mentre appare più limitata l'espansione delle aree cementificate nella fascia di territorio a maggiore distanza dalla costa (Tabella 14). Il periodo 1999-2006 si conferma, in tutte le aree valutate, come particolarmente critico per la crescita dei punti impermeabilizzati.

Tabella 14 - Incremento annuo percentuale dei punti impermeabili in Italia per distanza dalla linea di costa – Anni 1956-2006

Distanza (km)	1956 - 1994	1994 - 1999	1999 - 2006	1956 - 2006
< 10	0,12	0,06	0,20	0,12
> 10	0,07	0,04	0,10	0,07

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Il terzo gradiente analizzato rispetto all'impermeabilizzazione dei suoli riguarda l'altimetria (Tabelle 15 e 16). Come atteso, tale variabile costituisce di fatto un vincolo all'espansione urbana: è quindi ipotizzabile, nel tempo, un maggiore incremento di edificato nella fascia della pianura rispetto a quella collinare e montuosa. L'analisi dei dati conferma

questa ipotesi: il tasso di impermeabilizzazione passa dal 4% al 10% in pianura con un differenziale di 2 punti percentuali con la collina e 3 punti percentuali con la montagna nel 1956, differenziale che si consolida nel 2006 rispettivamente a 5 e 8 punti (Tabelle 17 e 18).

Tabella 15 - Statistiche sui punti-campione per la rilevazione dell'impermeabilizzazione dei suoli in Italia per fascia altimetrica – Anni 1956-2006

Fascia altimetrica	Quota (m)	Superficie (ha)	Punti valutati	Punti impermeabilizzati			
				1956	1994	1999	2006
Pianura	0 – 300	13.919.253	5.425	197	452	474	550
Collina	300 – 600	6.834.659	2.717	54	118	120	135
Montagna	> 600	9.434.962	3.773	32	61	61	71

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Tabella 16 - Percentuale di punti impermeabili in Italia per fascia altimetrica – Anni 1956-2006

Fascia altimetrica	1956	1994	1999	2006
Pianura	3,6	8,3	8,7	10,1
Collina	2,0	4,3	4,4	5,0
Montagna	0,8	1,6	1,6	1,9

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Tabella 17 - Ampiezza percentuale dell'intervallo di confidenza al 99% per la stima della percentuale dell'impermeabilizzazione in Italia per fascia altimetrica – Anni 1956-2006

Fascia altimetrica	1956	1994	1999	2006
Pianura	1,5	2,2	2,3	2,5
Collina	1,6	2,3	2,4	2,5
Montagna	0,9	1,2	1,2	1,3

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Tabella 18 - Incremento percentuale annuo dei punti impermeabili in Italia per fascia altimetrica – Anni 1956-2006.

Fascia altimetrica	1956 - 1994	1994 - 1999	1999 - 2006	1956 - 2006
Pianura	2,2	1,0	2,1	2,1
Collina	2,1	0,3	1,7	1,9
Montagna	1,7	0,0	2,2	1,6

Fonte: Elaborazione su mappe IGM e ortofoto AIMA, Compagnia Generale Ripreseaeree e Ministero dell'Ambiente

Un confronto tra i dati campionari e la classe 1.1.1 (Zone residenziali a tessuto continuo) della cartografia Corine Land Cover ha permesso di stimarne il grado di impermeabilizzazione e di effettuare una prima validazione dei risultati. Il valore ottenuto (86,5%) è infatti perfettamente congruente con le specifiche tecniche del progetto Corine. Allo stesso tempo, la caratterizzazione di tutte le 44 classi Corine sulla base del grado di impermeabilizzazione ha permesso di dimostrare i limiti, precedentemente esposti, di approcci di valutazione dell'impermeabilizzazione basati esclusivamente sull'uso di dati di copertura del suolo. Si evidenzia, a tal proposito, la presenza significativa di aree artificiali all'interno di classi di copertura normalmente considerate non impermeabilizzate come, ad esempio, le classi 2.2.2 (Frutteti e frutti minori), 2.3.1 (Prati stabili) e 2.4.2 (Sistemi colturali e particellari complessi) dove il *soil sealing* ha addirittura valori intorno al 10%.

4. Conclusioni

Questo lavoro rappresenta un nuovo ed originale contributo allo studio dell'impermeabilizzazione del suolo in Italia, e consente una sua valutazione diacronica di lungo periodo, sia a livello nazionale che ripartizionale. La proposta qui illustrata si basa su una metodologia campionaria che ha presentato una confidenza accettabile alla scala geografica considerata, suggerendo così un suo utilizzo routinario per il monitoraggio su base sia storica che congiunturale del fenomeno, ad esempio tramite una cadenza di rilevazione annuale (Alfsen e Saebo, 1993). Una rilevazione campionaria come quella proposta in questo contributo potrebbe supportare processi di valutazione del consumo di suolo a diverse scale geografiche, in base alla dimensione del campione scelto.

L'attività di foto-interpretazione a vista è stata svolta agevolmente navigando le foto aeree disponibili sul sito del Ministero dell'Ambiente nonchè una cartografia storica a copertura nazionale (la cartografia dell'Istituto Geografico Militare di Firenze, realizzata negli anni '50), utilizzando sostanzialmente materiale di ampia disponibilità e nella totalità dei casi gratuito. Visto il limitato costo di realizzazione, tale rilevazione può assicurare una base informativa piuttosto completa, con copertura nazionale e cadenza periodica. Lo sviluppo di una base informativa tramite supporto di geo-database consente un'immediata integrazione con altre fonti-dati (mappe Corine, rete Lucas, basi territoriali dei censimenti generali, etc.) permettendo, in tal modo, l'impiego del campione statistico anche per la validazione e per la valutazione dell'accuratezza di carte a diversi livelli di scala.

I risultati ottenuti rispecchiano il *pattern* spaziale dell'urbanizzazione Italiana nel secondo dopoguerra, da sempre prevalente nelle aree costiere, di pianura e collinari, attribuibile a motivazioni logistiche e fattori socio-economici tradizionali. Tra l'altro, l'evoluzione dell'uso del suolo, se correlata con le caratteristiche ambientali prevalenti alle diverse altitudini, si riflette in un consumo maggiore delle aree agricole rispetto a quelle boschive, le prime maggiormente diffuse nelle aree costiere e di pianura dove si concentrano i fenomeni di impermeabilizzazione (King et al., 1997; Ludlow et al., 2007).

Ciò conferma, indirettamente, i risultati precedentemente resi disponibili dalla cartografia del Progetto Corine Land Cover per quanto riguarda l'evoluzione recente dei principali usi del suolo in Italia (Gibelli, Salzano, 2006). La lettura dei risultati di questo lavoro è, dunque, pienamente in linea con le tendenze già rilevate nell'ambito di studi cartografici pregressi, ma aggiunge nuove informazioni circa un fenomeno trasversale ai diversi usi del suolo. Tutto ciò suggerisce, da una parte, una buona coerenza delle diverse fonti dati disponibili in Italia, dall'altro conferma la necessità e l'utilità di una specifica rilevazione sull'impermeabilizzazione del suolo.

L'indagine ha anche confermato le ipotesi di partenza circa la domanda di impermeabilizzazione del suolo in Italia dal secondo dopoguerra ad oggi. Questa coincide sostanzialmente con lo sviluppo urbano dapprima compatto e successivamente diffuso, trovando maggiore concentrazione, come già accennato, nelle aree pianeggianti e a ridosso delle zone costiere (Couch et al., 2007). Il nord è la zona con la più elevata percentuale di superficie impermeabile, ma è il sud che registra gli incrementi maggiori di consumo del suolo dal 1956 ad oggi (e.g. Agapito et al., 2009).

Questo contributo mette dunque in evidenza che, a fronte di un notevole tasso di impermeabilizzazione osservato nei primi anni del dopoguerra, il fenomeno del consumo di

suolo è tutt'altro che attenuato negli ultimi anni. E' necessario implementare, dunque, strumenti di monitoraggio permanente che siano in grado, da una parte, di sensibilizzare l'opinione pubblica sul possibile degrado ambientale derivante da un eccesso di cementificazione soprattutto in alcune aree del paese, dall'altro di fornire una base informativa tempestiva ai decisori, a tutti i livelli di *governance* (Giannakourou, 2005). L'auspicio è che simili strumenti possano stimolare politiche di autocontenimento dello *sprawl* urbano in grado di moderare la frammentazione del paesaggio, pur in un'ottica di sviluppo regionale policentrico (Munafò, 2008).

6. Riferimenti bibliografici

Agapito A., Alessi E., Battisti C., Benedetto G., Bologna G., Bulgarini F., Ciacci L., Costantini M., Fantilli P., Ferroni F., Ficorilli S., Fioravanti S., Lenzi S., Martinoja D., Meregalli D., Petrella S., Pratesi I., Rocco M., Romano B., Teofili C., Tosatti V., 2009, *2009: l'anno del cemento – Dossier sul consumo di suolo in Italia*, WWF Italia, Roma.

Alfsen K.H., Saebo H.V., 1993, *Environmental quality indicators: background, principles and examples from Norway*, Environmental and Resource Economics, 3.

Barberis R., 2005, *Consumo di suolo e qualità dei suoli urbani*, Rapporto ARPA Piemonte, Torino.

Bruegmann R., 2005, *Sprawl: a compact history*, Chicago: University of Chicago Press.

Couch C., Petschel-Held G., Leontidou L., 2007, *Urban sprawl in Europe: landscapes, land-use change and policy*, London: Blackwell.

European Environment Agency, 2007, *Urban Morphological Zones 2000*, EEA, Copenhagen,

Frenkel A., Ashkenazi M., 2007, *The integrated sprawl index: measuring the urban landscape in Israel*, Annals of Regional Science, 42(1), 99-121.

Giannakourou G., 2005, *Transforming spatial planning policy in Mediterranean countries: Europeanization and domestic change*, European Planning Studies, 13(2), 319 – 331

Gibelli M.C., Salzano E., 2006, *No Sprawl. Perché è necessario controllare la dispersione urbana e il consumo di suolo*, Alinea, Firenze.

Hasse J.E., Lathrop R.G., 2003, *Land resource impact indicators of urban sprawl*, Applied Geography, 23, 159-175.

Hough M., 2004, *Cities and Natural Process*, Routledge, London.

Insolera I., 1993, *Roma moderna*, Einaudi, Torino.

Istat, 2009, Atlante di geografia statistica e amministrativa. Istituto Nazionale di Statistica, Roma.

Johnson M.P., 2001, *Environmental impacts of urban sprawl: a survey of the literature and proposed research agenda*, Environment and Planning A, 33, 717-735.

King R., Proudfoot L., Smith B., 1997, *The Mediterranean. Environment and society*, Arnold, London.

Ludlow D., Fons J., Weichselbaum J., Kleeschulte S., Steinnocher K., Guerois M., 2007. *Environmental dimensions of territorial development in Europe, Espon Work Proposal*, European Environment Agency, Copenhagen.

Maricchiolo C., Sambucini V., Pugliese A., Munafò M., Cecchi G., Rusco E., 2005, *La realizzazione in Italia del progetto europeo Corine Land Cover 2000*, Apat, Rapporti n. 61/2005, Roma.

Munafò M. 2008, *Valutazione della sostenibilità ambientale ed integrazione di dati ambientali e territoriali*, Ispra, Rapporti n. 82/2008, Roma.

Perdigao V., Christensen S., 2000, *The Lacoast atlas: Land cover changes in European coastal zones*, Joint Research Centre, Ispra.

Pileri P., 2007, *Compensazione ecologica preventiva. Principi, strumenti e casi*, Carocci Editore, Roma.

Pileri P., Lanzani A., 2007, *Appunti per una proposta di legge. Limitare il consumo di suolo, riqualificare i suoli non edificati, dare primato alla formazione di natura e paesaggio, compensazione ecologica preventiva, promuovere un'urbanizzazione sostenibile e responsabile*, a cura di Legambiente e DIAP, Politecnico di Milano.

Schneider A., Woodcock C.E., 2008, *Compact, dispersed, fragmented, extensive? A comparison of urban growth in twenty-five global cities using remotely sensed data, pattern metrics and census information*, Urban Studies, 45(3), 659-692.

UNECE-Eurostat, 1998. Recommendations for the 2000 Censuses of Population and Housing in the ECE region. New York and Geneva: United Nations.

Berdini P., 2010, *Breve storia dell'abuso edilizio in Italia, dal ventennio fascista al prossimo futuro*, Donzelli editore, Roma.

Direct vs Indirect Forecasts of Foreign Trade Unit Value Indices*

Giancarlo Lutero and Marco Marini[1]

Abstract

This paper examines the forecasting approach of foreign trade unit value indices followed in the compilation of quarterly national accounts of Italy. Total imports and exports indices are indirectly obtained from the aggregation of ARIMA forecasts of disaggregated components, derived from the program TRAMO with automatic identification options. An out-of-sample forecasting exercise is performed to validate the automatic choices made by TRAMO and to evaluate the relative performance of a direct forecasting approach of imports and exports aggregates. Also, we show how the use of international raw commodity prices can improve the forecasting accuracy of aggregate unit value indices.

Keywords: Forecast aggregation, Foreign trade statistics, Flash estimates, Quarterly National Accounts
JEL Classification: C32, C43, C53, F17

1. Introduction

The compilation of quarterly national accounts (QNA) in Italy relies on a system of short-term indicators of economic activity (monthly industrial production indices, monthly foreign trade statistics, quarterly households budget survey, etc.). With the current timeliness some indicators are not available for the most recent quarter, generally the most interesting one for users. This is the case of foreign trade unit value indices (UVIs), which are used for the deflation of imports and exports of goods in QNA. One or two months of the current quarter are not available at the time of publication: the recourse to forecasting methods is thus necessary to fill in the missing information and proceed with the subsequent steps of the estimation process.
Foreign trade UVIs are used in QNA at a detailed level of the NACE classification, more than 60 products for both imports and exports. This is justified by the fact that UVIs cannot be considered as a proxy of import and export prices at an aggregated level. The forecasting exercise is repeated two times each quarter, before the publication of the GDP flash estimate (45 days after the end of the quarter) and the complete set of QNA (70 days). The program TRAMO (Gomez and Maravall, 1997) is used to forecast on the basis of estimated Reg-ARIMA models. Automatic modeling options are used, including the choice of the ARIMA order, log or level specifications and outliers. The aggregated indices for imports and exports result indirectly from the linear combination of the individual forecasts by

---

* The opinions expressed in this paper are those of the authors and do not necessarily reflect the official position of ISTAT.

[1] ISTAT, National Accounts Directorate, Methods Development of Quarterly National Accounts.

product, with weights given by the values at current prices of annual national accounts imports and exports of goods.
Both theoretical considerations and empirical results available in the literature do not seem to suggest a clear preference for direct or indirect forecasting approaches. Results depend on the type of model used, the forecasting horizon, the kind of time series, and other factors. For example Benabal *et al.* (2004) investigates whether the indirect forecast of the Euro area Harmonized Index of Consumer Prices (HICP) from its components improves upon the forecast of overall HICP. The direct approach provides better results than the indirect one (especially in the long-term); however, if the HICP excluding the unprocessed food and energy is considered then the indirect approach prevails.
Through an out-of-sample forecasting exercise, this work aims at evaluating the accuracy of the indirect forecasting approach of UVIs against other alternatives, including the direct modeling of aggregate import and export indices.
The paper is organised as follows. Section 2 gives a brief review on the theory of aggregation and disaggregation in the context of forecasting. The general principles of forecasting adopted in QNA are presented in section 3. The forecasting exercise is described in section 4, with presentation of data used, design of the experiment, and main findings. Section 5 concludes with a summary and future development of the work.

2. Forecasting and the aggregation problem: still an open issue

The problem of aggregation is a controversial and debated topic in the economic literature. An attempt to find a suitable microeconomic foundation of macroeconomics is done by Forni and Lippi (1997); other seminal works are those of Theil (1954), Grunfeld and Griliches (1960) and Zellner (1962). Behind the theoretical implications, the increasing availability of economic statistics at different detail levels makes the aggregation problem very interesting in practical applications too. A typical example is the forecast of key variables for the euro area, which influences the decisions of monetary policy makers (ECB) and operators. The choice between forecasting the euro area aggregate or aggregating forecasts of the Member states is in fact non-trivial and must be carefully analyzed (Marcellino, 2004).
A key question in this work is whether the point forecasts of an aggregate (direct method) improves upon those derived from an indirect approach. Aggregation can be performed along with different dimensions; they can be classified into:

- *contemporaneous aggregation*, where the aggregation is made across variables according to a given classification (i.e. sub-indices of inflation rate,[2] the *Composite Leading Indicators* released by OECD);
- *spatial aggregation*, that regards aggregation across space (i.e. GDP for the euro area, see Bacchini *et al.*, 2010 for an example);
- *temporal aggregation*, that implies the transformation of observations from higher to lower frequencies (i.e. quarterly to monthly, monthly to quarterly, etc.);

[2] Inflation rate is often considered in practical applications; examples are Benabal *et al* (2004), Demers and de Champlain (2005), Hubrich (2005) dealing with the forecast of the HICP index for the euro area.

Another important aspect is the role of the aggregation rule. When several forecasts are obtained for the same variable, their combination is usually done with weights estimated according to some optimization criteria. This certainly increases uncertainty of forecasts (Timmermann, 2006). Instead, the indirect forecast of aggregates from their components does not suffer this problem, because it can be derived on the basis of pre-determined weights given by, for example, the current values of the fixed base period or the relative weights of countries.
Hendry (2004) suggests several issues that influence the model predictability:

- model specification (choice of variables, functional form, model selection);
- estimation uncertainty;
- data measurement errors;
- structural breaks over the forecast horizon.

Similarly to Hendry and Hubrich (2007), we introduce the following taxonomy of the different forecasting approaches according to the kind of information set:

- $\hat{y}^a_{t+h} = f(y^a_{t+h} \mid \Omega_t)$, where the h-step ahead forecast of the aggregated variable is a function of its past values $\Omega_t = \{y^a_t, y^a_{t-1}, \ldots\}$, with $h = 1, 2, \ldots$;
- $\hat{y}^a_{t+h} = f(y^a_{t+h} \mid \Lambda^1_t, \Lambda^2_t, \ldots \Lambda^n_t)$, where Λ^i_t, for $i = 1, \ldots, n$ are the information sets of past values of components at a detailed level of disaggregation, with $\bigcup \Lambda^i_t \neq \Omega_t$;
- $\hat{y}^a_{t+h} = f(y^a_{t+h} \mid \Omega_t, X_{t+h})$, where X_{t+h} contains additional external variables up to period $t+h$.

Assuming a linear functional form, the minimum mean squared forecast of the (direct) aggregated variable y^a_{t+h} is the conditional expectation

$$\hat{y}^{a,d}_{t+h} = E(y^a_{t+h} \mid \Omega_t) \tag{1}$$

Following an indirect approach, the forecast is determined as the linear combination of forecasts of n sub-components

$$\hat{y}^{a,i}_{t+h} = \sum_{i=1}^{n} \kappa_i E(y^i_{t+h} \mid \Lambda^i_t) \tag{2}$$

where the weights κ_i are known and satisfy the following constraints

$$\kappa_i > 0 \quad \sum_{i=1}^{n} \kappa_i = 1 \quad i = 1, \ldots, n$$

The debate has been enriched in the recent years by the increasing interest for nonlinear models, in particular the switching regime models,[3] and the potential of nonlinear forecasting[4]. The aggregation operator induces the macro-variables parameters to be intrinsically *time-varying* and therefore this suggests to use the *State-Dependent* model

$$y_t + \sum_{i=1}^{p} \phi_i(I_{t-1}) y_{t-i} = \mu(I_{t-1}) + \varepsilon_t + \sum_{j=1}^{q} \theta_j(I_{t-1}) \varepsilon_{t-j} \qquad (3)$$

which consists of a set of autoregressive parameters $\phi_i(I_{t-1})$, a set of moving average parameters $\theta_j(I_{t-1})$, and a local intercept $\mu(I_{t-1})$, depending on past information I_{t-1}. They are a generalization of linear ARIMA models, which results assuming constant coefficients.[5] Combining together the various functional forms of parameters $\mu(.)$, $\phi(.)$ and $\theta(.)$, it is possible to obtain a wide range of nonlinear models.[6] Macroeconomic aggregates might be interpreted as the parametric aggregation of two or more stochastic, or deterministic, regimes that represent "cluster" of micro-units, homogeneous in relation to their behaviors. These models are also called *piecewise* linear models because they represent linear micro-relationships that assume nonlinear framework because of the aggregation in space and time.
Despite the unequivocal limits of nonlinear models,[7] one of the most promising frontier of aggregation theory in forecasting seems to be the pooling of linear and nonlinear forecasts. Stock and Watson (2001) and Marcellino (2004) use a large data set of macroeconomic variables for the US and euro area respectively, comparing three forecasting methodologies: linear, pooled linear-nonlinear and nonlinear forecasts. The results are encouraging, as pointed out by Marcellino: "*In other words, pooled forecasts, or simple AR models, have a stable performance over all the variables, but specific linear or non-linear models can do better for specific series.*". Similarly, Timmermann (2006) states that the combination of forecasts from linear and non-linear models with different regressors might prevail in certain circumstances. However, non-linear models are more difficult to implement and to maintain in a data production context; therefore we restrict our attention in this work to linear time series models.
A general opinion on aggregation problems is that the selection between direct and indirect forecasts should be done more on the basis of empirical exercises than theoretical considerations. As noted by Stock and Watson (2001): "*...time series models and forecasting methods, however appealing from a theoretical point of view, ultimately must be judged by their performance in real economic forecasting applications.*". The purpose of this work is just to compare the two alternatives on a practical case encountered in the Italian QNA.

---

[3] See Granger and Teräsvirta (1993) and Tong (1990) for an introductional survey on nonlinear modelization.
[4] See Stock and Watson (2001), Marcellino (2004) and most recently Granger (2008).
[5] Any nonlinear model can be approximated by linear time-varying parameters model, as demonstrated by the White theorem; see Granger (2008).
[6] For example bilinear models, threshold models, Markov-chain models, autoregressive with smooth transition, autoregressive with neural networks, etc.
[7] As is stressed in Granger (2008): "*...most nonlinear models are difficult to use to form point forecasts more than one step ahead and forecast confidence intervals are also typically difficult to obtain.*"

3. The practice of forecasting in QNA

In Italy, QNA are compiled through an indirect approach: quarterly time series of NA aggregates are derived from temporal disaggregation of annual data by means of short-term indicator series. Indicators are chosen according to well-founded statistical and economic relationships with aggregates (see Marini and Fimiani, 2006). For example, quarterly production (and value added) of manufacturing activities are based upon econometric relationships between annual NA data and industrial production indices; quarterly imports of goods are derived on the basis of monthly imports from external trade statistics; etc.
When yearly data are known, temporal disaggregation ensures their values are distributed across the quarters according to the movements of the chosen indicator series: long-term trends of NA variables and intra-year variations of short-term indicators are thus mixed together in QNA time series. When the annual figure is not yet available (normally the most recent year), short-term information are also employed to extrapolate the quarterly behavior of QNA aggregates during the year. This probably constitutes the most delicate and crucial task in the compilation of QNA, considering the prominent role of GDP and its components for purposes of economic analysis, decision-taking and policy-making.
Timeliness of indicators is of key importance in QNA. The preliminary estimate of GDP (the so-called flash estimate) is released by ISTAT after 45 days the end of the reference quarter; the complete set of production, expenditure and income accounts are published at 70 days. The acquisition of monthly and quarterly indicators carries on until the very last moment in both cases, in order to exploit as much as possible the information set available for the current quarter. Nevertheless, the latest observations of some indicators might still be missing due to collection and processing issues. For monthly indicators, this implies that only one or two months of the quarter are known: the remaining information must be predicted somehow to complete the quarterly information set.
The program TRAMO (Gomez and Maravall, 1997) is used to this purpose. This tool is a natural choice for ISTAT researchers, being TRAMO employed, along with the companion program SEATS, for seasonal and calendar adjustment of QNA indicators. TRAMO computes forecasts according to Reg-ARIMA models, which is a convenient way to model a time series with both deterministic and stochastic effects. A pure automatic modeling strategy is normally followed when the target is the prediction of missing information (instead, manual intervention of the user is preferred in seasonal adjustment processes): the order of ARIMA models, the type and number of outliers, level or log-level specifications are all chosen by the automatic routines available in TRAMO. Despite the reduced control this automatism implies, this practice allows to obtain reliable and prompt time series forecasts of the missing months in a very short time. Clearly, the recourse to forecasting is more frequent in flash estimates of GDP: this is the reason why preliminary estimates are affected by more uncertainty than the data published after 70 days.
Unit value indices (UVIs) of foreign trade statistics represent a typical information in QNA that needs to be forecasted. These indices are used in Italy to deflate current values estimates of imports and exports of goods, considered as a proxy of import and export prices. Moreover, they contribute to the construction of the system of input and output prices (along with domestic prices), being used for the deflation of output and intermediate consumption. On average, UVIs are published by ISTAT 50 days after the end of the month. This implies that only one month of UVIs is available for GDP flash estimates and two months for the complete estimation of quarterly accounts. One and two-step ahead forecasts are thus calculated to complete the information of the current quarter.

A description of UVIs (and their use in QNA) is provided in section 4.1. Here it is worth remarking the importance of such information in QNA. As stated above, the estimate of imports and exports in volume are obtained by applying UVIs to the current values' estimates. Poor forecasts of UVIs lead to bad volume estimates of external components of GDP, and thus of GDP itself. Moreover, forecasting errors of UVIs have a negative impact on the GDP deflator through the system of input and output prices. From our past experience it is possible to state that monthly UVIs are very difficult to predict: they are volatile, affected by structural breaks and outliers and sometimes present a highly unstable seasonal component.
These properties are particularly evident when indices are considered at the 3-digit NACE classification, that is currently used in the estimation of NA. At this detail, there are 60 products traded between Italy and foreign countries. Disaggregated UVIs are therefore taken into account in the deflation process: total imports and exports (of goods) in volume are indirectly derived by aggregating the volume estimates of such products.
Generally, disaggregated time series are less predictable than aggregated data. This seems confirmed in UVIs: the total UVI of imports and exports show certainly smoother movements than their components by sector. Therefore, the practice of forecasting disaggregated information when the primary target is the aggregate variable (in this case exports and imports in volume) might be questionable. In such cases, a direct forecasting model to predict total UVIs of imports and exports might outperform the indirect approach.
The use of time series models guarantees point forecasts in accordance with past movements of the individual series; no information is considered on the periods to be predicted. If available, gain accuracy can be achieved by considering exogenous information through appropriate specifications of regression models (possibly with a dynamic structure). Despite some attempts in the past,[8] forecasting models with exogenous information have never been used in the production process. Usually, the main difficulty is just connected with the lack of ready-to-use information on the missing months. However, the situation for UVIs of imports and exports is now different. For example, imports prices are likely to depend on world index prices of primary commodities, such as crude oil or steel, which are very rapidly available on international data warehouses (such as those of IMF or Eurostat); exports prices can be somehow related to domestic prices of manufactured goods (released by ISTAT after 30 days) or, even better, to producer price indices on foreign markets, recently made available by ISTAT.
Through a real-time forecasting exercise this work aims at assessing the current practice adopted in QNA to forecast foreign trade UVIs along different directions, summarized by the following questions:

- do the automatic routines in TRAMO guarantee a satisfactory out-of-sample performance?
- would a direct approach to forecasting total UVIs of imports and exports improve upon the results of an indirect approach?
- when available, can the use of additional information be effective to increase the forecasting accuracy of UVIs?

The results of the experiment presented in the next section provide useful information to answer each of these questions.

[8] Forecasting models with qualitative variables extracted from business and consumer surveys (available within a month) have been fitted to some indicators of production and expenditure components, generally with unsatisfactory results.

4. The real-time forecasting exercise

1. The data

Imports and exports UVIs are published every month by ISTAT. The calculation of UVIs have been recently revised (ISTAT, 2008), in order to comply with new international standards and introduce important methodological improvements. UVIs are now derived from a very detailed level of product disaggregation, which generates more than 220,000 elementary indices. The aggregation process of the elementary indices is done through the use of trimmed means, that smoothes the high volatility of the original flows.
UVIs in Italy are Fisher-type indices, namely they are obtained as the geometric mean of Laspeyres and Paasche indices. The base of the index shifts every year, with weights given by imports and exports of previous year at current prices. Chain-linked time series are derived using the annual overlap technique. Total imports and exports UVIs are shown in figures 1 and 2. Both series exhibit an upward long-term trend, with cyclical fluctuation (not exactly synchronized) and many spikes throughout the period. Taking the logarithms of the data and applying the first difference operator, non-stationarity is removed from both series (according to the Augmented Dickey Fuller test, not reported in this paper but available on request). A seasonal component is not clearly identifiable. From an exploratory analysis with TRAMO, it is found that the most suited model for imports is the classical Airline model (0,1,1)(0,1,1); instead, the exports series is well represented by the non-seasonal ARIMA model with order (0,1,1).
As said before, UVIs are considered at the three-digit level of the NACE classification. This is presented in table 1, reporting codes and descriptions of each sector of economic activity. This is part of the broader classification used in national accounts by ISTAT, made up of 101 branches. Clearly, disaggregated UVI time series at this detail level present common features and idiosyncratic movements: the relative shares vary according to the type of product.
Table 2 presents the current values (and their percentages over the total) of imports and exports by product in year 2005. Imports of crude petroleum and natural gas (product 6) has the largest share (about 13%), followed by products 51 (motor vehicles, 11.4%), 37 (iron, steel and ferrous materials, 8.9%), and 27 (chemicals, 6.4%). Concerning exports, the largest contribute is by far that one of production of machine and mechanical tools (16.9%); exports' shares of products 51 (7.9%) and 37 (5.8%) are also notable.
The sample used in the exercise covers monthly data from 1996:1 to 2007:12. The data are not seasonally adjusted, but seasonality is present in UVIs for some products: seasonal ARIMA models are occasionally identified by TRAMO. There are 62 imported products in the chosen classification, and 61 for exports (crude oil not exported by Italy).
The aggregate UVIs cannot be indirectly derived from the disaggregate UVIs. This happens because the indices are chain-linked, and so they suffer the additivity problem. This represents a problem in our exercise, because aggregate forecasts cannot be immediately derived from disaggregate forecasts. To overcome such problem, aggregation of forecasts is done by means of the transformed indices having the previous year as the base period (the inverse process of the annual overlap chain-linking). Next, these indices are applied to deflate monthly levels of imports and exports at current prices: the resulting estimates are volumes expressed at prices of the previous year, that can be added to achieve the aggregate imports and exports in volume (but with a shifting base year). Finally, the aggregate chain-

linked UVIs are obtained by applying the annual overlap technique to the aggregate at current prices and at previous year's prices.

2. The experimental design

An out-of-sample exercise is used to evaluate the accuracy of ARIMA forecasts resulting from the following two strategies:

- identify each time the order of the ARIMA model according to the automatic model identification implemented in TRAMO (strategy AMI), that corresponds to the current practice adopted in QNA;
- use the standard ARIMA model (0,1,1)(0,1,1) in all the experiments (strategy AIR), often chosen because it fits generally well many economic time series.

For each of these strategies two experiments are conducted to mimic the actual situations encountered in the estimation of QNA. In the former experiment the last two months of the quarter are considered as missing and need to be forecasted. To complete the quarterly information, it is thus necessary to calculate one-step and two-step ahead ARIMA projections. The quarters from 2002 to 2007 is used to evaluate the forecasting performance. The exercise starts with the forecasts of 2002:2 and 2002:3, on the basis of the sample 1996:1-2002:1. After that, the complete information for quarter 2002:Q1 can be calculated by averaging the actual value for the first month and the forecasts for the remaining two months. Next, the forecasts of 2002:5 and 2002:6 are calculated, shifting the in-sample period one quarter ahead (1996:1-2002:4). The sample is then extended sequentially by three months until 2007:10, from which the forecasts of 2007:11 and 2007:12 are derived. The parameters of the models are re-estimated each time; moreover, in the strategy AMI the order of the ARIMA model is chosen each time.
In the second experiment two months of the quarter are considered as known, with the last month to be predicted: then, the exercise begins with the prediction of 2002:3 on the basis of the in-sample period 1996:1-2002:2, then the prediction of 2002:6 with 1996:1-2002:5, and so on. In this case, only a one-step ahead forecast is necessary: this is in fact the problem actually faced at 70 days for the complete estimation of QNA. Overall, we compute 48 forecasts (one-step and two-step ahead) in the first exercise, 24 in the second one (only one-step ahead).
Forecasts are evaluated with standard measures of accuracy: the Root Mean Squared Forecast Error (RMSFE) and the Mean Forecast Error (MFE). They are both calculated on the year-on-year growth rates. It is useful to introduce a formal notation to define both measures properly. Denoting with h the forecast horizon, the forecast error is defined as follows

$$e_{t+h|t} = y_{t+h} - \hat{y}_{t+h|t} \quad h = 1,2$$

where y_{t+h} is the annual growth rate (in %) calculated from monthly data m_t as

$$y_{t+h} = \frac{x_{t+h} - x_{t+h-12}}{x_{t+h-12}} * 100$$

and $\hat{y}_{t+h|t}$ is the same rate obtained with the actual value x_{t+h} replaced by its forecast $\hat{x}_{t+h|t}$.
The MFE is calculated as

$$\text{MFE} = \frac{1}{TH}\sum_{t=1}^{T}\sum_{h=1}^{H} e_{t+h|t}$$

with the index *t* denoting all the months in the forecasting period and *H* equals to 1 or 2. This measure is useful to verify the presence of a forecast bias. The RMSFE is derived according to the following formula

$$\text{RMSFE} = \left[\frac{1}{TH}\sum_{t=1}^{T}\sum_{h=1}^{H} e_{t+h|t}^{2}\right]^{1/2}$$

that measures the average size of error, irrespective of their signs.
The same exercise is done for aggregated and disaggregated UVIs (imports and exports). Aggregate forecasts are also derived indirectly from the disaggregated forecasts, using the procedure described in the previous section.
A final remark concerns the software used in this work. We have already cited TRAMO: the Linux version of December 2005 has been used, available in the software Modeleasy+. The program *R* (version 2.8.0),[9] the well-known open-source environment that offers both a high-level programming language and a wide collection of statistical and mathematical libraries, has been used for data processing. Finally, the software *Gretl* has also been employed to estimate the dynamic model used in section 4.4: it is a very good and user friendly open-source econometric software, developed by Allin Cottrell and Jack Lucchetti.[10]

3. Results

Table 3 compares the out-of-sample results in terms of RMSFE and MFE of the approaches AMI and AIR. The two exercises (2 months missing and only one month missing) are considered apart. The table shows the number of times the approaches AMI and AIR obtains the minimum statistics. Considering the first exercise, the minimum RMSFE is achieved in 40 out of 61/62 cases for imports/exports. Instead, the MFE statistic does not show any significant difference. As far as the second exercise is concerned, AMI shows again a better performance relative to AIR for 35 products.
For completeness, table 4 and 5 present all RMSFE and MFE statistics for imports and exports UVIs by product. They are presented for both approaches AMI and AIR. The first four columns refers to the exercise with two months predicted for each quarter, the last four columns to the exercise with only one month missing. Large RMSFE statistics are found

[9] URL http://www.R-project.org.
[10] Both software are released under GNU General Public License.

for several products, but the most important ones are those relative to products with a high weight. Concerning imports, the RMSFE is very large for products 6 (around 7% in the first exercise), 26 (10.9%) and 60 (9.6%): prices of these products are strictly connected with the world energy market and therefore are subject to a higher price volatility. This certainly makes imports UVIs less predictable than exports; in fact, the RMSFE are often higher than that of the corresponding exports UVIs.
Overall, the AMI approach yields satisfactory results: this confirms the good properties of TRAMO as an automatic forecasting tool. To evaluate the stability of the selection process, we verify the sequence of ARIMA models chosen for each product in the simulation exercise. Table 6 and 7 shows the number of times in which selected ARIMA models are identified in the series. For imports, the order (0,1,1) is identified in about 40% of the cases: therefore, most of the series do not present a seasonal component. Considering the nature of the data, this result is quite reasonable. However, the classical Airline model is found in 22%: for products 17, 41, and 43 it is even the most frequent model. Regarding exports, the model (0,1,1) is again the most selected one but with a smaller percentage than imports (less than 27%). The Airline model is confirmed in the second position (24.4%).
The same forecasting experiment is replicated for the aggregate imports and exports UVIs (those shown in figures 1 and 2). The first row in tables 6-7 presents the ARIMA orders chosen by the AMI approach. For imports the most selected model is (0,1,1) (37 out of 48 cases); two seasonal models are instead identified for exports (the Airline and the model (0,1,0)(0,1,1)). At an aggregate level, seasonality is thus more visible in exports than imports UVI series. Table 8 compares the RMSFE and MFE statistics of the direct forecasts with those derived indirectly from the disaggregated forecasts. For imports, the indirect approach clearly prevails against the direct approach: 1.079% against 1.237% in the first exercise, and even 0.863% against 1.475% in the second exercise (the AMI and AIR approaches gives approximately the same results). The MFE is also lower following an indirect approach: in the second exercise, it drops from -0.22% to -0.02%. On the contrary, the two approaches provides very similar results for exports: it is worth noting that the direct approach provides the minimum RMSFE in the first exercise (0.709% against 0.727%).

4. Forecasting with exogenous information: a dynamic model for imports UVIs of crude oil and gas

As a final experiment, a dynamic regression model is used to forecast the imports UVI of crude oil and natural gas (product 6). A couple of useful world price indices are available from the IMF website (Primary Commodity Prices section): a crude oil (petroleum) price index and a natural gas price index. The former is calculated as a simple average of three spot prices: UK Brent, West Texas Intermediate, and the Dubai Fateh. It is published within a month from the reference period, so it might be used to forecast the missing information of the UVI. Since the prices are expressed in $ per barrels, the index must be first transformed in € before putting it into relationship with UVI. The euro-dollar exchange rate series is used to this end. The latter is computed as an arithmetic mean of Russian Natural Gas, Indonesian Liquified Natural Gas and Natural Gas spot prices at the Henry Hub terminal in Louisiana, expressed in US$ per cubic meters of liquid.
The crude oil price index (COPI), the natural gas index and the imports UVI are compared in figure 3 (in logs). The three series show very similar movements: the imports UVI of

petroleum products is strictly connected with both indices. Since the latters can be considered exogenous information of the former, it is useful to analyze its contemporaneous and delayed effects on the UVI. In fact, a change in the price index might not affect immediately the imports in Italy, but with a certain delay. Figure 4 shows the cross-correlogram of the stationary transformation (first-differences of log-levels) of UVI with leads and lags of COPI. Positive values in the x-axis indicate lags of COPI, whereas negative values indicate leads. The cross-correlogram is computed up to lag/lead 13. It is shown that COPI is a fairly coincident index relative to UVI, with large and positive correlation at lags 0 and 1 (0.59 and 0.68, respectively). Apart from lags 8 and 13, the correlation coefficients at other lags are also positive (even if not significant). Considering the dynamic relationship, an Autoregressive model with Distributed Lags (ADL) model is used to derive forecasts on the basis of COPI and gas.

To simplify notation, we denote by y_t the imports UVI of product 6 and by x_t the crude oil price index and by z_t the gas index. We start by fitting the general ADL model of order 13

$$\Delta y_t = \alpha_0 + \sum_{i=1}^{13} \alpha_i \Delta y_{t-i} + \sum_{j=0}^{13} \beta_j \Delta x_{t-j} + \sum_{j=0}^{13} \gamma_j \Delta z_{t-j} + \varepsilon_t$$

with the usual IID normal assumption for ε_t (the sample 1996:1-2002:1 is used for the specification). Then, the model is simplified by omitting the non-statistically significant lags, following a general-to-specific approach (Hendry, 2004). The sequential strategy implemented in the software *Gretl* is followed: the dependent variable with the highest *p*-value is omitted at each step, until all the remaining variables show *p*-values less than 0.10 per cent. The selection process yields the specific model presented in table 9: the first autoregressive term Δy_{t-1}, the contemporaneous term and 2 lagged terms of COPI ($\Delta x_t, \Delta x_{t-1}, \Delta x_{t-7}$) and one lagged term of gas (Δz_{t-2},) enter the final equation. The goodness of fit of the model is satisfactory ($\bar{R}^2 = 0.76$) and standard diagnostics on residuals are acceptable.

The equation model in table 9 is used throughout the out-of-sample period (2002:1-2007:12). The model parameters are estimated each time with additional observations: the values of the coefficients and the statistical properties of the model do not vary across the period, therefore the model can be considered sufficiently robust. The same forecasting exercises described in the previous section are performed, with prediction of two months of the quarter (one- and two-step ahead forecasts) and one month (one-step ahead forecast). Table 10 compares RMSFE and MFE statistics obtained from the specified ADL model against the ARIMA model (with the AMI approach). The RMSFE value is reduced from almost 7% to 4% in the first exercise, from 5.5% to 3.2% in the second exercise. Overall, the reduction of RMSFE for total imports UVI is strong, around 0.2% when two months are predicted (from 1.08% to 0.88%).

5. Conclusion

The aim of this paper is to assess the current practice of forecasting external trade UVIs in Italian QNA. The program TRAMO with automatic options is used to obtain one-step and two-step ahead forecasts of imports and exports UVIs disaggregated according to the NACE classification. Forecasts of total imports and exports UVIs are obtained from the aggregation of the disaggregated forecasts.
Through an out-of-sample exercise, this practice is assessed along three different directions. Firstly, the automatic selection strategy of TRAMO is evaluated in comparison with a standard ARIMA model (the Airline model). Then, a direct forecasting approach is experimented. Finally, the use of exogenous information to improve the forecasting accuracy is investigated.
The main findings shown in the paper suggest that:

- the automatic selection process of the ARIMA model carried out by TRAMO provides acceptable forecasts, on average better than those from the classical Airline model. In this way, we have certified the opportunity to adopt TRAMO as a pure forecasting tool;
- the indirect forecasting approach outperforms the direct approach in the case of total imports UVI; for exports, the two approaches give approximately the same results. This is probably connected with the higher volatility of imports UVIs of certain products (i.e. crude oil and gas), that worsen the predictability of the aggregate series. Therefore, a direct approach does not ensure any gain in accuracy with these data;
- the RMSFE is markedly reduced when a simple ADL model for imports UVI of crude oil and gas products is used, based on world market crude oil and natural gas price indices.

The last finding seems very interesting and promising for the future. For example, imports UVIs (but also exports) disaggregated by product can be put into relationships with other primary commodity prices (steel, iron, agricultural products, etc.). This practice would be simple to implement and maintain, fruitful and even feasible considering time and resource constraints of a data producer. We believe that this practice is likely to improve forecasting accuracy of UVIs and, more generally, the accuracy of QNA.

References

BACCHINI, F., CIAMMOLA, A., IANNACCONE, R. AND MARINI, M. (2008). "Combining Forecasts for Producting Flash Estimates of Euro area GDP", presented at Eurostat colloquium on "Modern Tools on Business Cycle Analysis", Luxembourg, September.

BENALAL, N., DEL HOYO, J. L. D., LANDAU, B., ROMA, M. AND SKUDELNY, F. (2004). "To aggregate or not to aggregate? euro area inflation forecasting". *Working Paper Series* 374, European Central Bank.

CLEMEN, R. (1989). "Combining forecasts: A review and annotated bibliography". *International Journal of Forecasting*, 5, pp. 559–583.

CLEMENTS, M. AND HENDRY, D. (1998). "Forecasting Economic Time Series",. Cambridge University Press, Cambridge, UK.

DEMERS, F. AND DE CHAMPLAIN, A. (2005) "Forecasting core inflation in canada: should we forecast the aggregate or the components?" Working Paper 44, Bank of Canada.

FORNI, M. AND LIPPI, M. (1997). "Aggregation and the Microfoundations of Dynamic Macroeconomics". Oxford University Press.

GOMEZ, V. AND MARAVALL, A. (1997). Programs TRAMO and SEATS: Instructions for the User. Bank of Spain.

GRANGER, C. (1990). "Aggregation of time series variables: a survey". In "Disaggregation in Econometric Modelling" (edited by BARKER, T. AND PESARAN, M. H.), pp. 17–34. Routledge London and New York.

GRANGER, C. W. J. (2008). "Non-linear models: where do we go next-time varying parameter models". Studies in Nonlinear Dynamics & Econometrics, 12, n. 3, Article 1.

GRANGER, C. W. J. AND BATES, J. (1969), "The combinations of forecasts", Operations Research Quarterly, 20, pp. 451–468.

GRANGER, C. W. J. AND NEWBOLD, P. (1986), "Forecasting Economic Time Series", Academic Press Inc, San Diego.

GRANGER, C. W. J. AND TERSVIRTA, T. (1993), "Modelling Non-Linear Economic Relationship", Oxford University Press.

GRUNFELD, Y. AND GRILICHES, Z. (1960), "Is aggregation necessarily bad?", Review of Economics and Statistics, 42, pp. 1–13.

HENDRY, D. (2004), "Unpredictability and the foundations of economic forecasting", Working paper, Economics Department, Oxford University. 15

HENDRY, D. F. AND CLEMENTS, M. P. (2002), "Pooling of forecast", Econometrics Journal, 5, pp. 1–26.

HENDRY, D. F. AND HUBRICH, K. (2007), "Combining disaggregate forecasts or combining disaggregate information to forecast an aggregate", presented at Conference in honour of David F. Hendry, Oxford University 23-25 august 2007.

HUBRICH, K. (2005), "Forecasting euro area inflation: does aggregating forecasts by hicp component improve forecasts accuracy?", International Journal of Forecasting, 21(1), pp. 119–136.

HYNDMAN, R. J. (1995), "Highest-density forecast regions for non-linear and non-normal time series models", Journal of Forecasting, 14, pp. 431–441.

ISTAT (2008), "Quarterly National Accounts Inventory - Sources and methods of Italian Quarterly National Accounts", EUROSTAT, Luxembourg.

LEE, K., PESARAN, M. AND PIERCE, R. (1990), "Testing for aggregation bias in linear models", The Economic Journal (Supplement), 100, pp. 137–150.

MARCELLINO, M. (2004), "Forecast pooling for european macroeconomic variables", Oxford Bulletin of Economics and Statistics, 66(1), pp. 91–112.

MARINI, M. AND FIMIANI, C. (2006), "Le innovazioni introdotte nelle tecniche di stima della contabilità nazionale", Presented at ISTAT Conference "La revisione generale dei conti nazionali 2005", Rome 21-22 june 2005.

ORCUTT, G., H.W.WATTS AND EDWARDS, J. (1968), "Data aggregation and information loss", The American Economic Review, 58, pp. 773–787.

R DEVELOPMENT CORE TEAM (2008), "R: A Language and Environment for Statistical Computing", R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

STOCK, J. AND WATSON, M. (2001), "A comparison of linear and nonlinear univariate models for forecasting macroeconomic time series", in "Cointegration, Causality and Forecasting", a Festschrift in Honour of Clive W.J. Granger (edited by ENGLE, R. AND WHITE, H.). Oxford University Press.

THEIL, H. (1954), "Linear Aggregations of Economic Relations", Amsterdam, North Holland.

TIMMERMANN, A. (2006), "Forecast combinations", in "Handbook of economic forecasting" (edited by ELLIOT, G., GRANGER, C. AND TIMMERMANN, A.), vol. 1, pp. 135–196. Amsterdam, Elsevier. 16

TONG, H. (1990), "Nonlinear time series, a dynamical system approach", Oxford University Press.

ZELLNER, A. (1962), "An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias", Journal of the of the American Statistical Association, 57, pp. 348–368.

Table 1 - NACE-rev.1.1 classification used in National Accounts (only imported and exported products)

Codes	Description
1	Growing of crops; market gardening; horticulture; agricultural and animal husbandry service activities, except veterinary services
2	Farming of animals; hunting, trapping and game propagation; growing of crops combined with farming of animals; related service activities
3	Forestry, logging and related service activities
4	Fishing, operation of fish hatcheries and fish farms; service activities incidental to fishing
5	Mining and agglomeration of coal, lignite and peat
6	Extraction of crude petroleum and natural gas; mining of uranium and thorium ores; incidental service activities
7	Mining of iron ores; mining of non-ferrous metal ores, except uranium and thorium ores
8	Quarrying of stone, gravel, sand and clay and other quarried minerals; production of salt
9	Mining of chemical and fertilizer minerals
10	Production, processing and preserving of meat and meat products
11	Processing and preserving of fish and fish products; manufacture of vegetable and animal oils and fats; manufacture of other food products
12	Processing and preserving of fruit and vegetables
13	Manufacture of dairy products
14	Manufacture of grain mill products, starches and starch products
15	Manufacture of prepared animal feeds
16	Manufacture of tobacco products
17	Manufacture of beverages
18	Preparation and spinning of textile fibres; textile weaving; finishing of textiles
19	Manufacture of made-up textile articles; manufacture of knitted and crocheted fabrics; manufacture of knitted and crocheted articles
20	Manufacture of wearing apparel; dressing and dyeing of fur
21	Tanning and dressing of leather; manufacture of leather products
22	Manufacture of footwear
23	Sawmilling and planing of wood; impregnation of wood; manufacture of builders' carpentry and joinery; wooden containers; panels and boards; plywood; carpentry and joinery
24	Manufacture of pulp, paper and paper products
25	Publishing, printing and reproduction of recorded media; related activities
26	Manufacture of coke oven products; manufacture of refined petroleum products; processing of nuclear fuel
27	Manufacture of basic chemicals
28	Manufacture of chemical products for agriculture, building, printing and various other uses
29	Manufacture of pharmaceuticals, medicinal chemicals and botanical products; manufacture of soap and detergents, toilet preparations
30	Manufacture of man-made fibres
31	Manufacture of rubber products
32	Manufacture of plastic products
33	Manufacture of glass products
34	Manufacture of ceramic products
35	Manufacture of cement, lime and plaster; manufacture of articles of concrete, plaster and cement
36	Manufacture of other non-metallic mineral products; cutting, shaping and finishing of stone
37	Production of iron, steel and ferro-alloys (ECSC); manufacture of basic precious and non-ferrous metals first processing
38	Manufacture of structural metal products
39	Forging, pressing, stamping and roll forming of metal; treatment and coating of metals; manufacture of various metal tools
40	Manufacture, installation, repair and maintenance of machine tools and machinery for the production and use of mechanical power, manufacture of weapons and ammunition
41	Manufacture of agricultural and forestry machinery
42	Manufacture of domestic appliances
43	Manufacture of office machinery and computers
44	Manufacture of electric motors, generators and transformers
45	Manufacture of electricity distribution and control apparatus, accumulators, primary cells and primary batteries, and lamps and lighting fittings
46	Manufacture of electronic valves and tubes and other electronic components
47	Manufacture of television and radio transmitters and apparatus for line telephony and line telegraphy
48	Manufacture of television and radio receivers, sound or video recording apparatus
49	Manufacture of medical and surgical equipment and orthopaedic appliances; manufacture of instruments and appliances for measuring, checking, testing, navigating and the like

Table 1 continued **- NACE-rev.1.1 classification used in National Accounts (**only imported and exported products

Codes	Description
50	Manufacture of optical instruments and photographic equipment; manufacture of watches and clocks
51	Manufacture of motor vehicles, trailers and semi-trailers, including coachwork, parts and accessories
52	Manufacture of motorcycles, bicycles and other transport equipment
53	Building and repairing of ships and boats
54	Manufacture of locomotives and rolling stock
55	Manufacture of aircraft and spacecraft
56	Manufacture of furniture and musical instruments
57	Manufacture of jewellery and related articles
58	Manufacture of sports goods, games and videogames; miscellaneous manufacturing n.e.c.
60	Production and distribution of electricity, steam and hot water
88	Computer and related activities
90	Professional and business activities
99	Recreational and cultural activities

Table 2 - Imports and exports of goods by product in 2005. Current values in billions of €

	Imports		Exports			Imports		Exports	
Sector	billions €	%	billions €	%	Sector	billions €	%	billions €	%
1	6206	2.004	3944	1.317	34	814	0.263	4241	1.417
2	2268	0.732	100	0.034	35	416	0.134	482	0.161
3	562	0.182	109	0.036	36	702	0.227	2356	0.787
4	846	0.273	207	0.069	37	27626	8.922	17430	5.822
5	1791	0.578	6	0.002	38	839	0.271	2879	0.961
6	39473	12.749	0	0.000	39	4218	1.362	10545	3.522
7	1379	0.445	74	0.025	40	19570	6.321	50640	16.914
8	1119	0.361	440	0.147	41	607	0.196	2987	0.998
9	130	0.042	61	0.020	42	1952	0.630	7167	2.394
10	4977	1.607	1749	0.584	43	8222	2.655	2111	0.705
11	7674	2.478	6521	2.178	44	2265	0.731	3122	1.043
12	1255	0.405	1959	0.654	45	6124	1.978	8051	2.689
13	2972	0.960	1506	0.503	46	3368	1.088	3066	1.024
14	502	0.162	809	0.270	47	5876	1.898	2792	0.933
15	596	0.193	206	0.069	48	4517	1.459	1476	0.493
16	1781	0.575	20	0.007	49	6730	2.174	4684	1.565
17	1334	0.431	4228	1.412	50	2040	0.659	2750	0.918
18	3581	1.157	7850	2.622	51	35438	11.446	23841	7.963
19	3790	1.224	6539	2.184	52	1634	0.528	2130	0.711
20	8418	2.719	12421	4.149	53	1030	0.333	2972	0.993
21	2962	0.957	5649	1.887	54	361	0.116	481	0.161
22	3696	1.194	7370	2.462	55	2102	0.679	2396	0.800
23	3885	1.255	1444	0.482	56	1647	0.532	8949	2.989
24	5909	1.908	4887	1.632	57	1012	0.327	4145	1.384
25	957	0.309	1653	0.552	58	3037	0.981	2644	0.883
26	6211	2.006	10153	3.391	60	2187	0.706	64	0.021
27	19843	6.409	10181	3.401	88	916	0.296	93	0.031
28	6087	1.966	4965	1.658	90	6	0.002	18	0.006
29	14636	4.727	14503	4.844	99	96	0.031	273	0.091
30	1424	0.460	1104	0.369	100	3	0.001	4	0.001
31	2474	0.799	3087	1.031	Total	309616	100.000	298892	100.000
32	4107	1.327	8368	2.795					
33	1419	0.458	1991	0.665					

Table 3 - AMI versus AIR strategies: number of times with minimum RMSFE and MFE

	2 months missing		1 month missing	
Index	AMI	AIR	AMI	AIR
Imports				
RMSFE	40	22	35	27
MFE	29	33	28	34
Exports				
RMSFE	40	21	35	26
MFE	31	30	30	31

Table 4 - Out-of-sample performances of disaggregated Imports UVIs

	2 months missing				1 month missing			
Sector	*RMFSE*		*MFE*		*RMFSE*		*MFE*	
	AMI	AIR	AMI	AIR	AMI	AIR	AMI	AIR
1	2.946	2.644	0.459	0.273	1.477	1.368	0.091	-0.058
2	2.717	2.708	0.406	0.627	2.508	2.466	-0.027	0.073
3	2.170	2.356	0.399	-0.074	2.065	2.004	0.683	-0.010
4	1.843	1.834	0.291	0.432	1.954	1.967	0.644	0.780
5	10.160	10.081	1.629	0.471	6.773	6.107	2.187	1.244
6	6.962	7.820	1.103	-0.939	5.532	5.987	-0.309	-0.738
7	7.026	8.061	-0.137	-1.969	5.433	5.863	-1.303	-0.809
8	2.167	2.279	-0.619	-0.653	2.022	2.069	0.003	-0.261
9	4.670	5.097	1.352	-0.411	4.315	4.319	0.955	-0.203
10	3.094	2.807	1.170	0.411	1.957	1.775	1.025	0.271
11	1.819	1.869	0.278	-0.047	1.650	1.650	0.037	-0.130
12	1.964	2.197	0.822	0.427	1.713	1.950	0.130	-0.126
13	1.183	1.557	0.360	0.267	1.050	1.088	-0.038	0.144
14	1.728	1.790	0.139	0.064	1.272	1.239	0.075	-0.307
15	2.155	2.246	0.169	-0.207	2.407	2.359	0.918	0.222
16	4.733	3.894	0.300	-0.387	3.584	3.957	-0.414	-0.642
17	2.982	2.900	0.237	-0.314	2.981	2.987	-0.393	-0.733
18	1.159	1.300	-0.060	-0.248	0.896	0.795	0.016	0.129
19	1.656	1.716	0.109	-0.147	1.359	1.159	0.120	-0.071
20	2.020	2.005	-0.128	-0.087	1.784	1.834	0.201	0.164
21	4.986	4.680	0.509	0.346	2.836	3.094	0.569	0.518
22	2.930	2.608	0.545	0.029	2.866	2.516	0.108	-0.072
23	1.520	1.059	0.362	0.162	1.214	0.789	-0.072	0.024
24	1.390	1.670	-0.140	-0.285	1.295	1.226	-0.441	-0.291
25	5.821	6.015	1.211	0.872	5.563	5.672	1.647	1.455
26	10.930	12.397	-0.989	-1.128	6.212	6.484	-0.427	-0.503
27	2.638	2.435	0.117	-0.068	1.715	1.304	0.110	-0.272
28	2.176	2.161	-0.083	-0.436	2.333	2.289	0.157	-0.007
29	4.558	4.333	0.744	-0.790	3.778	3.444	-0.202	-1.410
30	1.476	1.549	0.145	-0.055	1.785	1.741	-0.490	-0.519
31	1.650	1.699	0.490	-0.083	1.763	1.644	0.470	0.111
32	1.044	1.304	0.244	-0.085	0.956	1.217	-0.083	-0.473
33	1.473	1.582	0.136	-0.485	1.704	1.809	-0.145	-0.290
34	2.440	2.340	0.859	0.669	2.188	2.258	0.382	0.216
35	3.087	3.318	0.323	-0.225	2.196	2.740	-0.353	-0.365
36	2.073	2.228	0.096	0.304	2.016	2.181	0.150	0.302
37	2.364	2.711	0.309	-0.695	2.024	2.266	-0.037	-0.475
38	3.055	3.376	0.046	-0.930	2.295	2.551	0.585	0.238
39	1.612	1.638	0.154	0.048	1.368	1.442	0.284	0.192
40	2.233	2.339	-0.100	-0.375	2.139	1.748	0.262	-0.056
41	2.849	2.925	-0.341	-0.241	2.653	2.822	-0.504	-0.334
42	2.597	2.854	0.035	-0.430	2.135	2.371	0.634	0.170
43	4.062	4.110	-0.397	-0.413	3.889	3.878	-1.704	-1.735
44	3.672	3.984	0.753	-0.256	3.101	3.593	-0.216	-1.191
45	1.836	1.900	0.234	0.029	1.585	1.706	0.347	0.431
46	4.235	4.685	-1.535	-1.077	2.695	3.093	-0.785	-0.132
47	7.725	8.306	-0.865	-1.419	8.454	8.298	-2.108	-1.996
48	2.657	2.587	-1.185	-0.541	2.424	2.427	-0.885	-0.204
49	3.016	3.008	-0.870	-0.514	2.500	2.832	0.686	0.335
50	4.222	5.195	-0.054	-0.987	4.074	3.915	1.181	0.134
51	1.282	1.173	-0.203	-0.250	1.077	1.054	0.243	0.073
52	2.310	2.332	-0.058	-0.044	2.213	2.209	0.503	0.476
53	13.795	13.781	2.146	3.000	12.329	12.287	0.164	3.266

Table 4 continued **- Out-of-sample performances of disaggregated Imports UVIs**

	2 months missing				1 month missing			
Sector	*RMFSE*		*MFE*		*RMFSE*		*MFE*	
	AMI	AIR	AMI	AIR	AMI	AIR	AMI	AIR
54	11.010	10.701	1.838	2.133	11.796	11.699	1.396	-0.125
55	11.085	11.983	1.565	4.595	10.581	10.084	0.429	2.601
56	1.942	1.992	-0.160	-0.585	1.513	1.661	0.259	-0.279
57	10.709	10.060	1.345	1.073	10.560	8.993	-1.269	-0.301
58	2.977	3.180	-0.494	-0.519	2.926	2.931	0.098	0.068
60	9.628	10.938	1.192	0.847	10.616	12.520	1.221	-1.702
88	18.174	19.206	-2.529	2.306	17.760	20.311	4.082	6.522
90	23.852	25.089	1.421	-4.563	23.381	23.546	3.790	0.888
99	23.580	23.418	-2.307	-1.596	16.547	16.581	0.483	1.446

Table 5 - Out-of-sample performances of disaggregated Exports UVIs

	2 months missing				1 month missing			
Sector	*RMFSE*		*MFE*		*RMFSE*		*MFE*	
	AMI	AIR	AMI	AIR	AMI	AIR	AMI	AIR
1	4.077	3.213	0.225	-0.087	3.333	3.051	-0.490	-0.079
2	5.853	6.826	0.066	-1.690	6.043	6.174	1.623	0.224
3	2.250	2.517	0.208	-0.052	2.682	2.597	-0.011	-0.157
4	8.007	8.590	0.264	-1.756	6.399	7.152	-1.542	-2.669
5	73.855	71.615	28.936	23.188	49.350	42.918	21.201	13.273
7	12.551	11.713	-0.312	-1.900	10.639	11.360	3.049	0.469
8	3.227	3.364	0.675	-0.675	3.313	3.027	0.859	-0.218
9	4.342	5.098	-0.329	-1.526	3.828	4.348	0.719	-0.272
10	0.924	1.125	0.116	0.039	0.918	1.178	-0.005	-0.321
11	1.244	1.140	0.223	-0.265	0.975	0.976	0.261	-0.095
12	1.082	1.327	0.075	-0.131	0.865	0.964	-0.045	-0.278
13	1.101	1.140	0.223	-0.236	1.001	1.203	-0.262	-0.559
14	1.805	2.136	0.522	-0.169	1.139	1.476	-0.189	-0.520
15	1.969	2.159	0.470	-0.625	1.784	1.900	0.206	-0.105
16	5.653	6.034	-0.165	-1.100	4.687	4.939	1.201	-0.528
17	1.131	1.145	0.161	0.154	1.221	1.219	-0.039	-0.048
18	1.001	1.091	0.062	-0.026	1.205	1.168	0.111	0.138
19	1.635	1.470	0.632	0.270	1.276	1.576	0.080	-0.109
20	1.732	2.007	-0.100	-0.312	1.498	1.599	0.321	0.146
21	3.502	3.264	1.217	0.461	2.360	2.244	0.426	0.018
22	2.333	2.386	-0.167	-0.321	1.776	1.756	0.156	-0.017
23	2.222	1.775	0.670	0.363	2.348	1.967	-0.246	-0.356
24	0.942	1.149	-0.317	-0.530	0.696	0.659	-0.073	-0.110
25	2.637	2.630	0.242	0.281	2.532	2.402	-0.364	-0.175
26	9.265	8.263	2.421	0.615	9.901	9.284	-0.271	-2.037
27	1.927	1.840	-0.065	-0.263	1.517	1.365	-0.006	-0.077
28	1.449	1.608	-0.058	0.000	1.188	1.251	-0.044	-0.338
29	3.343	3.248	0.256	0.298	3.055	2.810	-0.036	-0.106
30	1.623	1.724	0.420	0.250	1.601	1.713	-0.133	-0.474
31	1.489	1.496	-0.044	0.091	1.124	1.128	0.274	0.376
32	1.114	1.177	0.019	-0.070	1.208	1.235	0.402	0.092
33	1.309	1.274	0.355	0.093	0.949	1.050	0.010	-0.342
34	1.143	1.159	-0.048	-0.269	0.939	0.984	-0.151	-0.339
35	1.812	1.865	0.190	-0.233	1.640	1.606	0.523	-0.314
36	1.299	1.302	-0.007	-0.035	1.214	1.200	0.132	0.117
37	2.049	2.148	-0.029	-0.195	1.739	1.734	0.292	-0.238
38	2.618	2.530	-0.124	-0.096	2.251	2.513	-0.151	-0.126
39	1.398	1.011	0.503	0.023	1.098	0.997	0.261	-0.005
40	1.461	1.706	0.095	0.350	1.552	1.650	0.446	0.491
41	1.820	1.671	-0.038	-0.021	1.516	1.374	-0.519	-0.527
42	1.280	1.201	0.180	0.038	1.136	1.088	0.511	0.169
43	6.881	6.933	-0.329	-0.084	5.777	6.503	0.873	1.830
44	3.815	3.433	1.337	0.545	4.732	4.238	1.833	1.205
45	1.715	1.733	0.201	-0.089	2.065	1.850	0.481	0.326
46	6.069	6.480	-1.388	-0.806	6.716	6.726	-0.896	-0.206
47	6.394	6.782	-0.898	0.111	6.208	6.611	-1.569	-0.600
48	6.074	6.789	-0.484	1.857	5.982	6.274	-1.412	0.953
49	2.196	2.402	0.089	0.475	2.076	2.250	-0.373	0.073
50	2.827	2.981	0.394	-0.235	3.245	2.973	-0.026	-0.901
51	0.958	0.970	-0.014	-0.213	0.790	0.783	-0.019	-0.187

Table 5 continued **- Out-of-sample performances of disaggregated Exports UVIs**

	2 months missing				1 month missing			
Sector	*RMFSE*		*MFE*		*RMFSE*		*MFE*	
	AMI	AIR	AMI	AIR	AMI	AIR	AMI	AIR
52	1.960	2.176	0.013	0.194	1.847	2.058	-0.043	0.054
53	9.458	10.942	0.061	2.037	8.755	9.936	0.705	2.054
54	6.993	7.005	1.352	1.000	8.175	8.276	0.947	0.553
55	6.990	7.663	-0.330	1.890	7.856	7.927	0.079	2.141
56	1.274	1.222	0.144	-0.060	1.378	1.411	0.133	-0.005
57	7.288	7.143	0.003	-0.450	6.833	6.913	2.130	2.068
58	1.450	1.362	0.121	-0.097	1.214	1.171	0.100	-0.165
60	17.746	19.815	5.358	1.794	16.626	17.182	1.665	2.302
88	18.428	19.685	5.455	5.656	14.033	13.387	7.756	5.727
90	12.796	13.484	4.440	-0.436	12.205	12.604	4.520	1.000
99	13.814	12.814	-2.469	-1.775	10.455	11.872	-4.235	-5.602

Table 6 - Frequency table of SEAS. ARIMA models identified by TRAMO: Imports

	(1,1,1)	(1,1,0)	(0,1,1)	(1,0,0)	(1,0,1)	(1,0,0)	(0,1,1)	(0,1,1)	(0,1,1)	(0,1,0)	others
Sector	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,0)	(0,0,1)	(0,1,1)	(1,0,0)	(0,0,1)	(0,1,1)	
Total	-	37	-	-	-	-	4	-	-	-	7
1	-	-	-	-	-	-	16	7	-	5	20
2	-	5	42	-	-	-	-	-	-	-	1
3	-	-	-	-	6	29	-	-	8	-	5
4	-	-	-	-	-	-	41	-	-	3	4
5	-	9	37	-	-	-	-	-	-	2	-
6	-	-	48	-	-	-	-	-	-	-	-
7	-	19	26	-	-	-	1	-	-	2	-
8	-	-	-	-	-	-	1	-	-	-	47
9	-	-	45	-	-	-	3	-	-	-	-
10	-	-	9	18	13	-	-	-	-	5	3
11	-	3	9	-	24	-	9	-	-	3	-
12	-	-	48	-	-	-	-	-	-	-	-
13	-	5	-	-	-	-	4	-	-	-	39
14	-	-	37	-	-	-	11	-	-	-	-
15	-	-	38	-	-	-	7	-	-	-	3
16	-	-	30	-	-	-	6	-	-	-	12
17	-	-	-	-	-	-	45	2	-	-	1
18	-	5	-	-	-	-	13	-	-	12	18
19	-	-	-	-	-	-	5	43	-	-	-
20	-	-	-	-	-	-	34	-	-	13	1
21	-	-	7	-	-	-	11	-	-	14	16
22	8	-	21	-	-	-	12	2	-	-	5
23	-	-	33	-	-	-	11	-	4	-	-
24	-	19	-	-	-	-	-	-	-	-	29
25	-	2	36	-	3	-	1	-	-	-	6
26	-	-	32	-	-	-	16	-	-	-	-
27	-	10	2	-	-	-	4	-	-	-	32
28	-	-	12	-	-	-	36	-	-	-	-
29	-	-	25	-	-	-	3	10	1	-	9
30	-	31	-	-	14	-	-	-	-	1	2
31	-	-	40	-	-	-	-	-	-	-	8
32	-	-	48	-	-	-	-	-	-	-	-
33	-	16	25	-	-	-	7	-	-	-	-
34	4	-	40	-	-	-	4	-	-	-	-
35	-	-	-	-	-	-	4	19	-	-	25
36	-	-	38	-	-	-	-	10	-	-	-
37	-	30	-	-	-	-	2	-	-	-	16
38	-	-	24	-	-	-	15	9	-	-	-
39	-	-	42	-	-	-	6	-	-	-	-
40	-	2	-	-	-	-	18	21	-	-	7
41	-	-	-	-	-	-	44	-	-	-	4
42	-	-	24	-	-	-	6	1	3	-	14
43	-	-	-	-	-	-	45	-	-	2	1
44	-	-	42	4	-	-	2	-	-	-	-
45	-	-	48	-	-	-	-	-	-	-	-
46	-	-	19	-	-	-	-	24	5	-	-

Table 6 continued **- Frequency table of SEAS. ARIMA models identified by TRAMO: Imports**

Sector	(1,1,1) (0,0,0)	(1,1,0) (0,0,0)	(0,1,1) (0,0,0)	(1,0,0) (0,0,0)	(1,0,1) (0,0,0)	(1,0,0) (0,0,1)	(0,1,1) (0,1,1)	(0,1,1) (1,0,0)	(0,1,1) (0,0,1)	(0,1,0) (0,1,1)	others
47	-	-	42	2	-	-	3	-	-	-	1
48	-	-	39	-	-	-	9	-	-	-	-
49	-	-	8	-	-	-	12	28	-	-	-
50	-	-	-	-	-	-	-	34	-	-	14
51	-	-	46	-	-	-	2	-	-	-	-
52	-	-	35	-	-	-	13	-	-	-	-
53	-	-	13	-	-	-	13	4	1	-	17
54	11	-	23	-	-	-	12	-	-	-	2
55	-	-	18	-	-	-	29	-	-	-	1
56	-	-	6	-	-	-	30	-	-	-	12
57	-	-	-	-	-	-	42	-	-	-	6
58	-	24	8	-	-	-	-	-	-	-	16
60	-	-	15	-	-	-	3	4	-	-	26
88	1	-	5	-	-	-	-	-	6	-	36
90	-	-	8	-	-	-	19	6	14	-	1
99	-	-	5	-	-	-	37	5	-	-	1
Sum	**24**	**180**	**1198**	**24**	**60**	**29**	**667**	**229**	**42**	**62**	**461**
%	0.81	6.05	40.26	0.81	2.02	0.97	22.41	7.69	1.41	2.08	15.49

Table 7 - Frequency table of SEAS. ARIMA models identified by TRAMO: Exports

Sector	(1,1,1) (0,0,0)	(1,1,0) (0,0,0)	(0,1,1) (0,0,0)	(1,0,0) (0,0,0)	(1,0,1) (0,0,0)	(2,1,0) (0,0,0)	(0,1,1) (0,1,1)	(0,1,1) (1,0,0)	(0,1,1) (0,0,1)	(0,1,0) (0,1,1)	others
Total	-	-	-	-	-	-	26	-	-	21	1
1	-	-	10	-	-	-	-	-	-	7	31
2	-	-	43	-	-	-	1	-	-	-	4
3	-	-	-	-	-	-	32	3	-	-	13
4	-	-	-	2	-	-	13	5	-	-	28
5	7	-	10	-	4	-	1	-	-	-	26
7	-	8	1	-	-	-	23	-	-	-	16
8	-	-	2	-	-	34	3	-	-	-	9
9	-	36	2	-	-	-	-	-	-	-	10
10	24	4	-	-	-	-	-	-	-	3	17
11	-	-	27	-	1	-	7	1	-	-	12
12	-	38	-	-	10	-	-	-	-	-	-
13	-	1	28	-	-	-	-	-	-	-	19
14	-	7	-	-	-	7	4	-	-	2	28
15	-	-	20	1	-	-	23	-	-	-	4
16	5	-	28	-	-	-	1	-	-	-	14
17	-	-	-	-	-	-	48	-	-	-	-
18	-	-	-	-	-	-	9	-	-	13	26
19	-	-	-	-	-	-	-	-	-	-	48
20	-	-	-	-	-	-	-	-	-	-	48
21	-	-	-	-	-	-	17	-	-	-	31
22	-	-	-	-	-	-	47	-	-	-	1
23	-	-	10	18	-	-	10	-	-	-	10
24	1	-	-	-	-	12	-	-	-	-	35
25	-	-	-	-	-	-	41	-	-	-	7
26	-	-	34	-	-	-	4	-	-	-	10
27	-	2	16	-	-	-	-	1	7	-	22
28	-	-	37	-	-	-	11	-	-	-	-
29	-	-	20	-	21	-	1	-	-	-	6
30	-	24	7	-	-	-	17	-	-	-	-
31	-	-	33	-	-	-	4	-	-	-	11
32	-	-	2	-	-	-	-	43	-	3	-
33	-	-	11	-	-	-	37	-	-	-	-
34	-	-	-	-	-	-	43	-	-	1	4
35	7	-	19	-	-	-	4	6	2	-	10
36	-	-	-	-	-	-	40	-	-	-	8
37	16	-	-	-	-	8	-	-	-	-	24
38	-	-	20	6	17	-	5	-	-	-	-
39	-	2	38	-	-	-	3	5	-	-	-
40	-	-	30	-	-	-	5	12	-	-	1
41	2	-	-	-	-	-	6	15	9	-	16

Table 7 continued **- Frequency table of SEAS. ARIMA models identified by TRAMO: Exports**

Sector	(1,1,1) (0,0,0)	(1,1,0) (0,0,0)	(0,1,1) (0,0,0)	(1,0,0) (0,0,0)	(1,0,1) (0,0,0)	(2,1,0) (0,0,0)	(0,1,1) (0,1,1)	(0,1,1) (1,0,0)	(0,1,1) (0,0,1)	(0,1,0) (0,1,1)	others
42	8	-	13	-	11	-	10	-	-	-	6
43	-	-	16	2	-	-	22	-	-	-	8
44	-	-	9	-	-	-	12	27	-	-	-
45	-	2	22	-	-	-	21	-	-	-	3
46	-	-	48	-	-	-	-	-	-	-	-
47	-	-	17	-	-	-	4	-	27	-	-
48	-	-	15	19	10	-	3	-	-	-	1
49	-	-	20	13	12	-	3	-	-	-	-
50	-	-	-	-	-	-	19	13	-	-	16
51	-	-	-	-	-	-	30	-	-	-	18
52	4	-	2	-	-	-	31	1	-	-	10
53	3	-	33	-	-	-	-	-	-	-	12
54	-	-	35	-	-	-	8	-	-	-	5
55	-	-	44	-	-	-	4	-	-	-	-
56	-	-	-	-	-	-	38	10	-	-	-
57	7	-	1	-	-	-	22	12	-	-	6
58	-	-	-	-	-	3	1	12	-	-	32
60	-	-	17	-	22	-	3	-	-	1	5
88	-	-	37	-	-	-	1	-	-	-	10
90	-	-	2	-	17	-	-	-	-	-	29
99	-	-	-	-	-	-	24	-	-	-	24
Sum	**84**	**124**	**779**	**61**	**125**	**64**	**716**	**166**	**45**	**30**	**734**
%	2.87	4.23	26.61	2.08	4.27	2.19	24.45	5.67	1.54	1.02	25.07

Table 8 - Out-of-sample performance of aggregated UVIs: direct and indirect approaches

	2 months missing				1 month missing			
	RMFSE		*MFE*		*RMFSE*		*MFE*	
	AMI	AIR	AMI	AIR	AMI	AIR	AMI	AIR
Imports								
direct	1.237	1.294	0.268	-0.266	1.475	1.494	-0.220	-0.298
indirect	1.079	1.300	0.159	-0.269	0.863	0.875	-0.021	-0.223
Exports								
direct	0.709	0.706	0.059	-0.046	0.635	0.633	0.112	0.073
indirect	0.727	0.734	0.112	0.030	0.606	0.603	0.140	0.045

Table 9 - OLS estimates using the 143 observations 1996:02–2007:12 Dependent variable: Δy_t

	Coefficient	Std. Error	t -ratio	p-value
c	0.001199	0.002360	0.5080	0.6123
Δx_t	0.352671	0.028835	12.2304	0.0000
Δx_{t-1}	0.479958	0.035712	13.4395	0.0000
Δx_{t-7}	0.089261	0.028601	3.1209	0.0022
Δz_{t-2}	0.087039	0.035445	2.4556	0.0153
Δy_{t-1}	-0.165582	0.052856	-3.1327	0.0021

Mean of dependent variable	0.009346
S.D. of dependent variable	0.055938
Sum of squared residuals	0.102845
Standard error of the regression ($\hat{\sigma}$)	0.027399
Unadjusted R^2	0.768536
Adjusted $\overline{R}^2$	0.760088
$F(5,137)$	90.9767
Durbin–Watson statistic	2.18120
First-order autocorrelation coeff.	-0.09441

Table 10 - Improvements of forecasting accuracy using the ADL model

	2 months missing		1 month missing	
	RMFSE	*MFE*	*RMFSE*	*MFE*
ARIMA model				
product 6	6.962	1.103	5.532	0.309
total imports	1.079	0.159	0.863	-0.021
ADL model				
product 6	4.037	0.508	3.200	0.445
total imports	0.883	0.083	0.817	-0.018

Figure 1 - Monthly UVI of imports. Period: 1996:01-2007:12.

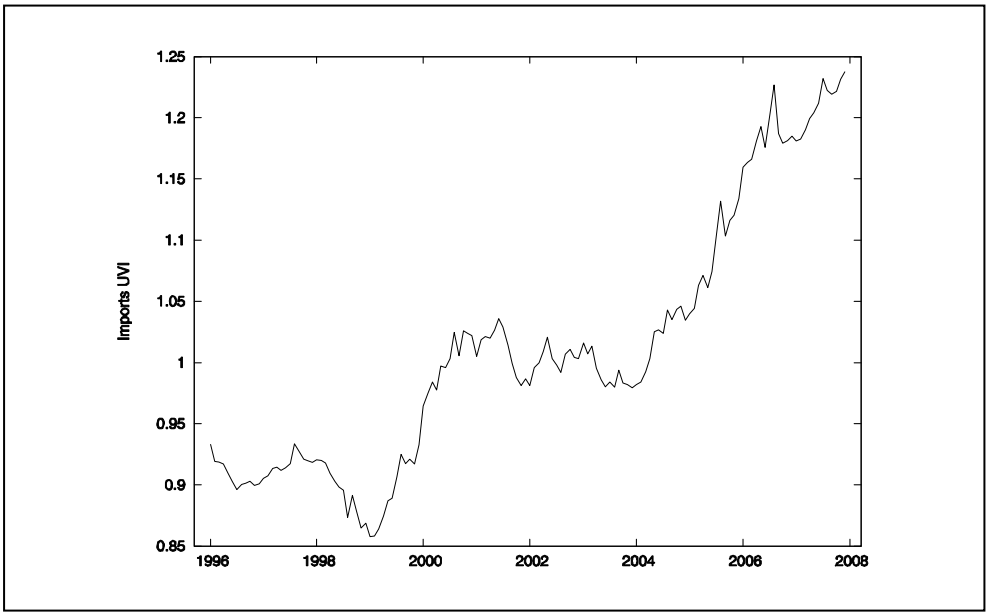


Figure 2 - Monthly UVI of exports. Period: 1996:01-2007:12.

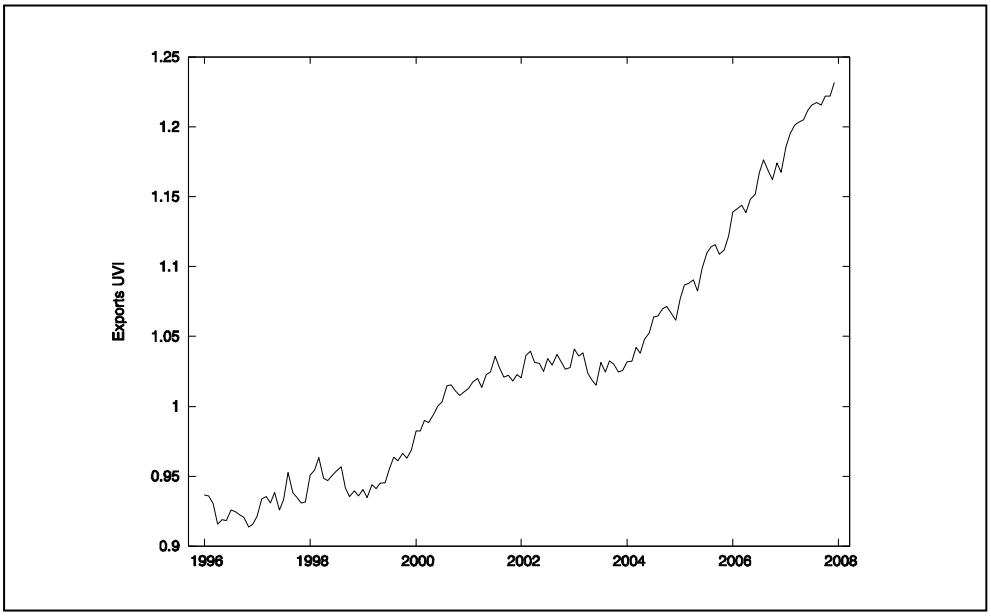


Figure 3 - Imports UVI, crude oil price index and gas price index. Period: 1996:1-2007:12

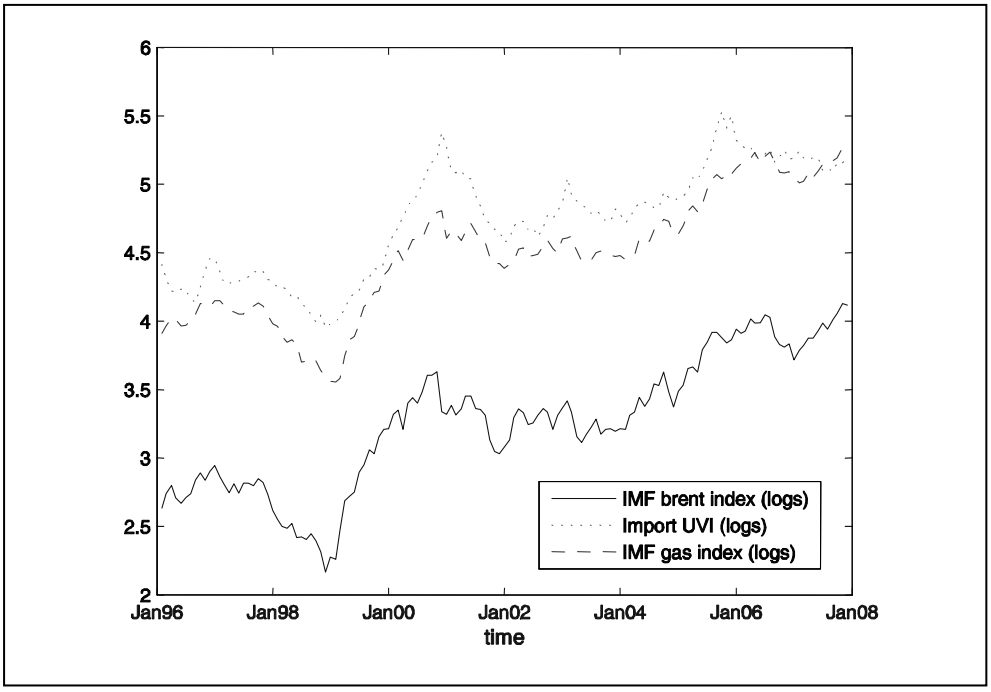
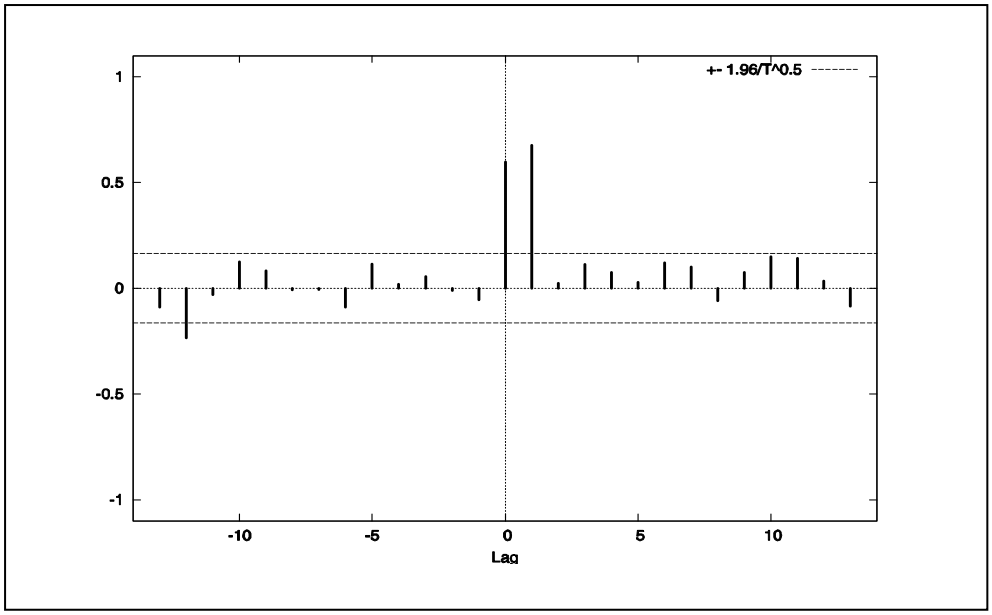


Figure 4 - Correlation of imports UVI and lags of crude oil price index



Norme redazionali

La Rivista di Statistica Ufficiale pubblica contributi originali nella sezione "Temi trattati" ed eventuali discussioni a largo spettro nella sezione "Interventi". Possono essere pubblicati articoli oggetto di comunicazioni a convegni, riportandone il riferimento specifico. Gli articoli devono essere fatti pervenire al Comitato di redazione delle pubblicazioni scientifiche Istat corredati, a parte, da una nota informativa dell'Autore contenente: appartenenza ad istituzioni, attività prevalente, qualifica, indirizzo, casella di posta elettronica, recapito telefonico e l'autorizzazione alla pubblicazione firmata dagli Autori. Ogni articolo prima della pubblicazione dovrà ricevere il parere favorevole di un referente scelto tra gli esperti dei diversi temi affrontati. Gli originali, anche se non pubblicati, non si restituiscono.

Per l'impaginazione dei lavori gli autori sono tenuti a conformarsi rigorosamente agli standard editoriali fissati dal Comitato di redazione e contenuti nel file Template.doc disponibile on line o su richiesta. In base a tali standard la lunghezza dei contributi originali per entrambe le sezioni dovrà essere limitata entro le 30-35 pagine.

Tutti i lavori devono essere corredati di un sommario nella lingua in cui sono redatti (non più di 12 righe); quelli in italiano dovranno prevedere anche un *Abstract* in inglese. La bibliografia, in ordine alfabetico per autore, deve essere riportata in elenco a parte alla fine dell'articolo. Quando nel testo si fa riferimento ad una pubblicazione citata nell'elenco, si metta in parentesi tonda il nome dell'autore, l'anno di pubblicazione ed eventualmente la pagina citata. Ad esempio (Bianchi, 1987, Rossi, 1988, p. 55). Quando l'autore compare più volte nello stesso anno l'ordine verrà dato dall'aggiunta di una lettera minuscola accanto all'anno di pubblicazione. Ad esempio (Bianchi, 1987a, 1987b).

Nella bibliografia le citazioni di libri e articoli vanno indicate nel seguente modo. Per i libri: cognome dell'autore seguito dall'iniziale in maiuscolo del nome, il titolo in corsivo dell'opera, l'editore, il luogo di edizione e l'anno di pubblicazione. Per gli articoli: dopo l'indicazione dell'autore si riporta il titolo tra virgolette, il titolo completo in corsivo della rivista, il numero del fascicolo e l'anno di pubblicazione. Nei riferimenti bibliografici non si devono usare abbreviazioni.

Nel testo dovrà essere di norma utilizzato il corsivo per le parole in lingua straniera e il corsivo o grassetto per quei termini o locuzioni che si vogliano porre in particolare evidenza (non vanno adoperati, per tali scopi, il maiuscolo, la sottolineatura o altro).

Gli articoli pubblicati impegnano esclusivamente gli Autori, le opinioni espresse non implicano alcuna responsabilità da parte dell'Istat.

La proprietà letteraria degli articoli pubblicati spetta alla Rivista di statistica ufficiale.

E' vietata a norma di legge la riproduzione anche parziale senza autorizzazione e senza citarne la fonte.

Per contattare il Comitato di redazione delle pubblicazioni scientifiche Istat e per inviare lavori: rivista@istat.it. Oppure scrivere a:
Segreteria del Comitato di redazione della Rivista di Statistica Ufficiale
All'attenzione di Gilda Sonetti
Via Cesare Balbo, 16
00184 Roma

La Rivista di Statistica Ufficiale accoglie lavori che hanno come oggetto la misurazione e la comprensione dei fenomeni sociali, demografici, economici ed ambientali, la costruzione di sistemi informativi e di indicatori come supporto per le decisioni pubbliche e private, nonché le questioni di natura metodologica, tecnologica e istituzionale connesse ai processi di produzione delle informazioni statistiche e rilevanti ai fini del perseguimento dei fini della statistica ufficiale.

La Rivista di Statistica Ufficiale si propone di promuovere la collaborazione tra il mondo della ricerca scientifica, gli utilizzatori dell'informazione statistica e la statistica ufficiale, al fine di migliorare la qualità e l'analisi dei dati.

La pubblicazione nasce nel 1992 come collana di monografie "Quaderni di Ricerca ISTAT". Nel 1999 la collana viene affidata ad un editore esterno e diviene quadrimestrale con la denominazione "Quaderni di Ricerca - Rivista di Statistica Ufficiale". L'attuale denominazione, "Rivista di Statistica Ufficiale", viene assunta a partire dal n. 1/2006 e l'Istat torna ad essere editore in proprio della pubblicazione.

1B012011003000000

ISSN 1828-1982

€ 10,00