

**Produzione di rifiuti in Italia:
una stima preliminare basata sul MUD 2010**

MUD 2010: un prototipo per la stima statistica

Ignazio Drudi
Università di Bologna, Dipartimento di Scienze Statistiche

5 novembre 2010, Ecomondo ~ Rimini Fiera, Padiglione D5 - Stand 115

Il gruppo di lavoro

ECOCERVED:

Manuela Medoro, Donato Molino, Jean Sangiuliano

DIPARTIMENTO DI SCIENZE STATISTICHE:

Ignazio Drudi, Giorgio Tassinari

RES Coop:

Fabrizio Alboni, Maria Vittoria Sardella

Avvertenza importante:

Lo scopo principale di questa relazione è la presentazione della metodologia di stima precoce per la produzione di rifiuti da attività produttiva. Come risulterà evidente, il problema è particolarmente complesso e richiede procedure statistiche non banali

Inoltre, per il fatto che la attendibilità delle previsioni dipende in modo cruciale dalla quantità di informazioni disponibili, si è scelto di utilizzare le dichiarazioni disponibili fino all'ultima data utile (circa 10 gg fa)

Pertanto, mentre la metodologia presentata è ormai stabilizzata e non richiede ulteriori affinamenti, la sua implementazione in termini di algoritmi di calcolo e verifica dei risultati è ancora "in progress". In più il processo di verifica a posteriori è reso ancora più difficoltoso dal fatto che i domini di stima previsti sono molto numerosi.

Quindi, le stime numeriche che qui sono presentate sono da intendersi come "ANTICIPAZIONI PROVVISORIE" e hanno principalmente il compito di validare la procedura proposta

Il rilascio ufficiale delle previsioni validate avverrà in una apposita pubblicazione, che sarà prodotta a breve

Sommario:

1. Premessa e finalità
2. L'informazione disponibile: dichiarazioni telematiche e "bonificate"
3. Il problema della previsione da più fonti e la strategia di previsione generale
4. Modelli di previsione per i dati bonificati
5. Modelli di previsione per le dichiarazioni telematiche
6. La combinazione delle stime

Premessa:

La relazione che segue richiede alcune precisazioni preliminari:

1. E' sintetica, anche se non brevissima, per problemi di tempo e quindi alcuni temi saranno solo accennati e, rispetto al lavoro compiuto, si presenteranno solo i risultati più significativi
2. E' rivolta ad un pubblico di "non statistici", pertanto si privilegerà la chiarezza e la semplicità, rispetto all'approfondimento formale delle tecniche utilizzate. Quando si dovranno affrontare aspetti di metodo, si cercherà di utilizzare un approccio "intuitivo" e non analitico
3. Il focus principale sarà rivolto al problema della previsione e alla spiegazione di come è stata ottenuta, dando conto solo di una parte delle scelte compiute
4. A questa relazione seguirà un rapporto dettagliato che consentirà agli interessati di approfondire la metodologia utilizzata

Finalità della collaborazione tra
ECOCERVED e Dipartimento di Scienze Statistiche
è la qualificazione di ECOCERVED
come produttore di dati e informazioni

Che cosa significa? Una prima precisazione:

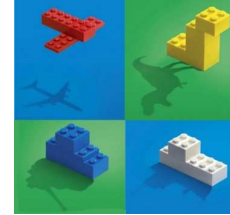
Dato (livello micro):

- Osservazione, rilevazione, misura diretta di un evento



Dato Statistico (livello meso):

- Insieme di uno o più dati messi in relazione e interpretati nell'ambito di un contesto in modo da avere un significato



Informazione (livello macro):

- Sistema di dati statistici finalizzato alla formulazione di un giudizio, di una strategia, di una previsione, di una decisione



Vi sono

Molti produttori di dati


Alcuni produttori di dati statistici

Pochi produttori di informazioni

Il passaggio da dati a informazioni è il contenuto della Statistica

I passi del processo:

Per ora la collaborazione si è occupata di questo

1) Produzione di dati 	a) Analisi ed "error profile" del processo di produzione dei dati b) Valutazione, validazione e miglioramento delle tecniche statistiche di "bonifica" c) Sistema statistico di stime "precoci" da dati parziali
--	--

Oggi presentiamo i primi risultati su questo punto

L'informazione disponibile oggi:

I Modelli MUD:

- Raccolgono una notevole quantità di dati sull'intero flusso del ciclo dei rifiuti, dalla produzione, al trasporto, al conferimento fino allo smaltimento.
- In sostanza, ogni soggetto coinvolto in ciascuna fase è tenuto a presentare un dichiarazione
- AI FINI DELLA PREVISIONE, sono state considerate le **quantità di rifiuti** dichiarati da unità locali **produttive**.
- Sono, quindi, state escluse le fasi "a valle" della produzione (trasporto, conferimento e smaltimento) che ai fini della stima costituirebbero delle duplicazioni
- Sono stati esclusi, altresì, i rifiuti di origine "**non produttiva**" ad es. i rifiuti solidi urbani

In dettaglio sono stati esclusi dalla stima:

Le attività:

Recupero e riciclaggio
Costruzioni
Commercio rottami, cascami
Smaltimento rifiuti

I rifiuti:

Rifiuti prodotti da impianti di trattamento dei rifiuti...etc.
Rifiuti delle operazioni di costruzioneetc.

Tranne quelle quantità dichiarate nelle attività produttive non escluse

L'informazione disponibile:

Le modalità di acquisizione delle dichiarazioni:

- Due modalità per la presentazione: invio telematico e spedizione via posta.
- Le modalità differiscono, ovviamente, per la tempestività e per il processo di revisione e controllo dei dati.
- **Dichiarazione Magnetica (*file*) consegnata per via telematica:** disponibile in "tempo reale" e controlli di coerenza formale immediati, mediante apposito software. D'ora in poi denomineremo tali dichiarazioni come "TELEMATICHE"
- **Dichiarazione Cartacea o Magnetica (*CD-ROM/floppy disk ecc.*) spedita via posta:** disponibile in tempi più lunghi (registrazione) e controlli "a posteriori" (correzioni => "bonifica") anch'essi più lunghi e difficoltosi.

Alcuni dati:

- Nel 2008 la percentuale di dichiarazioni telematiche è circa il 11% del totale
- Le dichiarazioni (presentate via telematica o per posta) sono disponibili in forma "BONIFICATA" circa 15-18 mesi dopo la scadenza per la loro presentazione

L'informazione disponibile:

Ci troviamo, dunque, di fronte a due fonti di informazioni con caratteristiche diverse:

L'una è tempestiva e parziale, l'altra è "completa" ma disponibile con ritardo

Le dichiarazioni telematiche hanno la capacità di evidenziare rapidamente cambiamenti di tendenza, che però potrebbero non rappresentare tutti i soggetti produttori di rifiuti

Le dichiarazioni bonificate mostrano il trend generale e verificato, ma lo fanno con molto ritardo

Il processo di previsione qui proposto cerca di sfruttare le caratteristiche positive delle due fonti, combinandole secondo criteri di ottimalità.

Anticipando la logica del processo, utilizzeremo le bonificate per una proiezione del trend generale e le telematiche per inglobare nella previsione i segnali precoci di cambiamento di quel trend

Gli andamenti, principali caratteristiche:

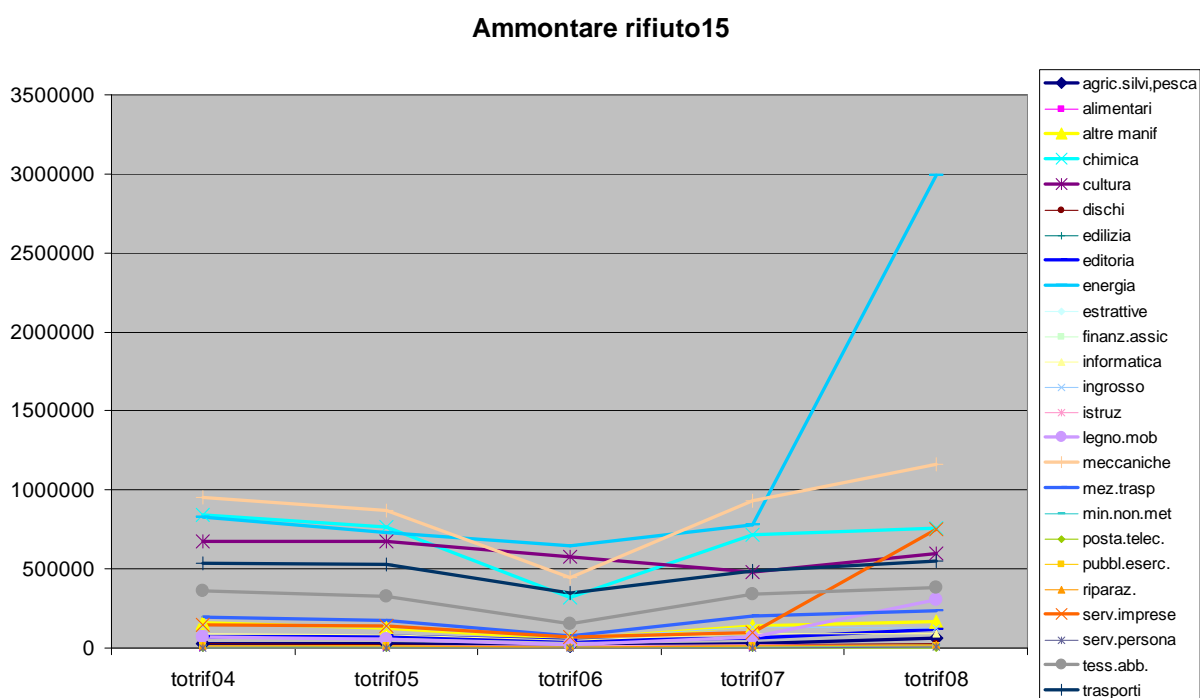
- 1) Almeno apparentemente, fenomeno con altissima variabilità: nel tempo, nello spazio, per settori
- 2) Variabilità comune sia alle dichiarazioni telematiche che alle bonificate
- 3) Fenomeno soggetto a variazioni normative e procedurali

⇒ Procedura di previsione complessa

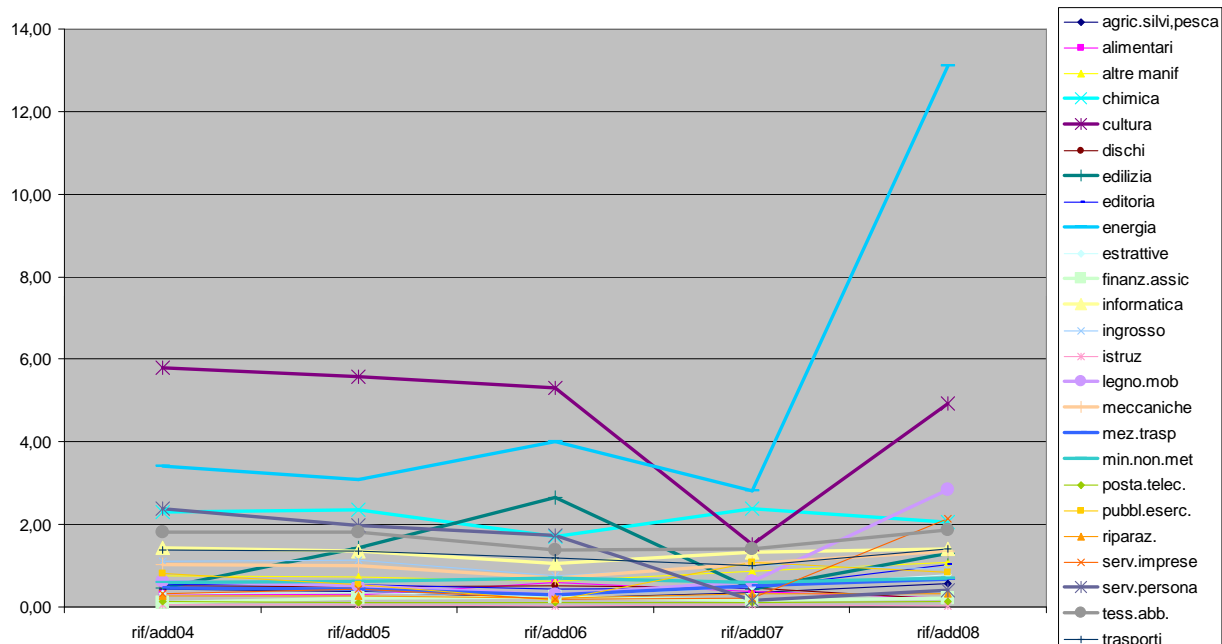
⇒ Utilizzare il massimo di informazione disponibile

Un esempio (il rifiuto "15")

La variabilità: alcuni esempi sul totale dei rifiuti dichiarati (BONIFICATE)



rifiuti per addetto



Omettiamo per brevità la variabilità per regione, ma naturalmente è maggiore

Il fenomeno è amplificato se si restringe il campo alle telematiche

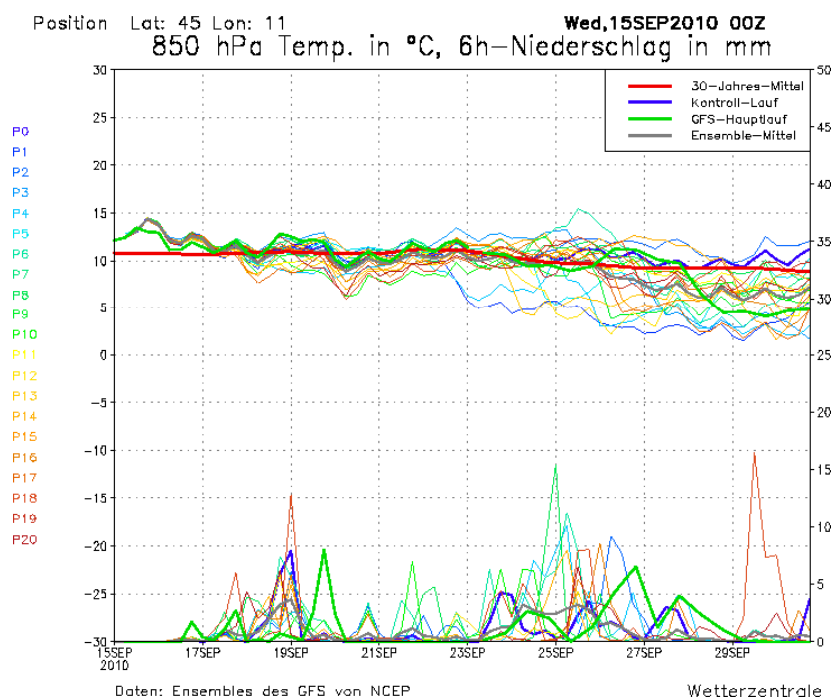


Stima in situazione di grande incertezza

Bilanciamento di diverse previsioni mediante un sistema di ponderazione

Meccanismi di bilanciamento differenti per dominio di stima

Situazione simile a quella delle previsioni del tempo.....



Come si bilanciano diverse stime ?

Molte proposte, essenzialmente, fanno riferimento alla cosiddetta stima dei

MINIMI QUADRATI GENERALIZZATI

In sintesi si tratta di:

Una tecnica di stima che attribuisce alle singole osservazioni un peso inversamente proporzionale alla loro variabilità

A prima vista sembra complicata (lo è...un po') ma vedremo che, in questo caso, il meccanismo dei pesi risulterà piuttosto semplice

Modelli di previsione per i dati bonificati

Obiettivo: estrarre il trend generale fino all'ultimo anno disponibile e "proiettarlo" in avanti di un anno

I dati: poche osservazioni temporali (2004-2008) molte osservazioni sezionali (per ciascun anno), circa 70.000 unità locali ogni anno, escluso il 2006 (che vedremo in seguito)

Prima approssimazione =

Tecnica di previsione: stima di modelli da dati di "panel", in particolare le cosiddette "Dynamic panel estimation". Cioè stima del trend mediante **variabili ritardate**, oppure mediante variabili strumentali.

In simboli:

$$y_{i,t} = \alpha + \beta y_{i,t-1} + \varepsilon_{i,t}$$

oppure

$$y_{i,t} = \alpha + \beta x_{i,t-1} + \varepsilon_{i,t}$$

Il valore y da prevedere per il rifiuto i nell'anno t dipende dal valore dello stesso rifiuto dell'anno precedente

Oppure

Il valore y da prevedere per il rifiuto i nell'anno t dipende dal valore di altre variabili (es. addetti o imprese) dell'anno precedente, oppure anche dello stesso anno se note.

Gli anni già disponibili funzionano come benchmark per trovare il "miglior" valore per α e β

La previsione con dati di panel consente anche di specificare delle caratteristiche "tipiche" di ogni soggetto (invarianti nel tempo) e caratteristiche tipiche di ciascun periodo (invarianti tra i soggetti)

Cioè:

$$y_{i,t} = \alpha_i + \gamma_t + \beta x_{i,t-1} + \varepsilon_{i,t}$$

Senza entrare nel dettaglio, questo consente di "scomporre" la variabilità complessiva dei dati, attribuendone una prima quota alla eterogeneità dei soggetti, una seconda alla eterogeneità dei tempi.

Tolte queste due fonti di variabilità, la quota residuale sarà allora quella "veramente" (cioè intrinsecamente) legata ai dati, cioè il principale ostacolo per la previsione)

.... *Una piccola complicazione tecnica: vi sono diversi algoritmi per stimare questi modelli. Essi variano in funzione delle ipotesi a priori sulla variabilità complessiva.....Sono stati utilizzati e comparati tutti (poco meno di 20)*

Il data-base utilizzato per la stima è costituito dai dati di produzione di rifiuti per gli anni 2004-2008 suddiviso per tipo di rifiuto (20 categorie), attività economica dell'unità locale (Ateco 2 cifre, 55 categorie) e regione (20 categorie).

Quindi il data-set individuale su cui sono state fatte le stime consta di oltre 100.000 dati elementari. E ciò che fin qui abbiamo denominato "soggetto" è identificato dall'incrocio di regione-attività-tipo di rifiuto

Ciò è importante al fine di interpretare i risultati, soprattutto per quanto concerne la quota di variabilità legata ai soggetti.

..... *Una avvertenza, forse ovvia, il livello micro dei dati utilizzati per la stima NON è il livello al quale le previsioni saranno significative. I dati analitici sono il materiale col quale si costruiscono le previsioni aggregate.....*

I risultati delle stime "panel". In sintesi:

1. Con pochissime eccezioni le stime forniscono risultati molto simili. Tra l'altro le eccezioni sono assai significative, nel senso che confermano quanto descritto nel punto successivo, ma non vengono approfondite qui
2. La distribuzione dei residui di tipo log-normale
3. La scomposizione della varianza complessiva nelle sue componenti fornisce una indicazione evidente e inequivoca:

La consistente variabilità complessiva è quasi totalmente da attribuire alle due componenti (soggetti e tempi), cioè gli andamenti macro sono influenzati da comportamenti micro che risultano sistematici e prevedibili.

In particolare per i soggetti risulta significativa la variabilità legata al settore di produzione (Ateco) mentre per i tempi è (quasi) sempre l'anno 2006 a risultare sistematicamente significativo

I diversi comportamenti micro, correttamente stimati, hanno una spiccata tendenza a compensarsi, cioè procedendo con opportune aggregazioni il loro effetto scompare

La semplificazione del processo di previsione:

Le evidenze che emergono dalle stime panel, consentono di semplificare l'approccio alla previsione: opportune trasformazioni e aggregazioni condotte sui dati elementari porteranno ad eliminare l'effetto delle eterogeneità sistematiche e, quindi, a "pulire" i dati in modo da poter utilizzare tecniche di previsione "semplici"

Le trasformazioni sono :

Per tutti i dati:

1. Passaggio ai logaritmi

Per i soggetti:

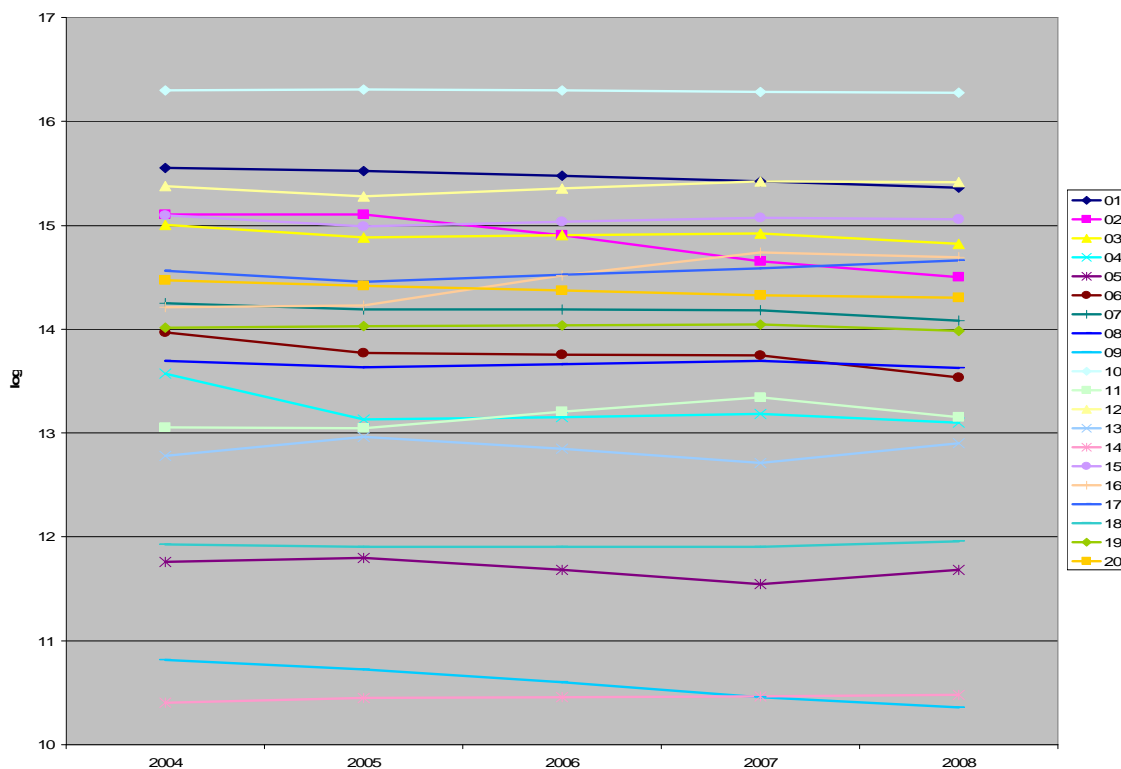
1. passaggio dai dati assoluti agli scarti dalla media (per ciascun soggetto)
2. Aggregazione dei dati in funzione delle stime sintetiche da ottenere
3. Ritorno ai dati assoluti mediante la somma della media aggregata

Per i tempi:

1. Perequazione del dato 2006 mediante imputazione della media degli anni adiacenti

Il risultato delle operazioni descritte ha un effetto evidente:
le serie sono "regolari" con lievi effetti di "trend"

Produzione rifiuti per tipo di rifiuto



La previsione:

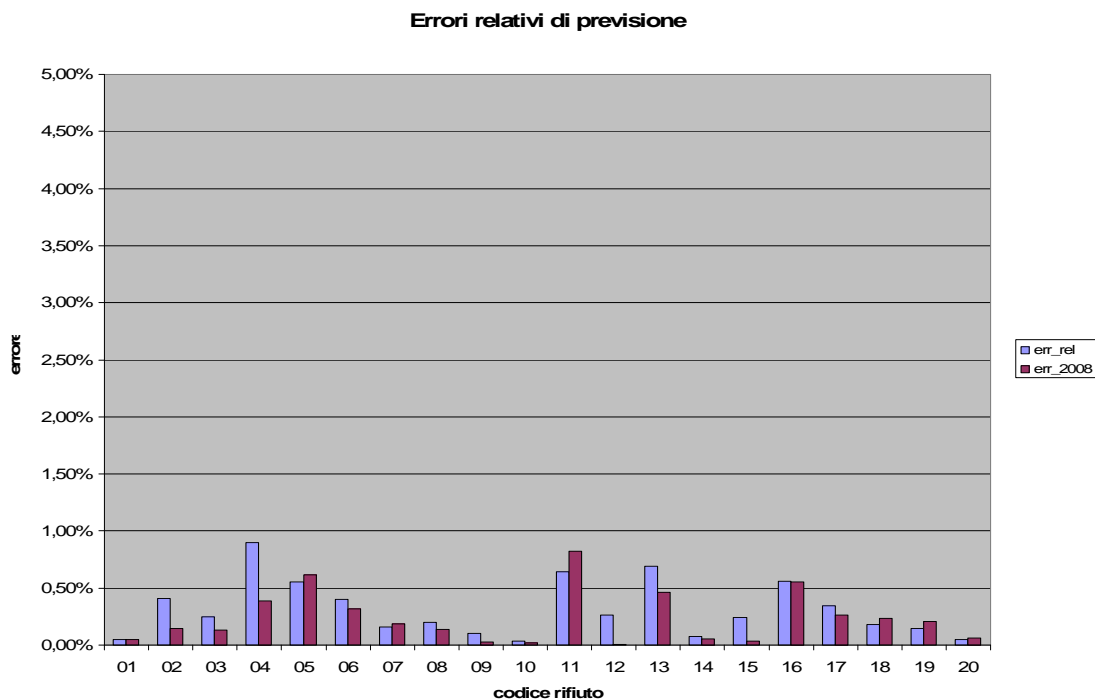
Di fronte alla regolarità evidente degli andamenti così modificati è giustificato un metodo di stima che fa perno sulla estrapolazione del trend, cioè una previsione basata sulla tendenza osservata.

D'altra parte è possibile verificare l'adeguatezza del modello di trend mediante 2 indicatori:

1. L'errore cumulativo di stima cioè la media (dei quadrati) delle differenze PER OGNI ANNO tra i valori osservati e i valori stimati col modello di trend, rapportato al valore osservato (nei grafici indicato con err.rel.)
2. L'errore relativo all'ultimo anno disponibile (2008), rispetto alla previsione ottenuta dal trend degli anni precedenti (nei grafici indicato come err2008)

L'ipotesi tipica è che minori sono gli errori, migliore è la stima
Vi sono alcune soglie "tipiche" (5%, 10%)

I risultati sono molto soddisfacenti, tutti gli errori sono < 1%



I domini di stima:

Questo processo di aggiustamento e di previsione che abbiamo descritto è stato utilizzato per tutti i domini di stima che costituiscono l'obiettivo del lavoro, che prevede previsioni per:

1. Tipo di rifiuto
2. Tipo di rifiuto e settore economico
3. Tipo di rifiuto e circoscrizione territoriale
4. Tipo di rifiuto, settore economico e circoscrizione territoriale

Com'è evidente, lo schema è gerarchico ed ogni livello superiore funge da vincolo per le previsioni del livello inferiore, nel senso che, ad esempio, la somma delle previsioni per settore deve corrispondere, rifiuto per rifiuto, al totale stimato per tipo di rifiuto

I domini di stima, il problema della numerosità:

Per poter procedere alla previsione nei domini di stima elencati occorre premettere una considerazione.

E' ben noto che il fattore principale che determina la stabilità delle previsioni è il fatto di poter disporre di una quantità sufficiente di informazioni elementari

E' evidente che più si specializza il dominio di stima, minore è la numerosità delle informazioni che si troveranno in ciascun dominio. Banalmente il numero di osservazioni disponibili per la stima del rifiuto 1 prodotto nel settore AB sarà inferiore a quello disponibile per l'intero rifiuto 1. Anzi, in diversi casi il dominio risulterà vuoto o del tutto inconsistente

Ciò significa che in alcuni domini la stima è impossibile o basata su talmente pochi da essere inutilizzabile. Si impone, quindi, la determinazione di una soglia al di sotto della quale la procedura non si applica

I domini di stima, la soglia:

Non vi sono criteri predefiniti per determinare una soglia di numerosità per le stime. Di norma si procede sulla base dell'esperienza e si valutano i risultati di diverse alternative.

Nel nostro caso la scelta migliore appare quella di:

1. Calcolare la soglia in base alla quantità di rifiuto prodotta (non in base, ad es. al numero di imprese - potremmo infatti avere una sola impresa che produce quasi tutto il rifiuto)
2. Fissare convenzionalmente la soglia al 5%, escludendo dalla stima i domini che non raggiungono il 5% di produzione del rifiuto, per ogni corrispondente sovra-dominio in cui sono inseriti
3. Ad esempio nel dominio tipo di rifiuto-settore-circoscrizione verranno esclusi tutti quei settori che non raggiungono il 5% della produzione di un dato rifiuto in una data circoscrizione

I domini di stima, il vincolo di coerenza:

Ovviamente, fissando una soglia minima il contributo alla produzione totale di alcuni domini, pur presenti con un peso modesto nella produzione del rifiuto, viene azzerata in termini di previsione

Pertanto il vincolo costituito dalla stima nel dominio di livello superiore non verrebbe rispettato.

Per ovviare a questo inconveniente, per qualsiasi livello superiore, viene creato un dominio residuale che garantisce il rispetto del vincolo.

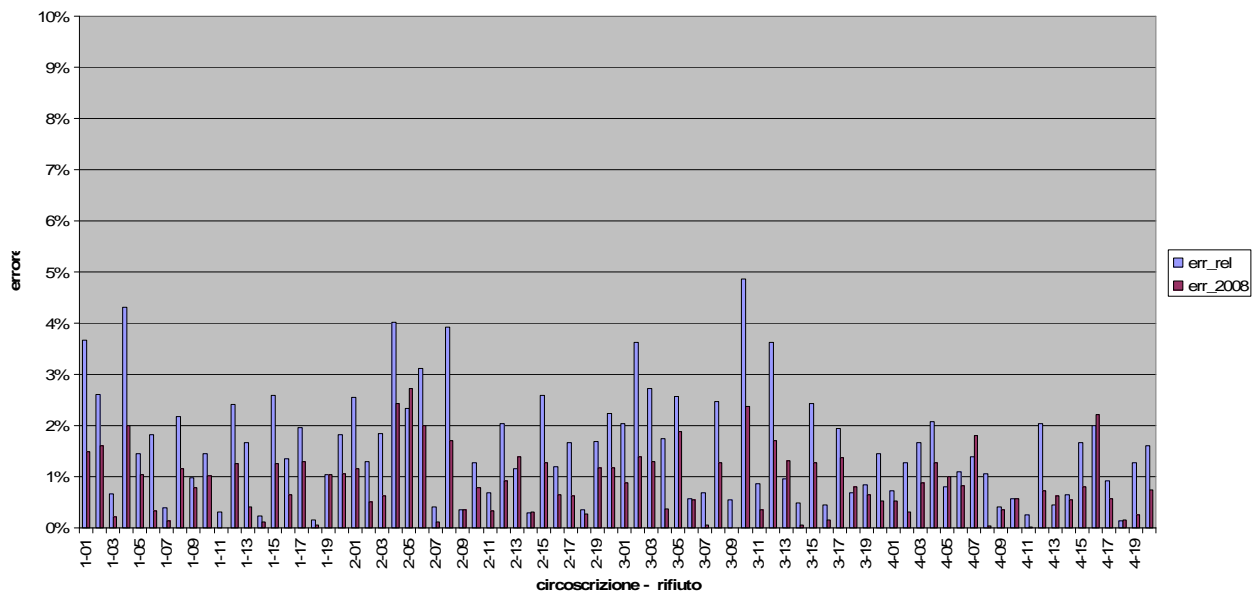
Un esempio serve a chiarire meglio il meccanismo: supponiamo che nel dominio "tipo di rifiuto" per il rifiuto 1 si sia stimata una produzione totale di 100 e che, sommando le stime di tutti i settori che superano la soglia del 5% per il rifiuto 1, si ottenga un totale di 98. La differenza di 2 unità sarà allora attribuita ad un settore creato "ad-hoc" denominato "altri settori"

Lo stesso processo verrà applicato a tutti i domini

Dominio Circoscrizione-rifiuto: i risultati del modello

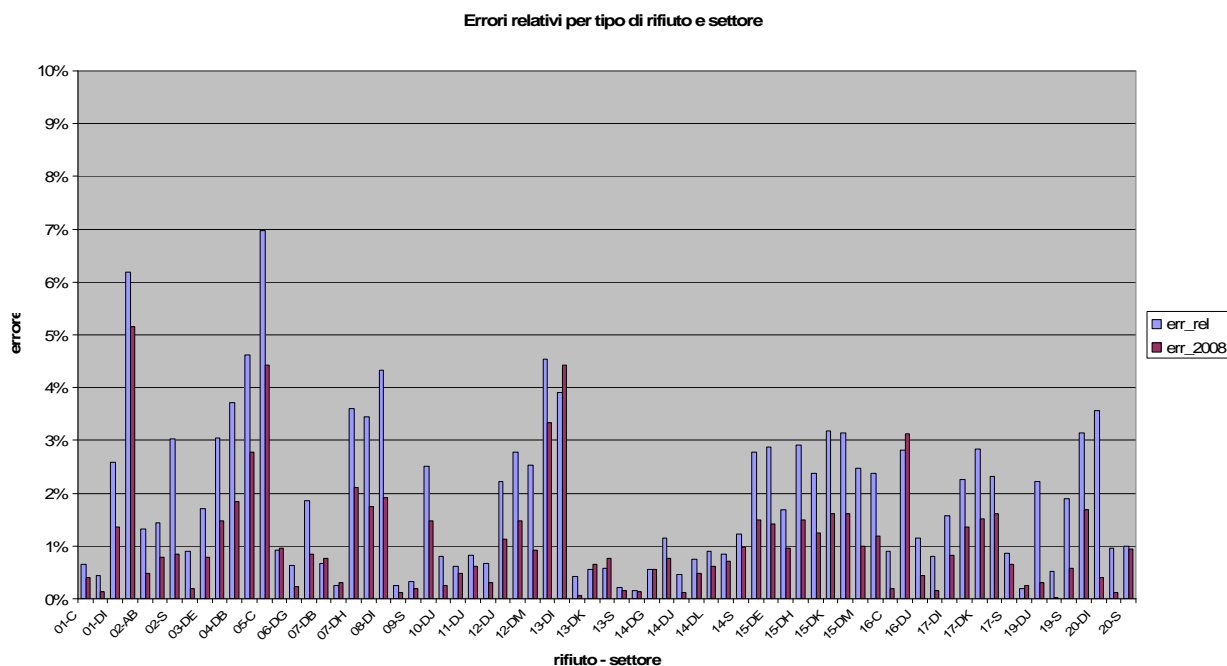
Ovviamente errori lievemente più alti, legati alla numerosità minore, tuttavia nessuna previsione si discosta dal dato "vero" per più del 5% e gli errori relativi al 2008 non arrivano mai al 3%

Errori di previsione per circoscrizione e tipo di rifiuto



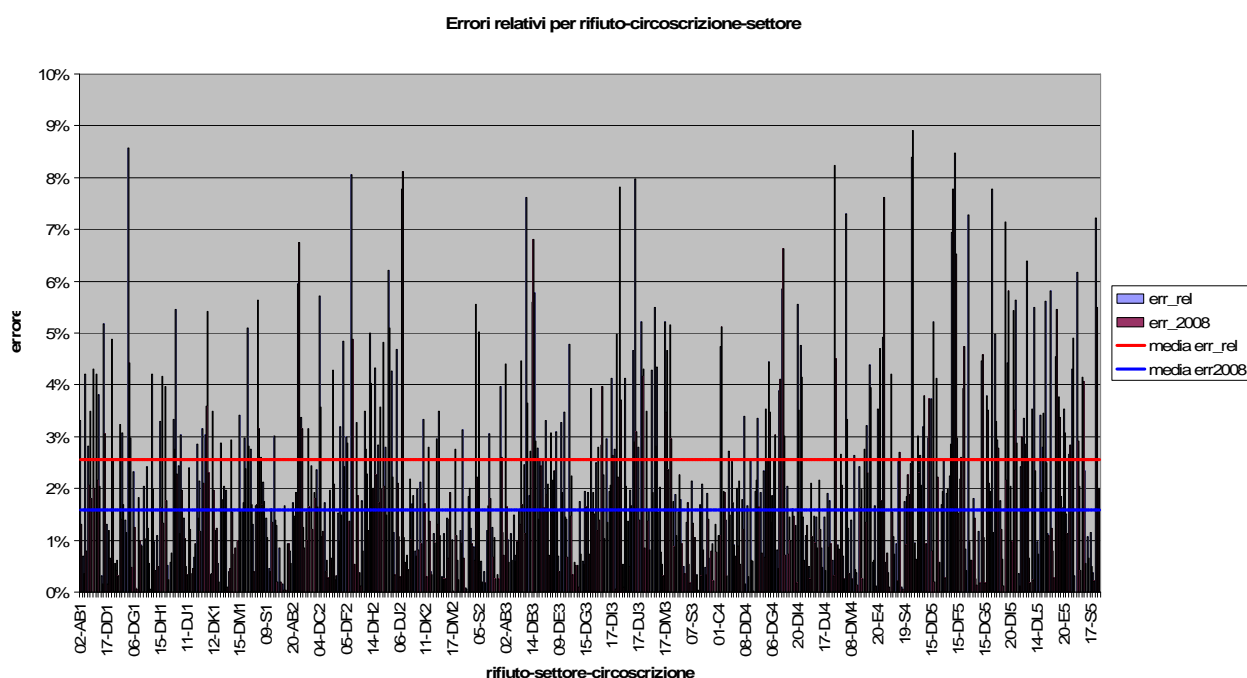
Dominio settore-rifiuto: i risultati del modello

Ovviamente errori lievemente più alti, legati alla numerosità minore, tuttavia nessuna previsione si discosta dal dato "vero" 2008 per più del 5%



Dominio circoscrizione- settore-rifiuto: i risultati

Ovviamente errori più alti, legati alla numerosità molto minore, tuttavia nessuna previsione si discosta dal dato "vero" 2008 per più del 10% e la media degli errori sui domini è sotto al 3%



Modelli di previsione per le dichiarazioni telematiche

Obiettivo: cogliere segnali "precoci" di modificazione del trend

I dati: sottoinsieme ridotto di dichiarazioni MUD disponibili con molta tempestività

Limiti: le Unità Locali che compilano la dichiarazione telematica hanno caratteristiche piuttosto diverse dalle altre

NON PUO' ESSERE TRATTATO COME UN CAMPIONE CASUALE

Tecnica di stima: "stima di modello", cioè stima che tenga conto della distorsione del sottoinsieme e della sua (parziale) rappresentatività.

Non si entra nel dettaglio dei fondamenti teorici della metodologia, si rimanda, tra gli altri a :

Drudi, Filippucci (2000) *Inferenza da campioni longitudinali affetti da selezione non casuale*, Franco Angeli.

Drudi, Ferrante (2002), *"Stima da fonti amministrative longitudinali con parziale sovrapposizione delle unità"*, Franco Angeli

Modelli di previsione per le dichiarazioni telematiche

La tecnica di stima è piuttosto complessa, tuttavia le idee di base si possono illustrare anche in forma intuitiva, si basano su alcune ipotesi-chiave:

1. Si suppone una RELAZIONE (anche complessa) tra il campione e la popolazione, e quindi della sua distorsione, sia descrivibile e quantificabile
2. Intuitivamente, se è vera l'ipotesi 1., allora è possibile individuare una forma di "correzione" per la distorsione.
3. Come per i modelli di panel, si ipotizza che l'analisi dell'evoluzione temporale consenta di studiare e quantificare tale relazione
4. La quantificazione della distorsione (e quindi la guida per la "correzione") è costituita dalla corretta determinazione della variabilità (anno per anno) dei dati e dalla loro correlazione (negli anni)
5. Si dimostra che sotto queste condizioni, è possibile procedere a stime dei parametri per l'intera popolazione, a patto di sfruttare tutta l'informazione disponibile

Modelli di previsione per le dichiarazioni telematiche

Si tratta, quindi, di specificare un modello relazionale tra variabilità-covariabilità del campione e quello della popolazione

A seconda della forma del modello che si ipotizza, si ricavano diversi stimatori, cioè diversi algoritmi di stima, taluni abbastanza complicati

La scelta del "miglior" algoritmo è questione complessa, tuttavia empiricamente viene risolta sulla base della capacità di prevedere ciò che "già successo", in analogia con quanto descritto per la previsione dai dati bonificati

Sono stati sperimentati diversi stimatori, ma quelli che hanno dato i migliori risultati sono:

Stimatore per rapporto

Stimatore GREG, nella versione Fuller-Battese

Modelli di previsione per le dichiarazioni telematiche

Stimatore per rapporto

Ipotizza una relazione lineare che viene stimata per ogni coppia di anni, in sostanza si determina la variazione nel campione e la si considera una buona stima della variazione della popolazione

Stimatore GREG, nella versione Fuller-Battese

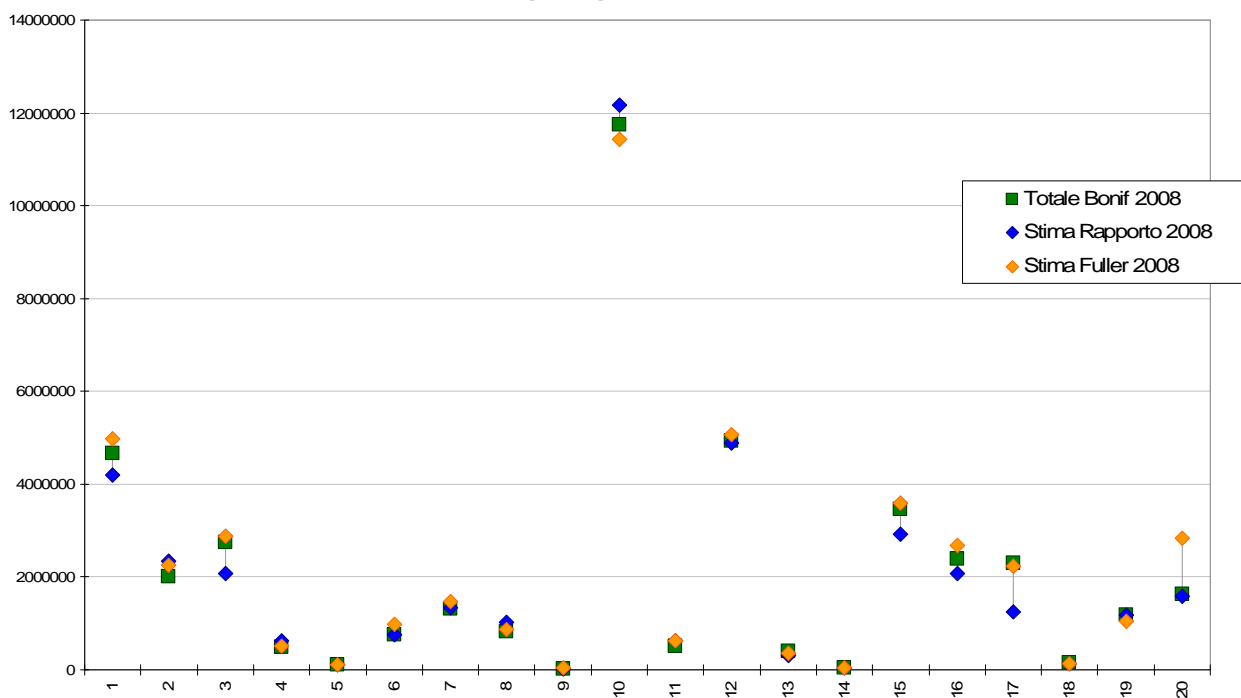
Ipotizza una legge di evoluzione temporale del rapporto campione-popolazione e sulla base di tale legge, quindi dell'intero periodo osservato, determina il valore più corretto per la variazione della popolazione

Il primo risulta (come vedremo) più adatto per quelle situazioni in cui si osservano cambiamenti consistenti e rotture nell'evoluzione temporale, il secondo, utilizzando una quota maggiore di informazione, ha performances migliori quando il trend nella relazione campione-popolazione, è stabile nel tempo.

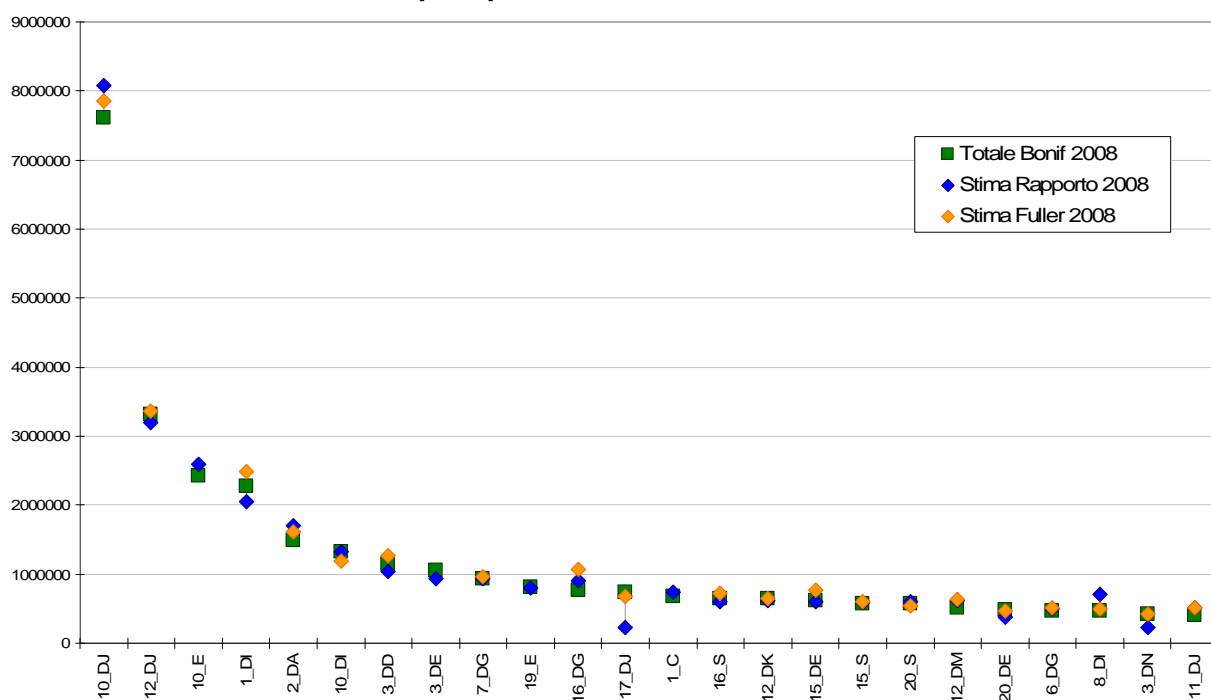
Di volta in volta, si sceglierà l'algoritmo che meglio approssima i dati noti

I risultati:

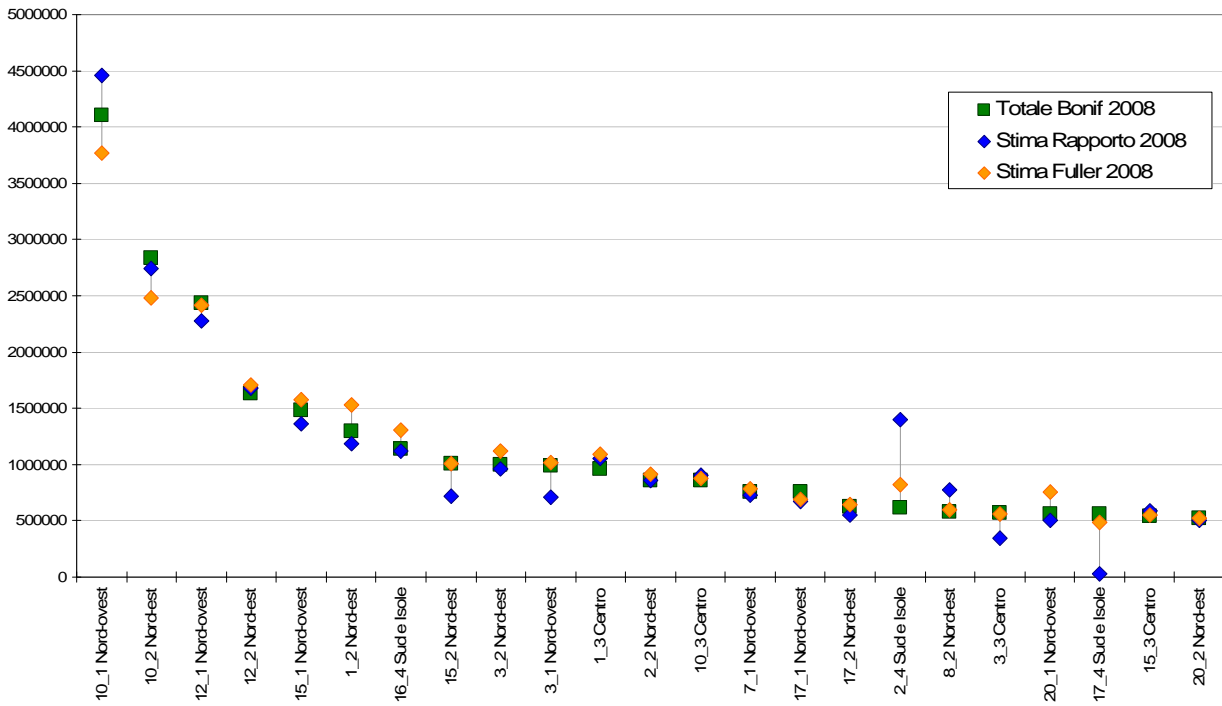
Confronto Totale Bonificato 2008 - Stime per Tipo di rifiuto



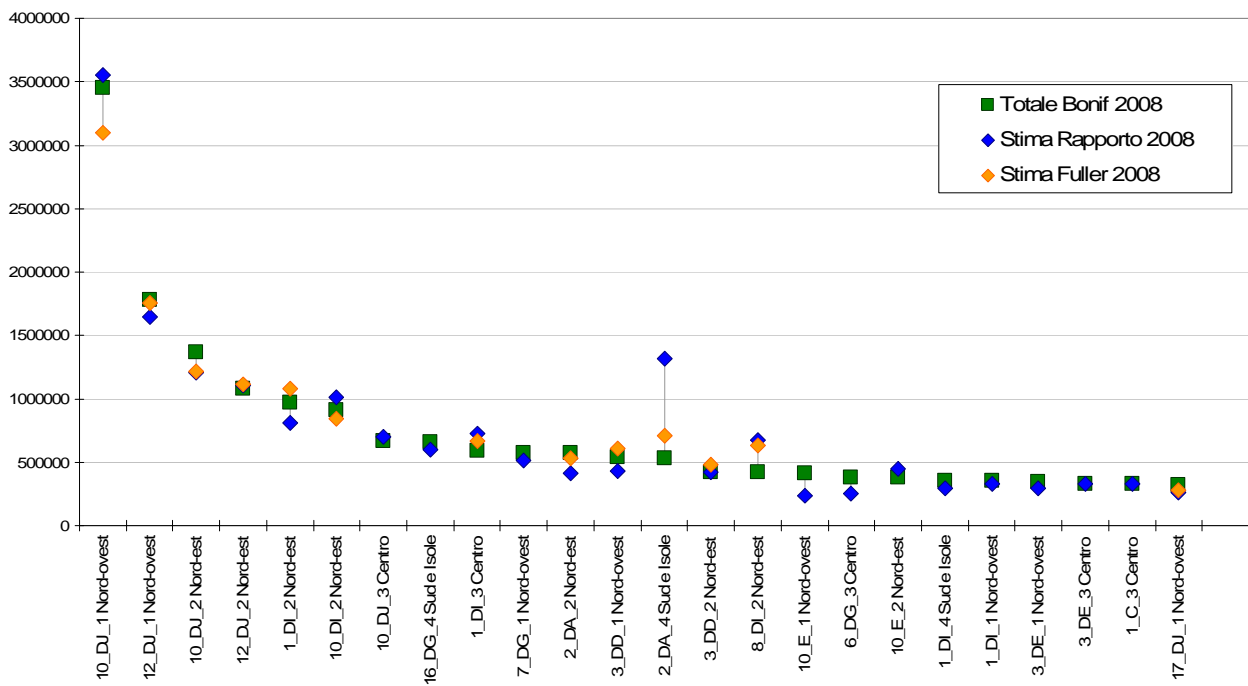
Confronto Totale Bonificato 2008 - Stime per Tipo di rifiuto e Settore di attività



**Confronto Totale Bonificato 2008 - Stime
per Tipo di rifiuto e Ripartizione geografica**



**Confronto Totale Bonificato 2008 - Stime
per Tipo di rifiuto-Settore di attività-Ripartizione geografica**



La combinazione delle stime "bonificate" e "telematiche"

Per combinare le stime ottenute a partire dalle 2 fonti occorre definire un sistema di pesi che rappresentino la quota relativa di informazione di ciascuna di esse, cioè l'ammontare del contributo alla stima sintetica finale

Coerentemente con quanto argomentato circa il criterio di definizione della soglia di "trattamento" dei domini di stima (calcolato sulla produzione totale di rifiuto per dominio) il criterio che è stato adottato per definire i pesi fa perno sull'ammontare della produzione.

In questo modo la stima derivante dalle dichiarazioni bonificate avrà un peso (nel dominio) pari a 1, mentre il peso delle telematiche sarà pari alla quota di produzione di rifiuto dichiarato telematicamente per ciascun dominio nell'ultimo anno disponibile. Le stime finali saranno allora una media ponderata delle due stime.

$$\hat{y}_{tot,t} = \frac{\left(\hat{y}_{bon,t} * 1 + \hat{y}_{tel,t} * \frac{Pr_{tel,t-1}}{Pr_{bon,t-1}} \right)}{\left(1 + \frac{Pr_{tel,t-1}}{Pr_{bon,t-1}} \right)}$$

I risultati:

la variazione 2008-2009 della produzione totale di rifiuti

Tipo di rifiuto	Previsione "bonificate"	Stima "telematiche"	Peso Telematiche	Stima sintetica
Non pericoloso	-2%	-10%	0,1191	-3%
Pericoloso	1%	-10%	0,1182	0%
Totale	-2%	-10%	0,1100	-2%

Grazie per l'attenzione !