Basi dati e integrazione informativa: cosa cambia per la ricerca sociale

Giuseppe Garofalo (ISTAT/DICA/Archimede)

LaRIS – LABORATORIO DI RICERCA E INTERVENTO SOCIALE

10° LaRIS day

Statistica e vita quotidiana: leggere la precarietà

Giovedì 10 marzo 2016 , rescia

Istat

Vision for the next decade

Il modello stovepipe, le *canne d'organo*

- Le statistiche nei singoli settori si sono evolute *indipendentemente* le une dalle altre
- *Processi di produzione* più adatti ai prodotti corrispondenti
- *Flessibile*: può essere adattato velocemente a cambiamenti «minori» nei fenomeni descritti dai dati
- A *basso rischio*: eventuali problemi in uno dei processi di produzione normalmente non hanno impatto sul resto della produzione
- Vantaggio di poter essere disciplinato da un regolamento relativamente *limitato e specifico*

VISION

Svantaggi delle «canne d'organo»

- *Onere sui rispondenti* (cioè sulle imprese, ma anche crollo di tassi di risposta … e dei budget per le rilevazioni)
- Non è adatto alla raccolta di dati sui *fenomeni multidimensionali*, «*quali la globalizzazione o il cambiamento climatico*»
- E' *inefficiente e costoso*: non utilizza la standardizzazione tra settori e la cooperazione
- *Duplicazioni* inevitabili nello sviluppo, nella produzione o nei processi di diffusione

Trascura la connessione fra fenomeni

- *Domanda vs. offerta*
- *Imprese vs. famiglie*
- *Economico vs. sociale (anche dentro l'Istat!)*

Eccessiva specializzazione delle risorse

VISION: da stovepipe a integrazione

- Le statistiche per settori specifici *come parti integrate* in sistemi di produzione completi per gruppi di statistiche
- Basati su una *comune infrastruttura*, con l'utilizzo di tutte le fonti disponibili con un livello di qualità adeguato

- **Combinare i dati di indagini con i dati amministrativi**
- Gestione dei rischi metodologici relativi a
  - concetti e definizioni e
  - rapporti con i proprietari dei dati riutilizzati
- Valutazione della qualità
  - Misura e stima dell'errore statistico

Integrazione

- C'è di nuovo che riguarda i processi di produzione
- Che si inserisce in una fase di evoluzione culturale nell'uso delle statistiche
  - Europa 2020, Beyond GDP & Stiglitz report negli stessi anni
- Che risponde a una domanda crescente (?), per la gestione delle policy anzitutto (il territorio!)
- Che riflette una certa tensione metodologica
  - A cui l'Istat ha partecipato (ESSnet data integration)
  - Ma che non ha ancora messo bene i piedi nel piatto, non si è trasformato in cultura (o scontro)
- E' un processo evolutivo e non reversibile
- Ma può essere un processo «pericoloso»

Il processo di modernizzazione dell'Istat

From multiple modes for surveys to multiple data sources for estimates
by *Constance F. Citro – Statistics Canada*

Register base statistics: Administrative data for statistical purposes
by Andres and Britt Wallgren– Statistics Sweden

Statistics 4.0 - **Are we at the edge of a new era for statistics?**
by Walter Radermacher – Eurostat

Towards an integrated statistics programe for the post-2015 development agenda
by Geet Bruinooge – Statistics Denmark

Towards a system of official statistics based on a coherent combination of data sources, including surveys and administrative data -
by Bo Sundgren, Stockholm University, 2011

Il processo di modernizzazione dell'Istat

Riduzione dei costi:

- diminuzione della raccolta "diretta" dei dati
- eliminazione delle ridondanze nei processi

contraddizione

Incremento dell'offerta *informativa* in termini:

- quantitativi
- qualitativi – quadri informativi più ampli capaci di rispondere alle domande con un approccio multidimensionale

Specialized corporate-level services units : Approccio "per funzioni centralizzate" a supporto di tutti i processi statistici e abbandono dei processi a "***silos***" (per singoli domini stat.)

Sfruttamento di tutte le informazioni disponibili per produrre statistiche "pubbliche" .

Uso ***massivo*** di dati ***non*** raccolti da indagine.

La conta delle popolazioni

Storicamente uso dell'indagine censuaria (pop. Residente/pendolarismo):

- Costi
- Ritardi nella diffusione
- Comportamenti «anagrafici distorti» dei rispondenti
- Impossibilità di analisi longitudinali

Integrazione più fonti amministrative:

- Informazioni «amministrative distorte»
- Parzialità dell'informazione disponibile
- Difficoltà a stimare la «frequenza» dell'uso di un territorio

Call Data Record:

- Incertezza della popolazione di riferimento
- Impossibilità nella «qualificazione»
- Impossibilità di stimare le incoerenze fra chi «possiede» e chi «usa» il cellulare

Uso integrato di fonti amministrative per la «conta»

Informazioni disponibili integrate:

- Registri anagrafici trattati nel sistema ANVIS
- Permessi di soggiorno, fonti sull'occupazione - dipendenti, parasubordinati, lavoro autonomo -, studenti – scuola primaria, secondaria e università – dichiarazioni dei redditi, pensionati, trattamenti per disoccupazione, mobilità,….

Sovra/sotto copertura dei registri anagrafici

Groppo A- Individui in ANVIS CON SEGNALI in altre fonti amministrative	**50.265.493**
Groppo B – Individui in ANVIS SENZA ALTRI SEGNALI	**10.477.172**
Groppo C - Individui NON in ANVIS CON SEGNALI in altre fonti amministrative	**1.889.994**

Gruppo B – Sottopolazioni critiche per la sovracopertura anagrafica

Of which: Children and Persons dependent in the Tax Register	*7.287.274*
Other persons	*2.544.061*
IN ANVIS WITH PERMIT TO STAY	
Of which: Children and Persons «dependent» in the Tax Register	*237.100*
Other persons	*408.737*

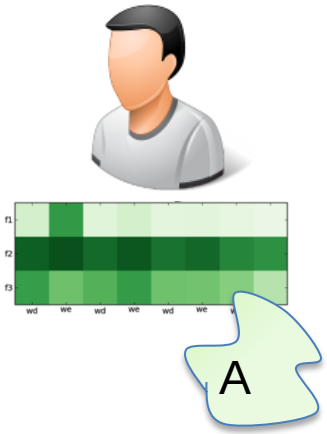
Gruppo C – Sottopolazioni critiche per la sottocopertura anagrafica

WITH SIGNALS OF WORK/STUDY BUT NOT IN ANVIS	**1.401.326**

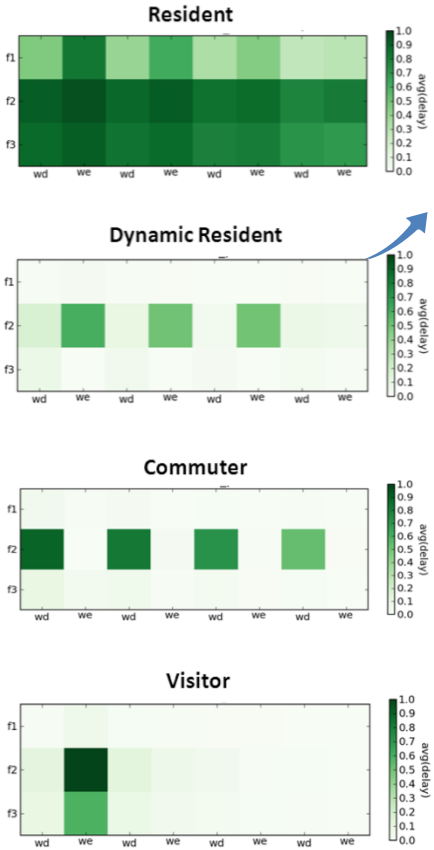
→ *WITH STRONG SIGNALS OF WORK/STUDY* ***384.637***

WITH LIGHT SIGNALS OF WORK/STUDY ***1 mln***

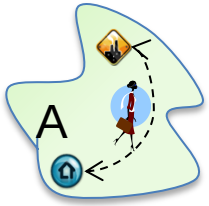
Profilo di chiamata individuale



Algoritmo di Classificazione

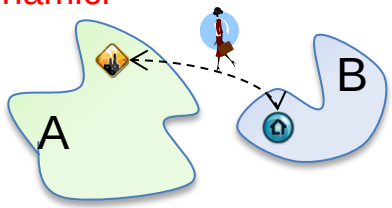


Residenti

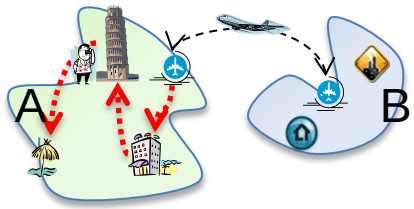


Residenti Dinamici

Pendolari

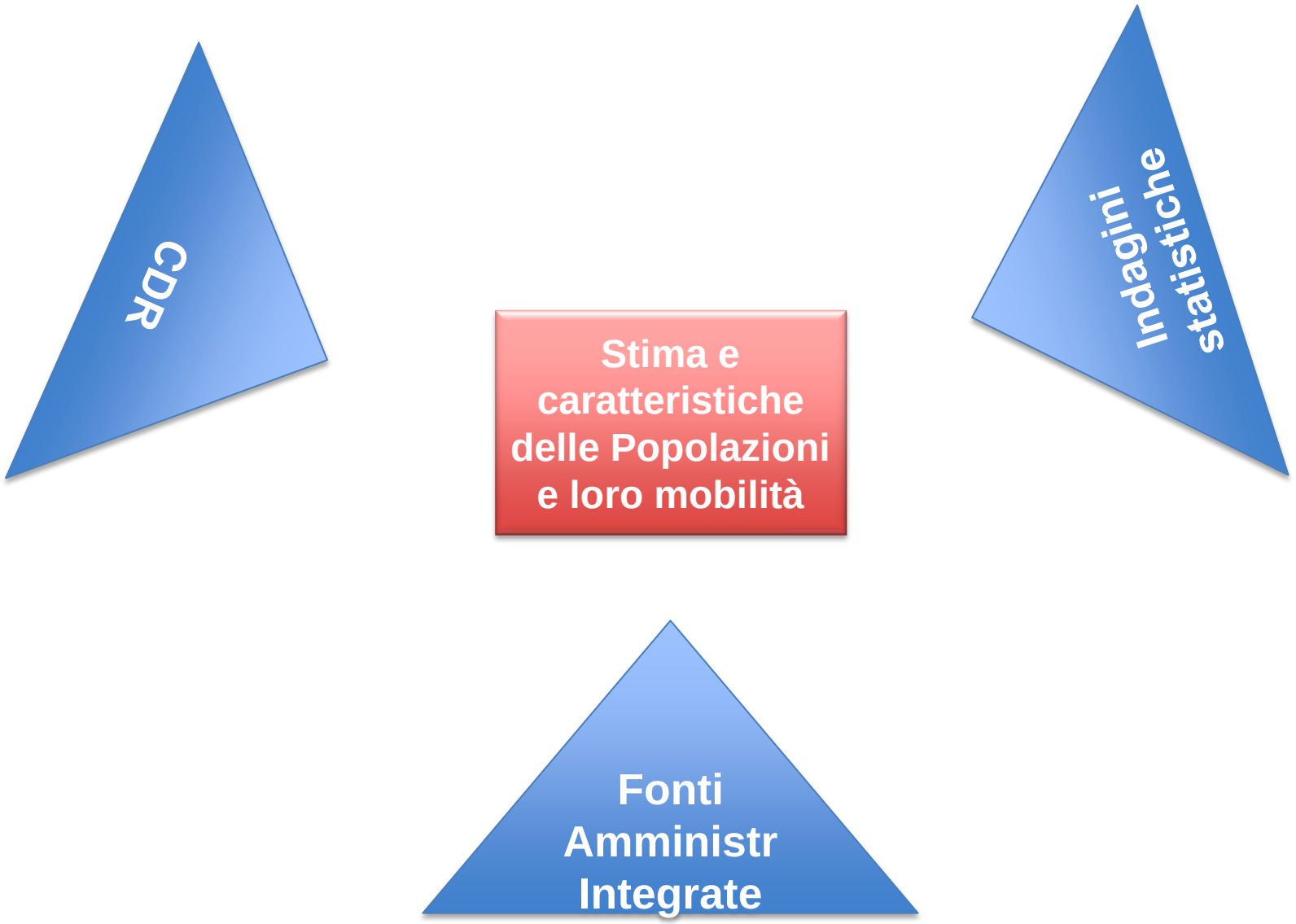


Visitatori



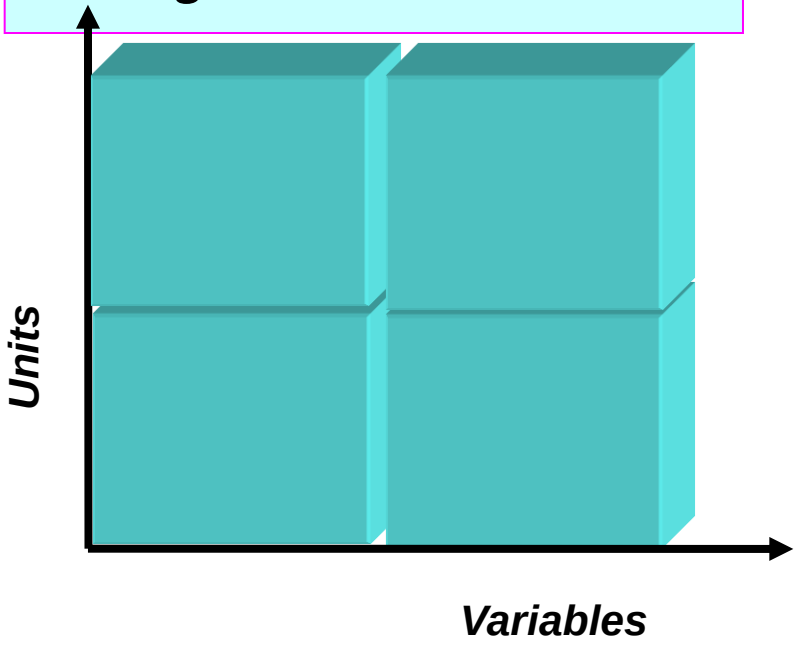
Istat

Multiple Integrated Data Collection

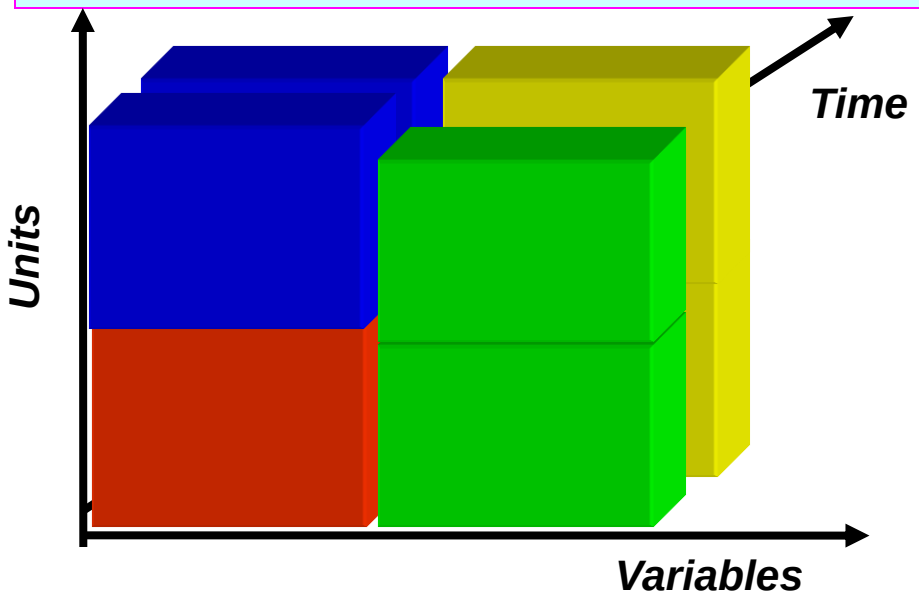


Multiple Integrated Data Collection

Single Data Collection



Multiple Integrated Data Collection



Un bisogno informativo

↓

Una indagine

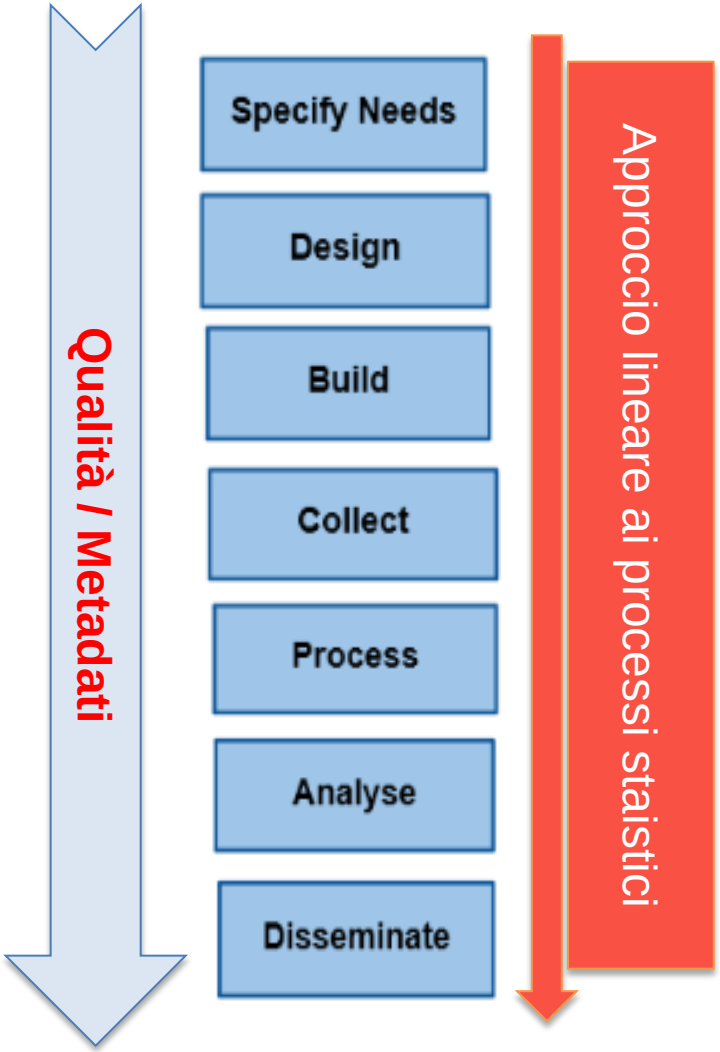
Un bisogno informativo

↓

Più fonti integrate

1:1 → 1:n

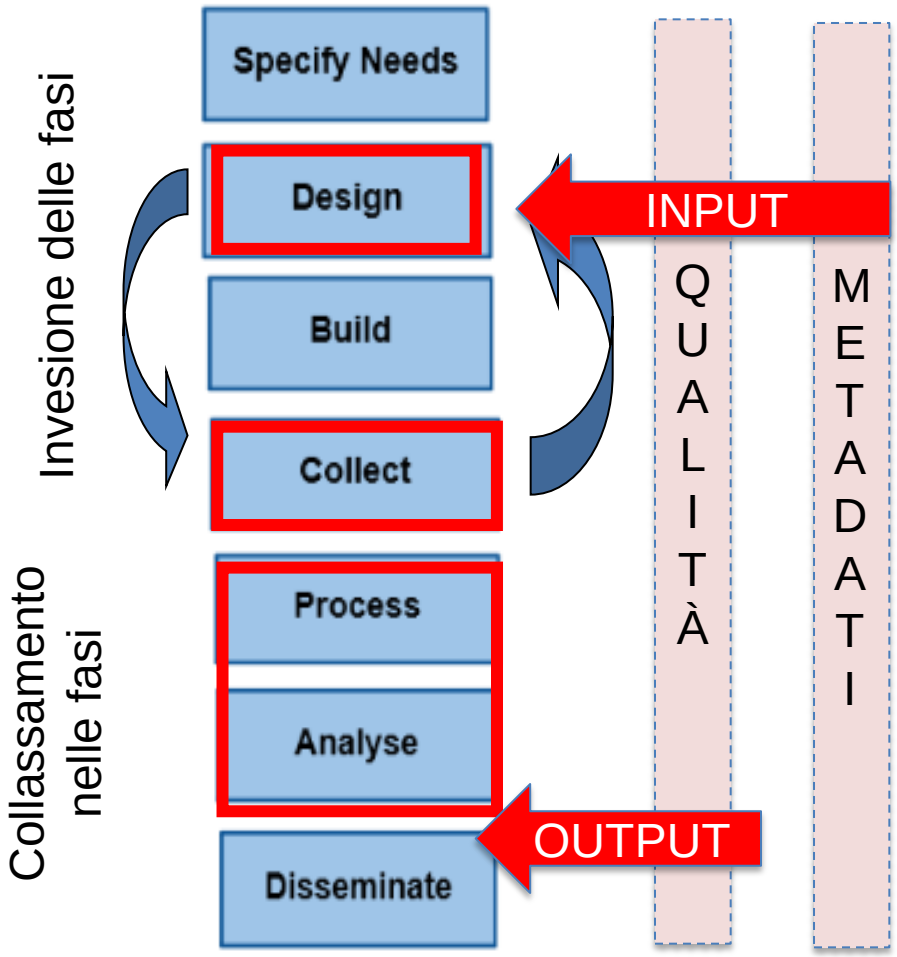
Cosa cambia nel processo produttivo statistico: Generic Statistical Business Process Model (GSBPM)



Nei processi classici di produzione statistica la conoscenza è definita ***a priori***.

- Sappiamo «prima» quale informazione produrre, con quali caratteristiche e come produrre l'informazione
- La modifica del dato è governata.

Cosa cambia nel processo produttivo statistico: Generic Statistical Business Process Model (GSBPM)



Con i nuovi processi ***il «dato» esiste già***:

- Senza (o con poche) informazioni sul processo di generazione
- A volte con una scarsa o nulla valutazione della qualità

Instabilità del dato (cambia per esigenze esterne) non governata da chi lo usa per finalità statistiche

Multiple Integrated Data Collection

- Grandi moli di informazioni
- Difficoltà nell'integrazione fisica
- Le diverse fonti possono non essere disponibili in tempi diversi
- Possono utilizzare concetti/classificazioni non coerenti fra loro
- Possono utilizzare gli stessi concetti ma con visioni differenti (oggettivo/soggettivo)
- Contengono differenti tipologie di errori (non campionari/campionari)
- Possono contenere differenti livelli di qualità intrinseca

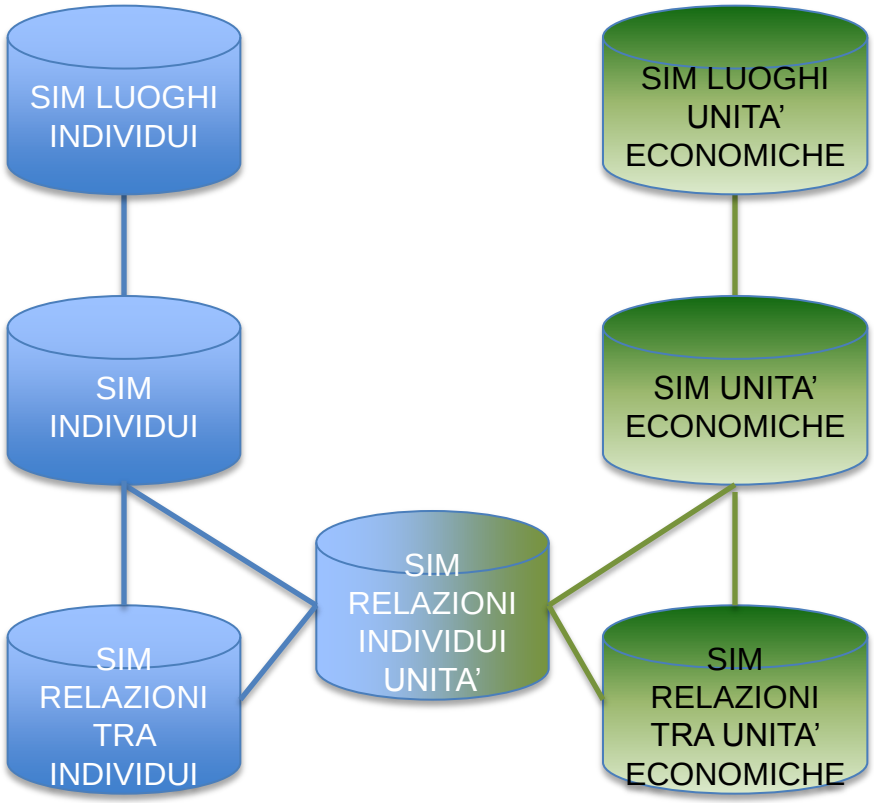
Processo produttivo complesso

Modifica nei processi produttivi, nell'organizzazione, nelle tecnologie e nelle metodologie statistiche ma anche nella «lettura» dei fenomeni.

SISTEMA INTEGRATO DEI MICRODATI (SIM)

Repository dei dati amministrativi acquisiti dall'Istituto,
organizzato con lo scopo di supportare i processi di produzione statistica dell'Istat.

Favorisce l'utilizzo di dati individuali, ***privi degli identificativi diretti***,
mantenendo inalterate le potenzialità informative derivanti dal processo di integrazione



Sistema Integrato di Microdati (SIM)

Tipologia delle fonti		Sottosistemi	
		Individui	Unità
Anagrafici	Anagrafi Comunali	X	
	Anagrafi Consolari	X	
	Anagrafe Tributaria	X	X
	Permessi di soggiorno	X	
Fiscali	Banca Dati Reddituale - MEF	X	
	Studi di settore		X
	Modello UNICO	X	X
	Modello 730	X	
	Modello 770	X	X
Formazione	Anagrafe degli studenti	X	
	Anagrafe degli studenti universitari	X	
	Anagrafe personale doc. e non doc. delle scuole	X	X
	Anagrafe personale doc. e non doc. delle università	X	X
Lavoro	Arch. INPS Emens (UNIMENS)/Parasubordinati/Cassa integrazione/Lav. agricoltura/Artigiani e commercianti/ Autonomi dell'agr./Lav Domestici	X	X
	Archivi INAIL	X	X
	Arch. Ex-INPDAP ed Ex-ENPALS	X	X
	Cedolini stipendiali (MEF)	X	X
Welfare	Casellario dei pensionati	X	
	ANF/Maternità	X	
	Mobilità/Disoccupazione/LSU	X	
Camerali	Registro delle Imprese		X
	Soci delle Imprese	X	X
	Persone con cariciche sociali	X	X
	Bilanci delle Imprese		X

Sistema Integrato di Microdati (SIM)

SIM	N. Fonti / Records
Individui	50 (600mln records)
Unità	42 (65mln records)
Luoghi individui	25
Luoghi unità	30
Relazioni individui	3
Relazioni unità	7
Relazioni individui_unità	12

100 mln di codici individui
10 mln di codici unità

Migliaia di variabili

Progetto ARCHivio Integrato di Microdati Economici e DEmografici (ARCHIMEDE)

Obiettivo: ampliamento dell'offerta informativa dell'ISTAT mediante produzione di collezioni di dati elementari di tipo longitudinale e *crossection*, da rendere disponibili all'utenza, utili alla ricerca sociale ed economica, alla programmazione territoriale e settoriale, politiche pubbliche a livello nazionale, regionale e locale.

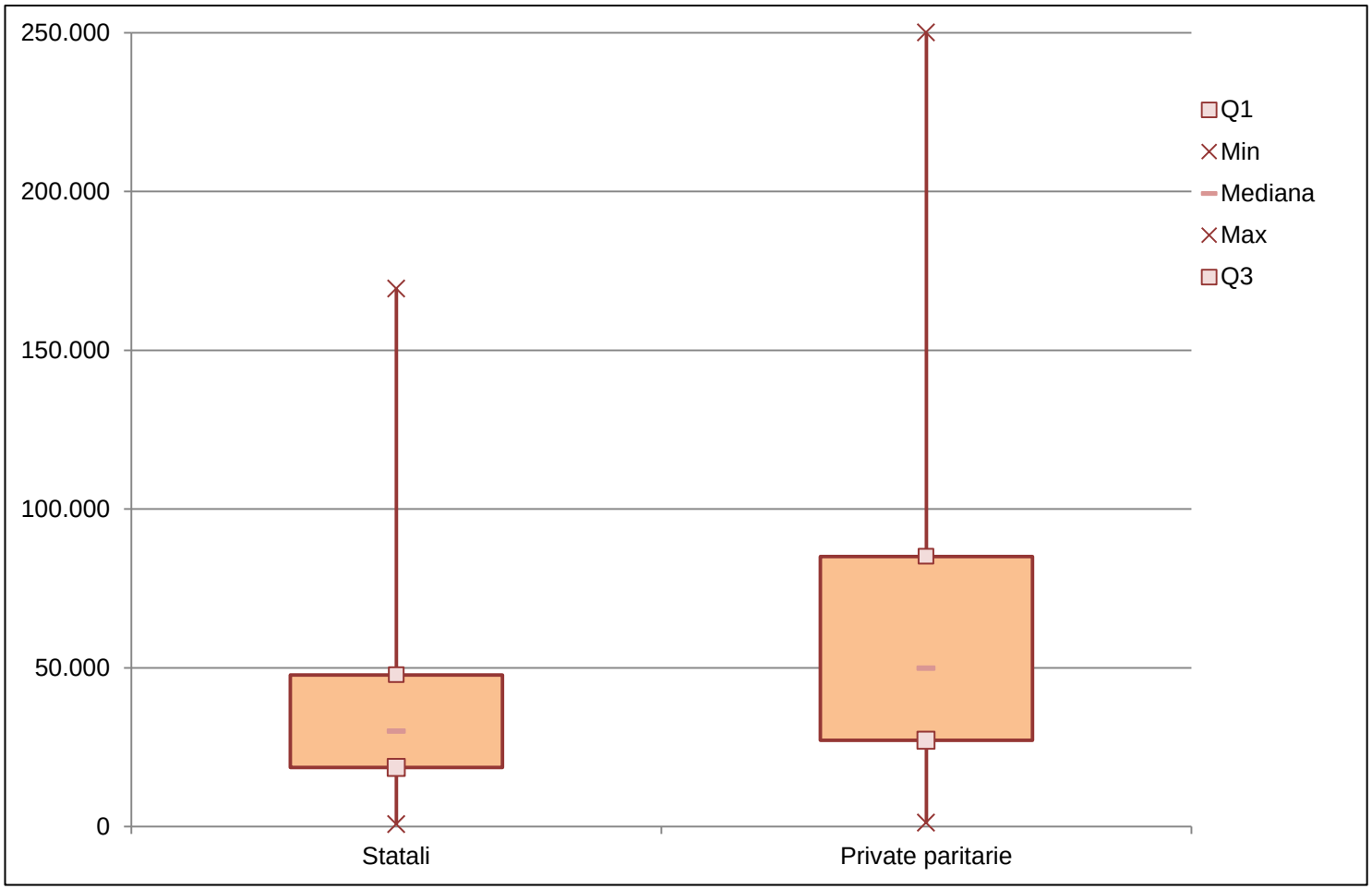
Sfruttamento dei contenuti informativi di fonti amministrative «integrate» presenti in SIM.

Massimizzare lo sfruttamento dei dati disponibili per massimizzare l'informazione statistica resa disponibile

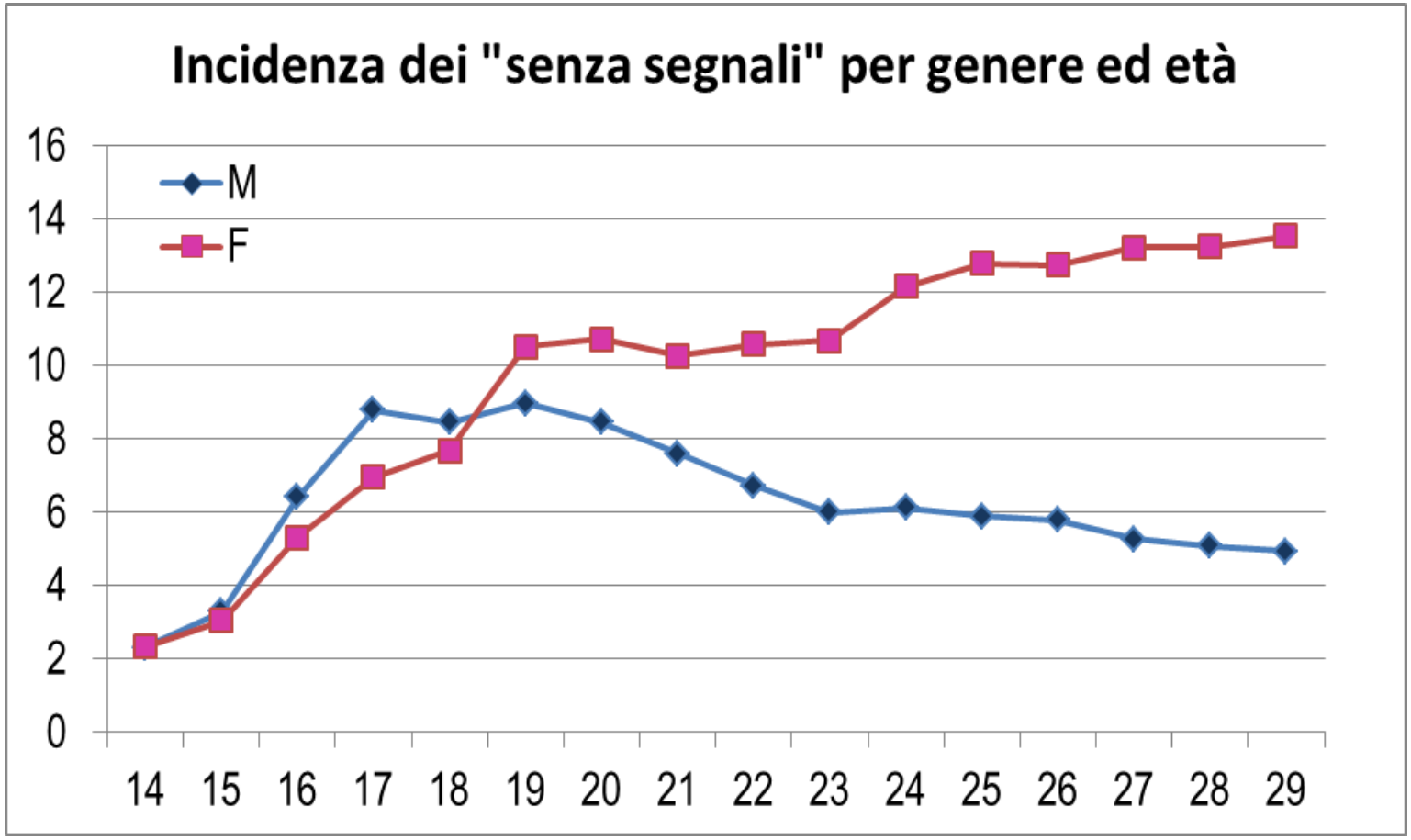
- Cercare di far «parlare» i dati disponibili
- Fare riferimento ad un insieme di unità che possono non rappresentare l'universo di una specifica popolazione
- Utilizzare dati amministrativi anche senza un «ossessivo» trattamento statistico (es. editing/imputation)

Come è nato: massimizzare

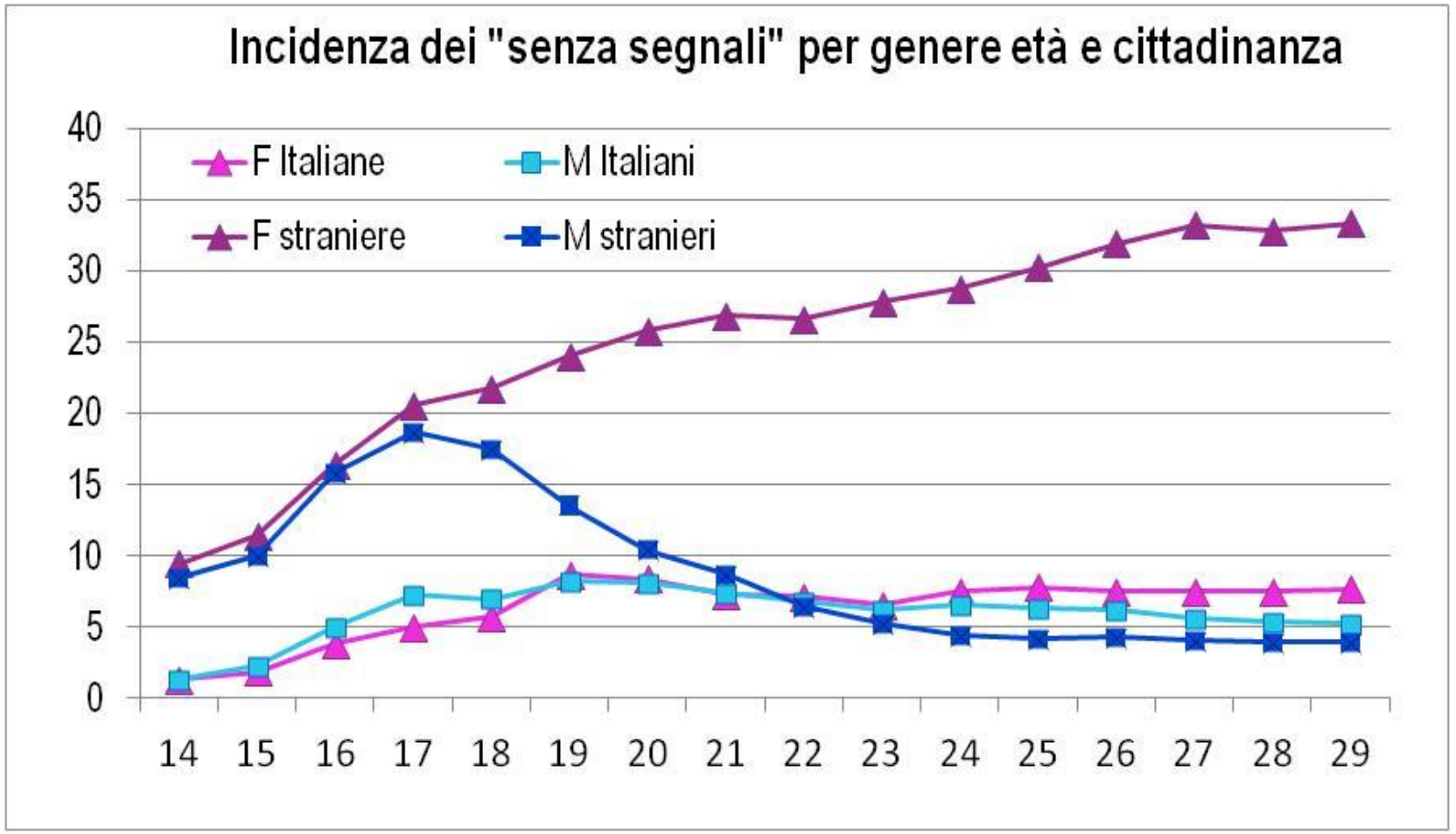
Box-plot del reddito dei genitori degli studenti per tipo di gestione delle scuole
(integrazione di dati fiscali/miur/anagrafici)



Distribuzione giovani 14/29 anni senza segnali di lavoro/studio nella Regione Lombardia
(integrazione di dati Miur/INPS/Fiscali/Anagrafici)

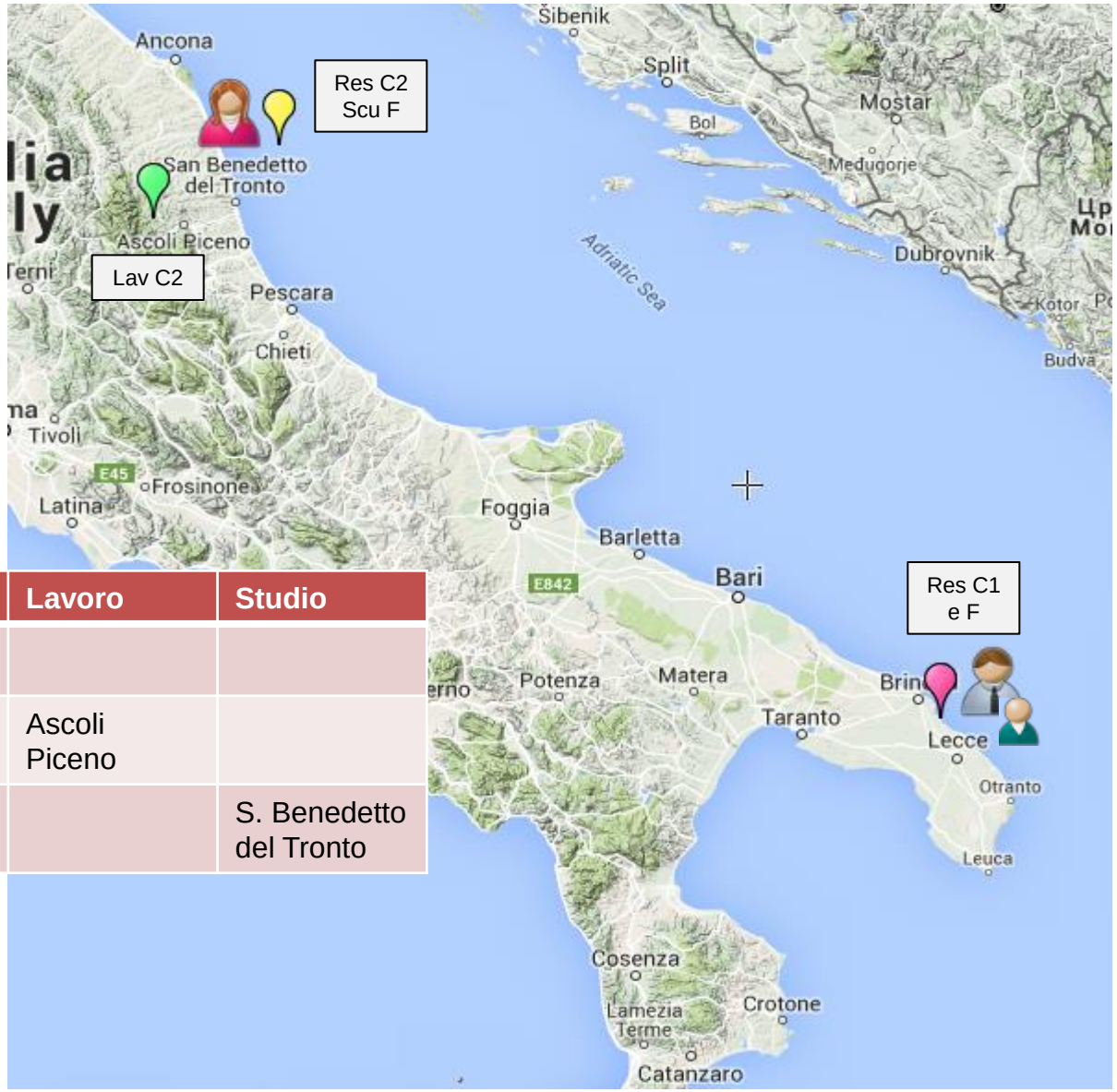


Distribuzione giovani 14/29 anni senza segnali di lavoro/studio nella Regione Lombardia
(integrazione di dati Miur/INPS/Fiscali/Anagrafici)



Come è nato: massimizzare

Identificazione del luogo di dimora abituale (integrazione di dati Miur/INPS/Anagrafici)



	Residenza	Lavoro	Studio
Coniuge 1	Lecce		
Coniuge 2	S. Benedetto del Tronto	Ascoli Piceno	
Figlio	Lecce		S. Benedetto del Tronto

Progetto Archimede - Attività

2013/14 : Costruzione di basi di microdati

1. ***Popolazioni che insistono su di un territorio***
   - Realizzazione Sis. Inf. *Persons&Places*
   - Matrici origine/destinazione per ambiti terr.
   - Identificazione delle tipologie di *«city users»: Residenti, Temporaneamente dimoranti e Pendolari*
2. ***Precarietà lavorativa*** – Analisi delle caratteristiche di un universo di individui definiti «lavoratori precari» osservandone le trasformazioni nel tempo (analisi longitudinale delle transizioni)
   - Concetti di **atipicità** e **professionalità** (elementi oggettivi)
   - Tipologie contrattuali utilizzate in forma impropria (P.IVA monocommittenti / Tirocini e stages)
3. ***Condizioni Socio-economiche delle famiglie*** - Classifica le famiglie secondo le dimensioni:
   - Tipologia della famiglia / Reddito / Condizione lavorativa / Disagio (disabilità, pensioni al minimo, cittadinanza) /Istruzione

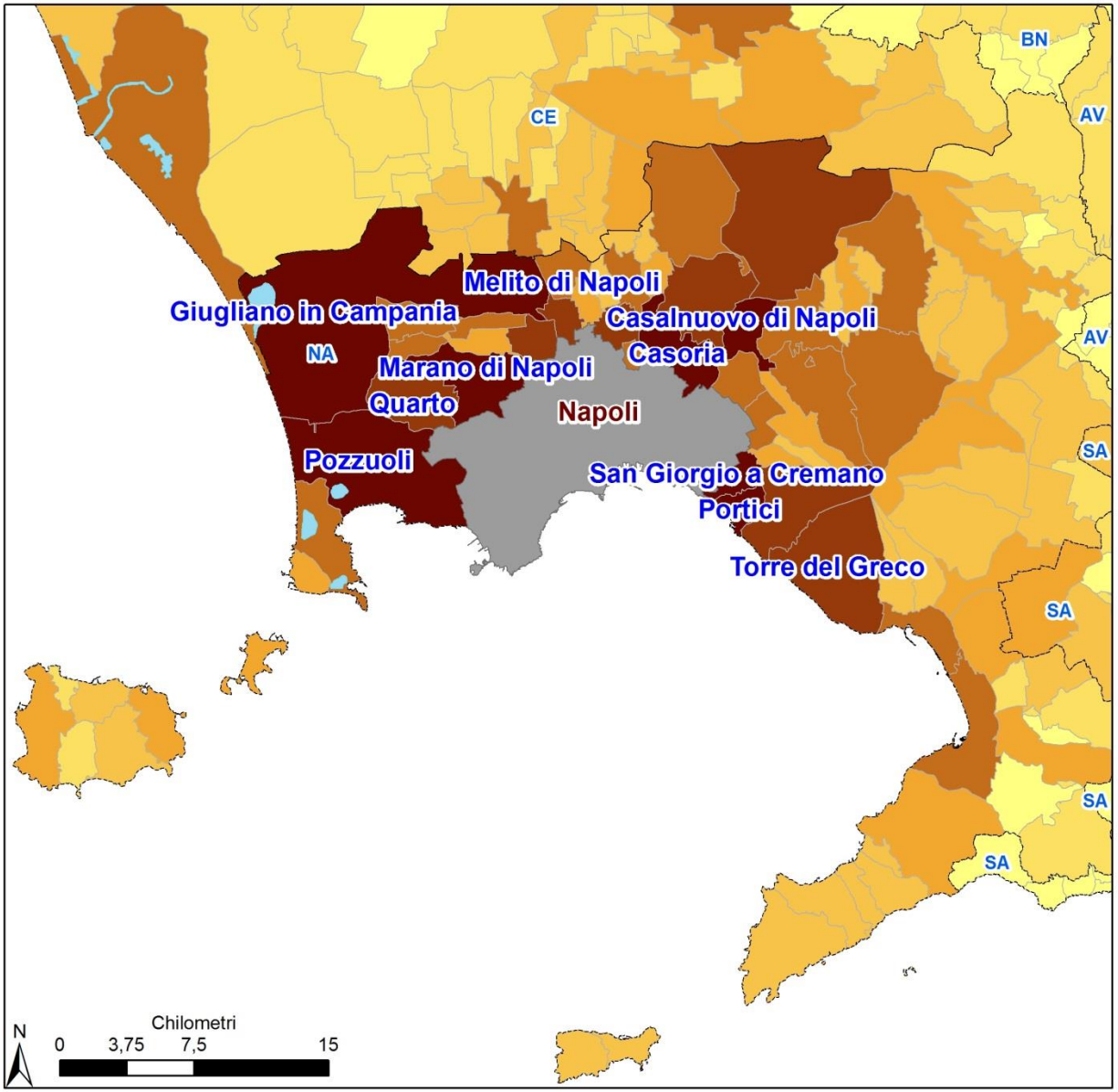
2014/15 : Valutazione

1. ***Definizione del «cruscotto» degli indicatori derivati***
2. ***Definizione di un nuovo progetto sperimentale sui «percorsi di istruzione/formazione/inserimento lavorativo»***
3. ***Valutazione degli output sperimentali in alcuni ambiti territoriali***

2016 : Diffusione interna la SISTAN

Popolazioni che insistono su di un territorio

Lavoratori in entrata nel **Comune di NAPOLI**. Anno 2012



Legenda

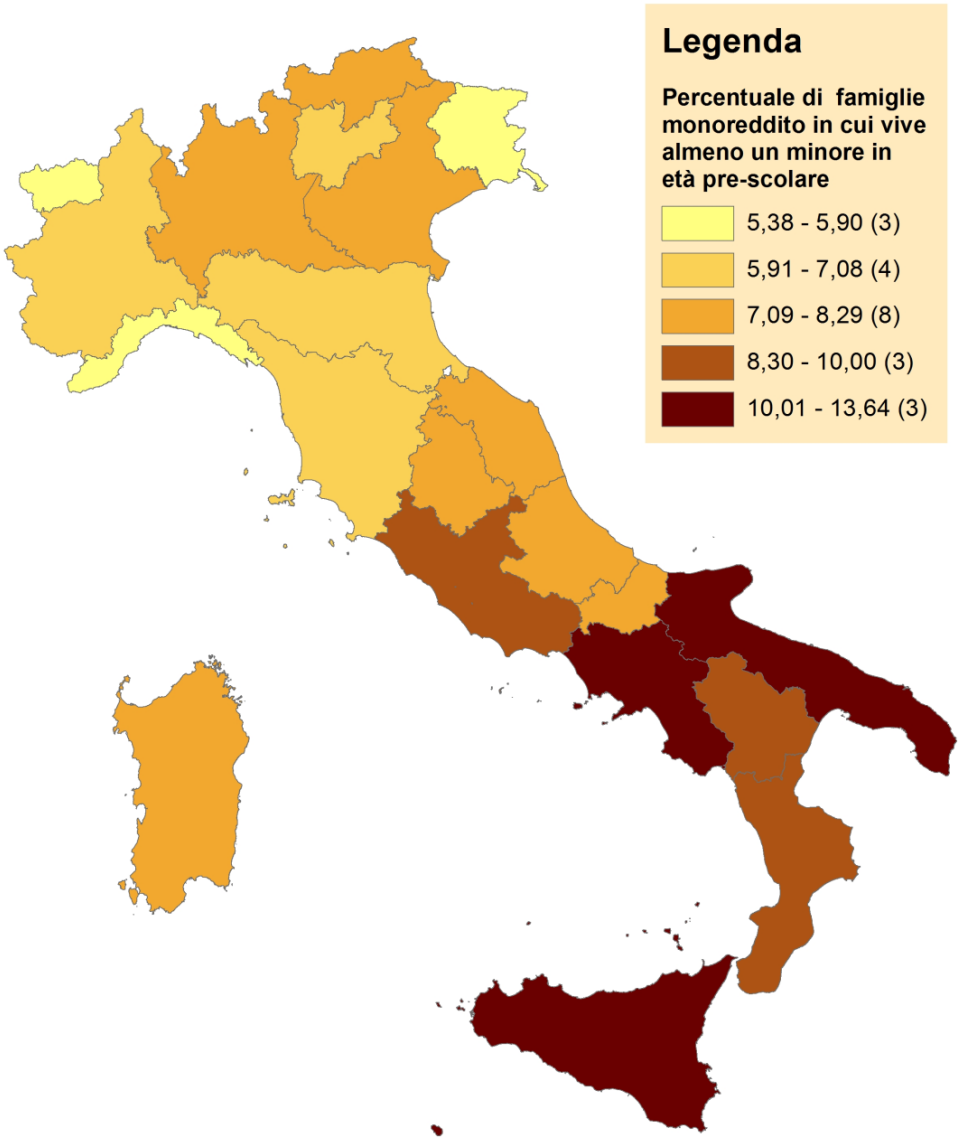
- Confini Provinciali
- Napoli

Occupati in entrata a Napoli

- 1 - 49 (21)
- 50 - 199 (41)
- 200 - 499 (46)
- 500 - 999 (21)
- 1.000 - 1.999 (18)
- 2.000 - 3.499 (8)
- 3.500 - 7.427 (7)

COMUNE ORIGINE *	N
Giugliano in Campania	7.427
Casoria	5.591
Pozzuoli	5.578
Marano di Napoli	5.125
Portici	4.595
San Giorgio a Cremano	4.408
Casalnuovo di Napoli	3.549
Quarto	3.391
Torre del Greco	3.276
Melito di Napoli	3.136
Acerra	2.903
Mugnano di Napoli	2.853
Roma	2.610
Ercolano	2.421
Afragola	2.303
Arzano	2.026
Villaricca	1.983
Pomigliano d'Arco	1.908
Caserta	1.799
Volla	1.702
Salerno	1.612
Cercola	1.530
Castellammare di Stabia	1.484
Aversa	1.479
Casavatore	1.398
Somma Vesuviana	1.344
Bacoli	1.324
Torre Annunziata	1.248
Marigliano	1.228
Sant'Anastasia	1.223
Sant'Antimo	1.144

Condizioni socio-economiche delle famiglie



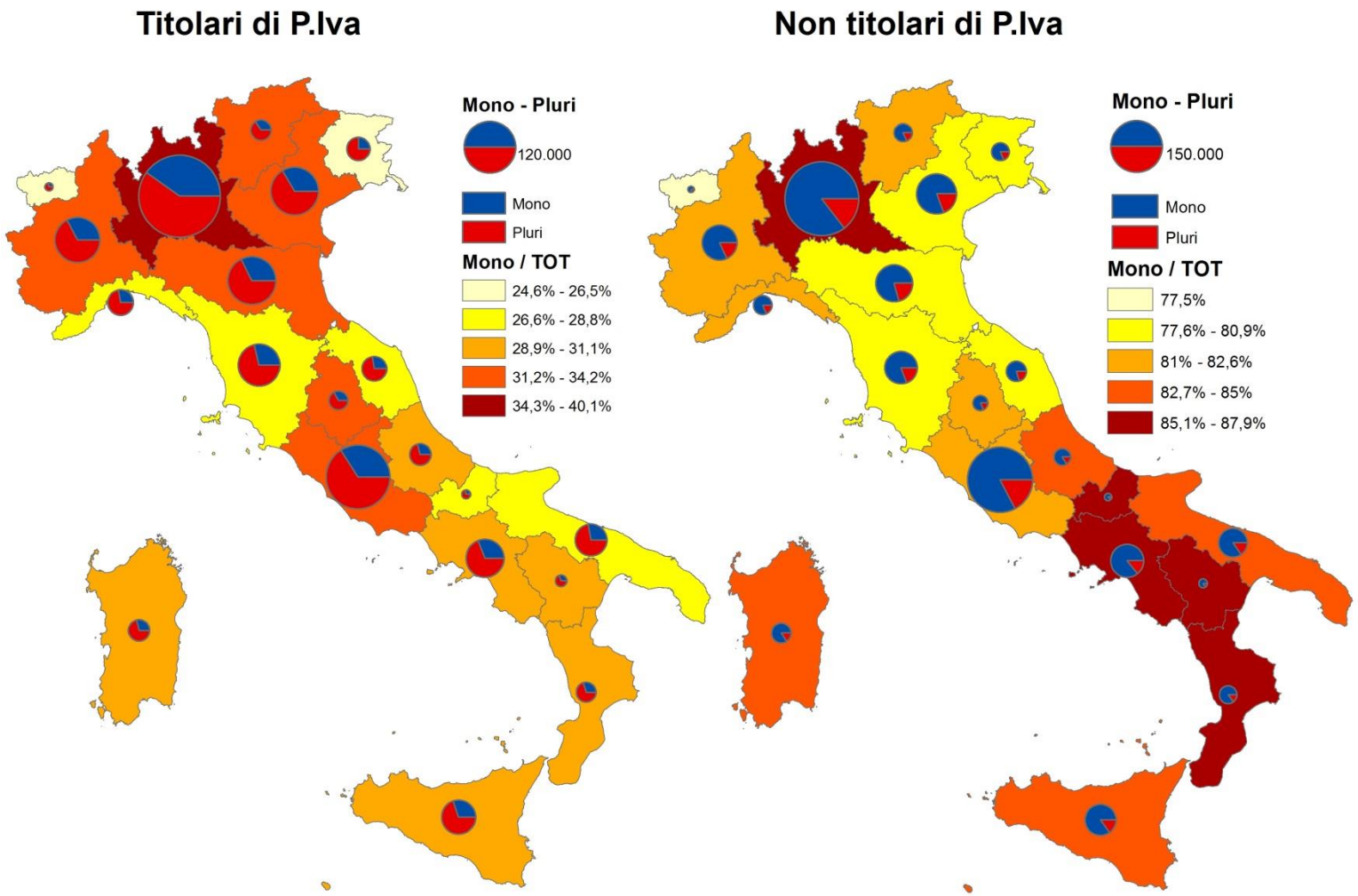
Livello regionale

Percentuale di famiglie monoreddito in cui vive almeno un minore in età prescolare - Anno 2012

Numeratore: numero famiglie monoreddito in cui vive almeno un minore di 6 anni.

Denominatore numero famiglie monoreddito.

Focus su lavoro autonomo: analisi territoriale* 2012



** Fonte: Registri ASIA*

Progetto Archimede – Punti di forza

- ***Verso l'esterno*** dell'Istituto:
  - Diffusione microdati ad un livello di aggregazione territoriale fine
  - Messa a disposizione di Enti Pubblici di basi di microdati per l'analisi di sottopopolazioni di interesse (schematizzazione delle popolazioni a cui sono rivolti interventi e politiche pubbliche)
  - Possibilità di identificare strumenti omogenei – indicatori – fra le varie realtà territoriali a supporto alla comparazione territoriale

- ***Verso l'interno*** dell'Istituto:
  - Esplorazione/identificazione di segnali delle fonti amministrative utili ai processi statistici.
  - Ponte fra statistiche sulle imprese e statistiche sugli individui/famiglie
  - Sperimentazione di nuove metodologie di integrazione e di analisi
  - Possibilità di analizzare fenomeni per «popolazioni» diverse

Archimede – Aspetti critici (o non ancora risolti)

Distanza con output statistici «simili» e «ufficiali»

Valutazione della qualità

Tempi di disponibilità delle fonti integrate

- **Dal punto di vista della conoscenza: *Illusione informativa***
- **Dal punto di vista statistico: *Moltiplicazione dei «rumori»***
- **Dal punto di vista della privacy: *De-anonimizzazione***

Principio di economia di William Ockham:

pluralitas non est ponenda sine necessitate ponendi

Grazie per l’attenzione