

Agglomeration and Local Dynamics in Italian Firms

Stefania Cardinaleschi – ISTAT

Andrea D'Orazio – ISTAT

Stefano De Santis – ISTAT

Marina Schenkel – Università di Udine

Francesco Giovanni Truglia – ISTAT

AISRE - XL Conferenza scientifica annuale

L'Aquila, 16-18 Settembre 2019

Introduzione: Obiettivi del lavoro

Individuando il territorio come soggetto tematico:

- Coniugare la discussione teorica e applicata con la crescente disponibilità di dati derivante dai processi di digitalizzazione (Big Data).
- Ampliare la platea degli interlocutori e fruitori a tutti i livelli (statistico, accademico, policy).
- Il progetto è perciò dichiaratamente multidimensionale (nei dati, nelle competenze, negli output) e a “geometria variabile”, ossia modulare per essere svolto nel tempo e nelle sue articolazioni da soggetti diversi per raggiungere i loro obiettivi specifici.

Introduzione: Fonti, Strumenti, Obiettivi

Costruzione di basi dati da Big Data, fungibili rispetto a molteplici scopi.

1. Costruzione della base dati sfruttando principalmente basi di microdati (archivi statistici, archivi amministrativi, indagini campionarie, Big Data).
2. Definizione di aree territoriali autoinclusive (“soggetti territoriali”), ossia aree funzionali coerenti dal punto di vista socio-demografico-economico per descrivere il modello di sviluppo e valutarlo in termini di politica
3. Analisi territoriali a livello socio-demografico ed economico (ad es.: Contesto e dinamica dei sistemi produttivi locali, Mercati locali del lavoro, ecc.)
4. Realizzazione di una serie di contributi tematici sul territorio, possibilmente con cadenza periodica
5. Realizzazione di basi di microdati per la ricerca e le valutazioni delle politiche, ad es. le valutazioni controfattuali richieste dai Rapporti Attuativi dei Fondi Strutturali

Introduzione: Definizione di aree territoriali autoinclusive

Caratterizzazione degli agenti (imprese, individui e famiglie), delle loro relazioni e del posizionamento per costruire partizioni del territorio compatte, attraverso le seguenti fonti di dati:

- Dati amministrativi
- Produzione di dati da banche dati/servizi web
- Big data (georeferenziazione)
- Indagini statistiche (statistical matching / stime per domini non pianificati)

Introduzione: Fonti, Strumenti, Obiettivi

E' possibile un tipo di analisi completamente nuova, in cui lo studio del territorio può basarsi interamente su una analisi a livello di microdato

La metodologia di individuazione delle unità funzionali minime e microfondate del territorio è di tipo data driven (partizioni costruite a partire dal punto impresa/individuo, aggregati con tecniche esplorative di data mining, es.: clustering supervised e unsupervised).

La metodologia è dichiaratamente multidimensionale: diverse partizioni minime ottenute dalle diverse fonti di dati e metodologie applicate sono poi sovrapposte al fine di consentire una lettura multidimensionale dei fenomeni sul territorio.

Una lettura multidimensionale del territorio.

Laddove possibile la metodologia sfrutta analoghe partizioni territoriali ufficiali già realizzate in ISTAT, a es. i Sistemi Locali del lavoro (SLL).

Gli SLL rappresentano una griglia territoriale i cui confini, indipendentemente dall'articolazione amministrativa del territorio, sono definiti utilizzando i flussi degli spostamenti giornalieri casa/lavoro (pendolarismo). Rappresentano perciò aree definite da un incontro (realizzato) fra domanda e offerta di lavoro. Si valutaranno congiuntamente offerta (gli spostamenti: i lavoratori) e domanda (i luoghi di lavoro: le imprese) in una regione in particolare.

Una lettura multidimensionale del territorio.

Il piano di lettura dei cluster di imprese, sovrapposto a quelle delle matrici di pendolarismo (qualsivoglia definiti) consentirebbe una lettura integrata e multidimensionale del territorio, con particolare riferimento alla domanda/offerta di lavoro sul territorio.

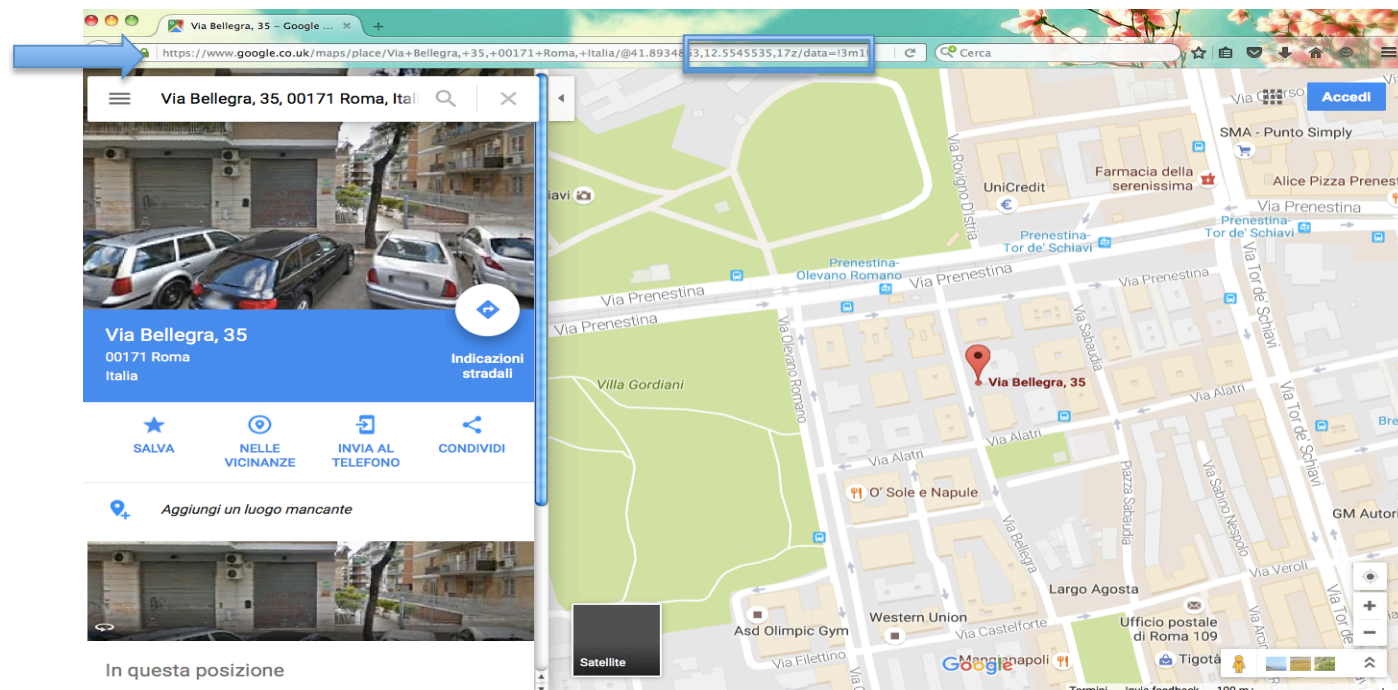
La domanda è definita rappresentando le imprese come punti e pertanto l'approccio seguito è noto in letteratura come *Point Pattern Analysis*. Lo scopo è ricavare una serie di informazioni sulla distribuzione spaziale delle imprese a partire dalla loro localizzazione sul territorio.

La georeferenziazione dei dati rappresenta un arricchimento dell'informazione statistica e consente di utilizzare specifici apparati metodologici e strumenti tecnici per l'analisi geo-spaziale di matrici complesse.

Standardizzazione e referenziazione degli indirizzi

Standardizzazione e suddivisione degli indirizzi nelle loro parti elementari e successiva referenziazione via servizi web

DUG	Toponimo	Civico	Den. Comune	CAP	Sigla Prov.
Via	Bellegra	35	Roma	00171	RM



Geolocalizzazione: referenziazione degli indirizzi

Esempio *response* json

```
{
  "results" : [
    {
      "address_components" : [
        {
          "long_name" : "22",
          "short_name" : "22",
          "types" : [ "street_number" ]

          "long_name" : "Viterbo",
          "short_name" : "Viterbo",
          "types" : [ "locality", "political" ]

          "long_name" : "Provincia di Viterbo",
          "short_name" : "VT",
          "types" : [ "administrative_area_level_2"

          "long_name" : "Italia",
          "short_name" : "IT",
          "types" : [ "country", "political" ]

          "location" : {
            "lat" : 42.4259435,
            "lng" : 12.0925167

            "long_name" : "Via Luigi Galvani",
            "short_name" : "Via Luigi Galvani",
            "types" : [ "route" ]

            "long_name" : "Viterbo",
            "short_name" : "Viterbo",
            "types" : [ "administrative_area_level_3", "political" ]

            "long_name" : "Lazio",
            "short_name" : "Lazio",
            "types" : [ "administrative_area_level_1", "political" ]

            "long_name" : "01100",
            "short_name" : "01100",
            "types" : [ "postal_code" ]

            "partial_match" : true,
            "place_id" : "types" : [ "street_address" ]
            "status" : "OK"
        }
      ]
    }
  ]
}
```

Geolocalizzazione Lazio: analisi dei risultati

Dimensione del campione **22.000** (distribuzione rettangolare, tasso di campionamento 1/20, bootstrap)

Casi Missing (indirizzo non geolocalizzabile) **0,03%**

Casi dubbi **0,5%** (localizzazione al di fuori del territorio della regione Lazio).

Casi approssimati **1,1%** (al baricentro del comune)

Geolocalizzazione Lazio: analisi testuale

L'informazione ottenuta è stata replicata in serie storica;

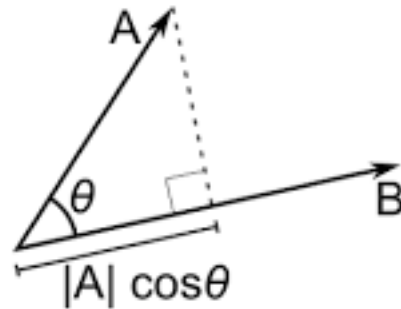
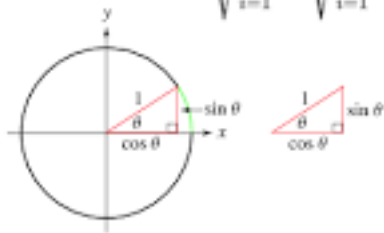
Per ottimizzare i risultati, un'analisi testuale è stata effettuata per confermare gli indirizzi delle stesse imprese negli anni precedenti

Uso di tre misure di similarità:

- Edit distance
- Jaro distance
- Cosin similarity (TF-IDF weighted)

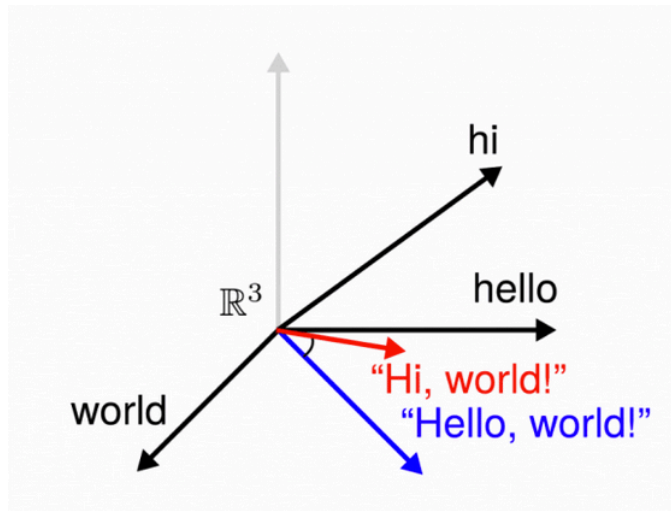
Geolocalizzazione Lazio: analisi testuale

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \|\mathbf{B}\|} = \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \sqrt{\sum_{i=1}^n B_i^2}}$$



$$w_{i,j} = tf_{i,j} \times \log \left(\frac{N}{df_i} \right)$$

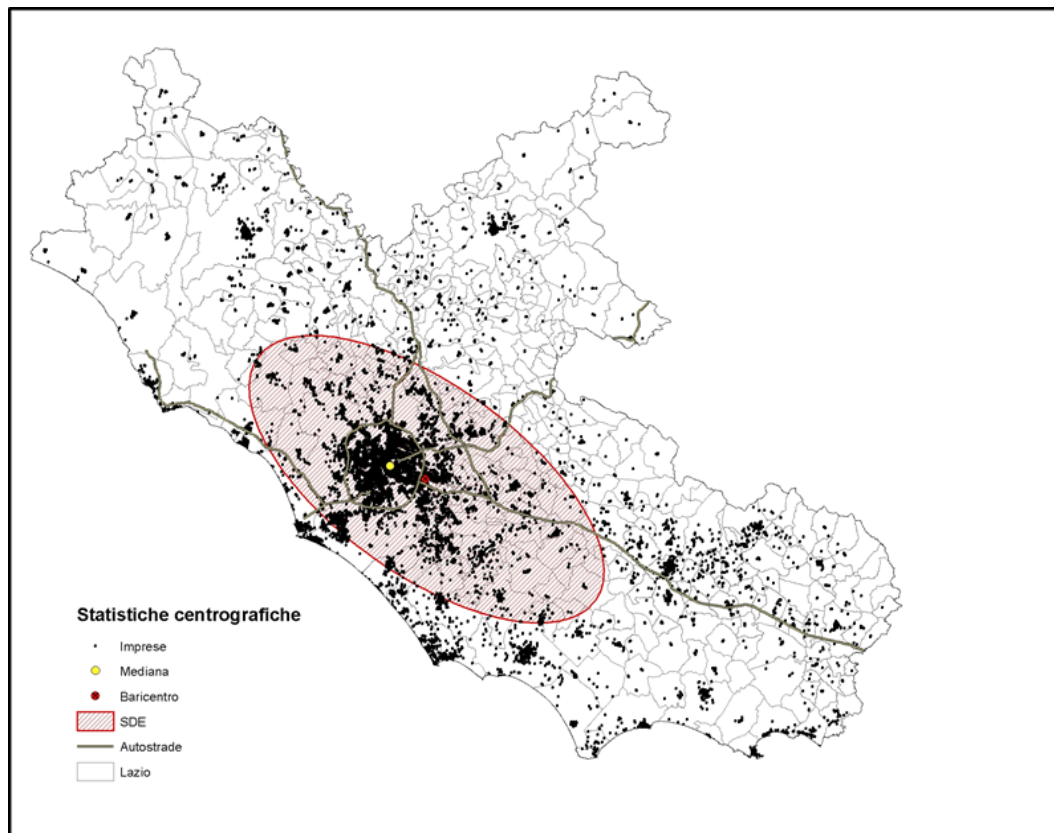
$tf_{i,j}$ = number of occurrences of i in j
 df_i = number of documents containing i
 N = total number of documents



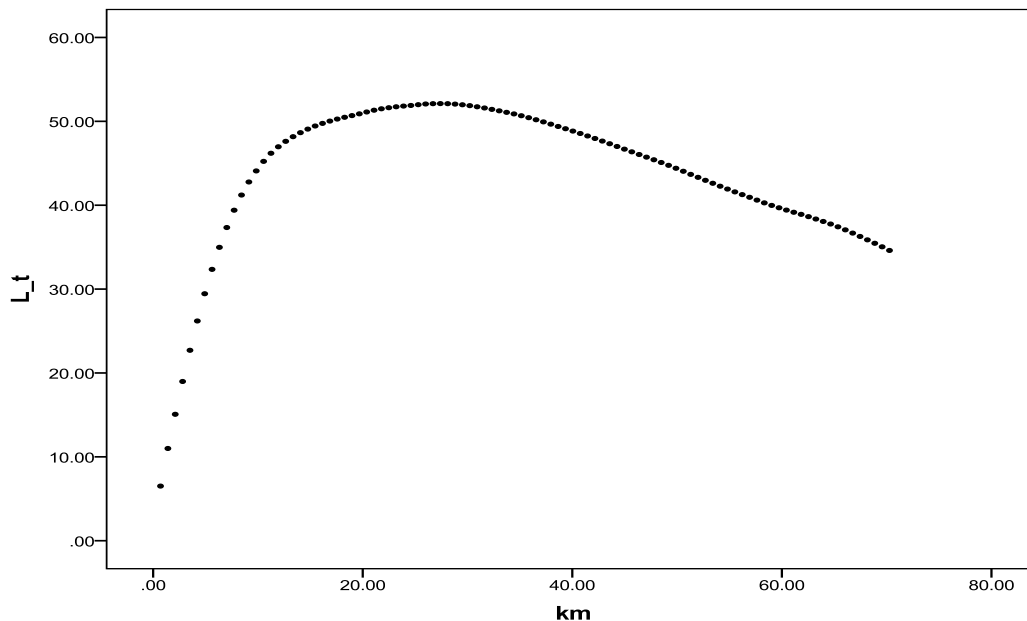
Indirizzo 1	Indirizzo 2
via	v.
Bellegra	Bellegra
35	35

Vettore	Freq1	Freq2	TFIDF
via	1	0	0,3456
v.	0	1	0,2345
Bellegra	1	1	3,487
35	1	1	1,87

Configurazione spaziale delle imprese: statistiche centrografiche



Configurazione spaziale delle imprese: analisi di vicinato e cluster



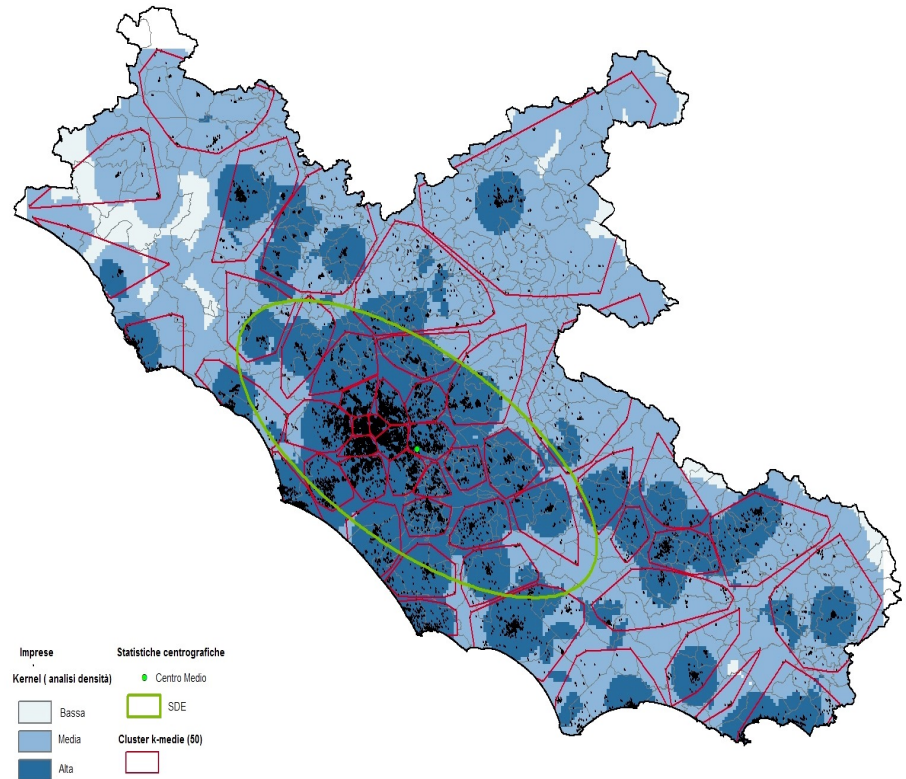
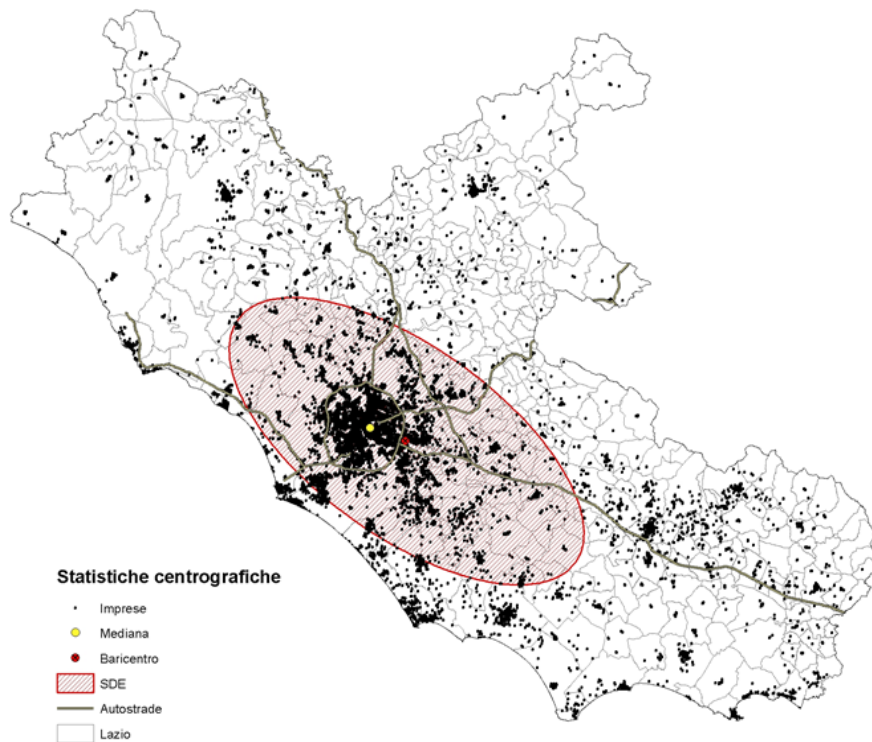
L'analisi di vicinato mette in evidenza la presenza di un processo di tipo aggregativo (le imprese tendono a localizzarsi vicine le une dalle altre) che aumenta fino a raggiungere la massima intensità ad una distanza di 23 km per poi diminuire molto lentamente, ma mantenendo sempre valori positivi e statisticamente significativi

Configurazione spaziale delle imprese: analisi di vicinato e cluster

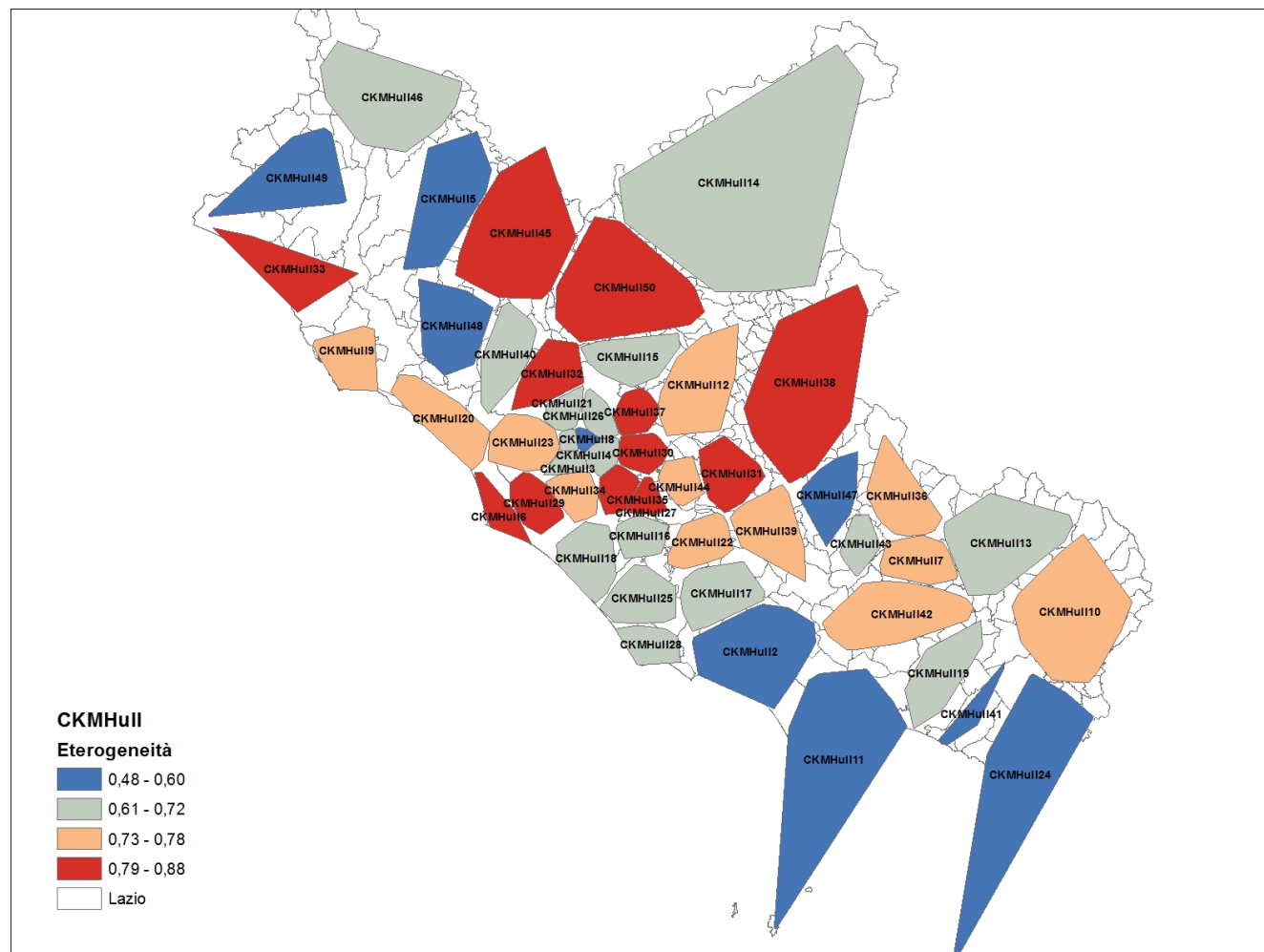
Seguendo le indicazioni venute dall'analisi di vicinato è sembrato interessante cercare di aggregare le imprese in cluster territoriali che segnano a loro volta una partizione del territorio Laziale. A tale scopo si è eseguita una prima cluster gerarchica con il metodo Nearest Neighbor Hierarchical Spatial Clustering - Nnh (Levin et. Al.), tra le diverse soluzioni si è scelta quella a 50 cluster. Si è quindi proceduto con una seconda analisi utilizzando il metodo k-means.

Per una quota residua di imprese pari al 2% la procedura di clusterizzazione non ha individuato nessun raggruppamento. Queste imprese sono imputate a un cluster fittizio (CKMHull 00).

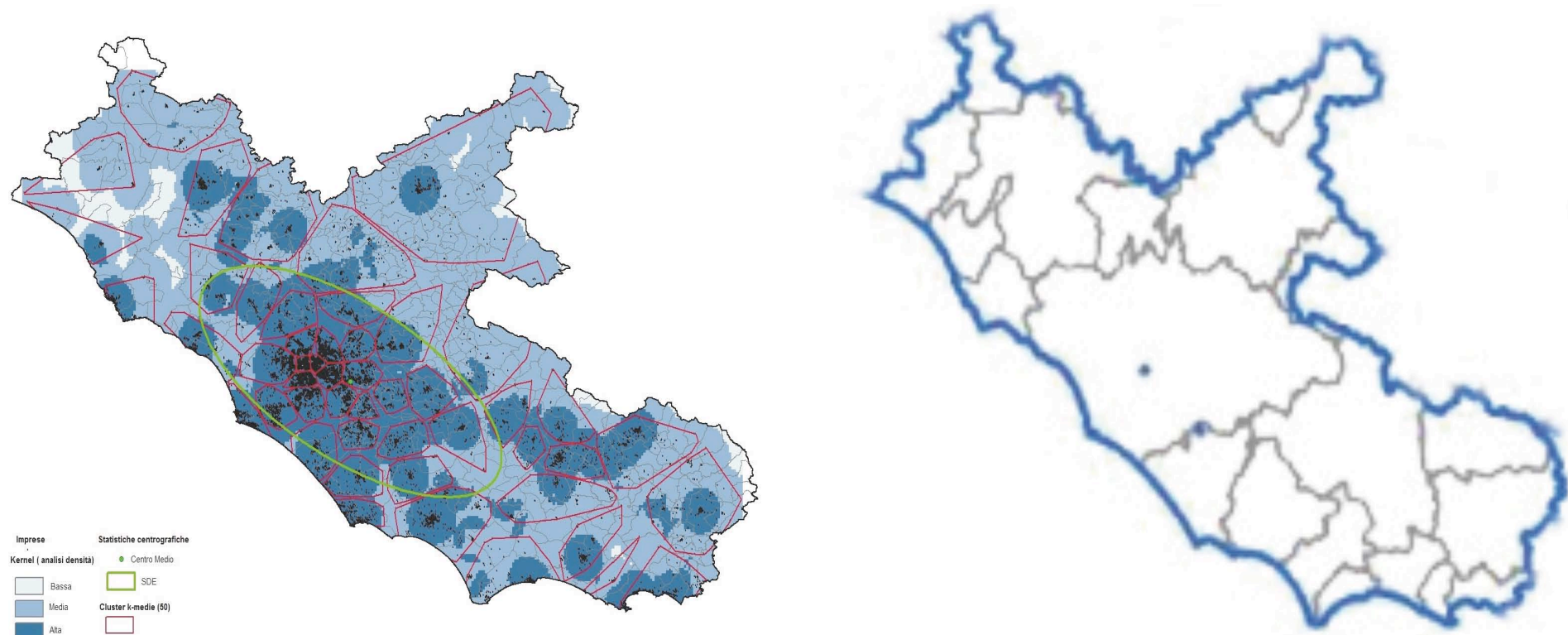
Configurazione spaziale delle imprese: cluster



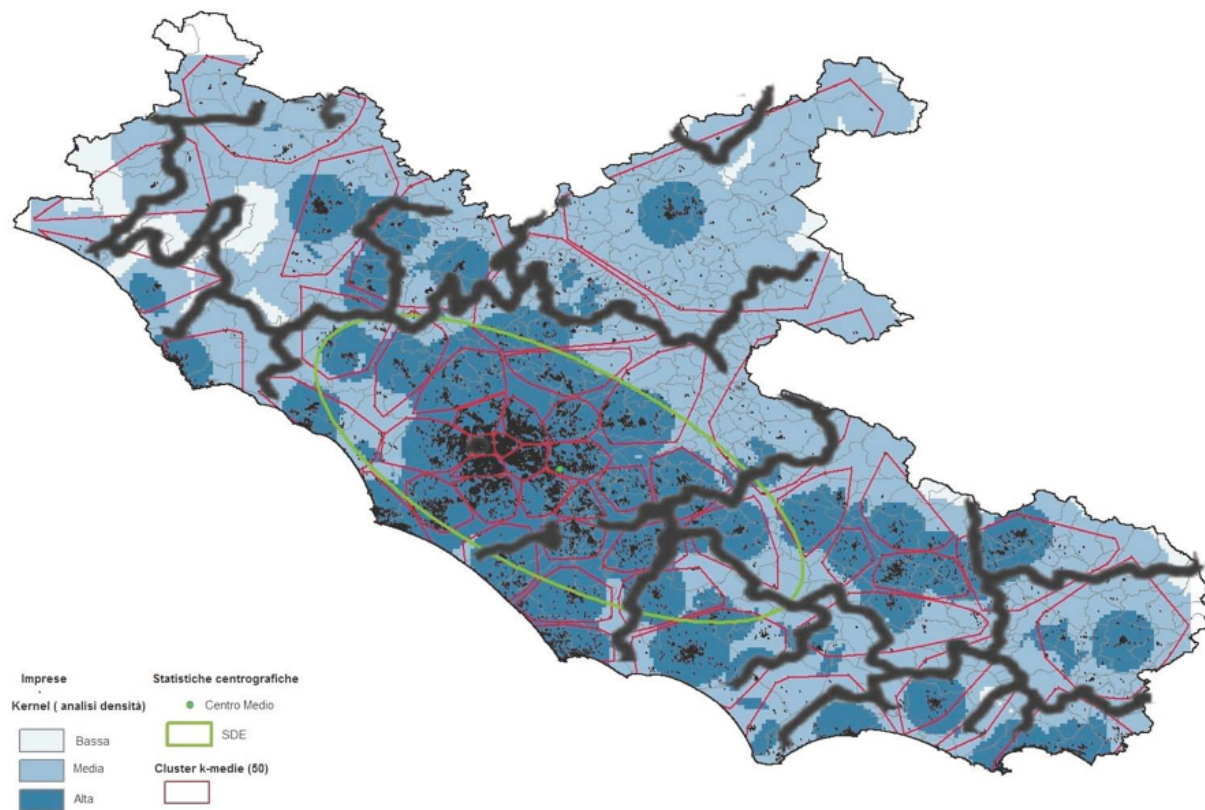
Configurazione spaziale delle imprese: eterogeneità



Partizioni del territorio: cluster di imprese (sx) e SLL (dx)



Analisi multidimensionale: sovrapposizione delle partizioni



Analisi multidimensionale: sovrapposizione delle partizioni

Si nota come il Sistema Locale di Roma possa essere partizionato in un elevato numero di cluster di imprese, peraltro di caratteristiche assai differenti in termini di produttività del lavoro e settore di attività economica.

Ma anche gli altri SLL possono essere ulteriormente definiti, per separazione o accorpamento.

La partizione multivariata del territorio è molto informativa, perché consente di individuare dei gruppi maggiormente compatti che possono essere sfruttati per molteplici fini di analisi e policy.

Ad esempio definendo diverse categorie di imprese/lavoratori, potenziali recipients di politiche differenziate, che potrebbero notevolmente migliorare il *tuning* delle policy regionali (Fondi Strutturali Europei), spesso basate su basi conoscitive *ad hoc* e scarsamente standardizzate.

Geolocalizzazione Italia: analisi dei risultati

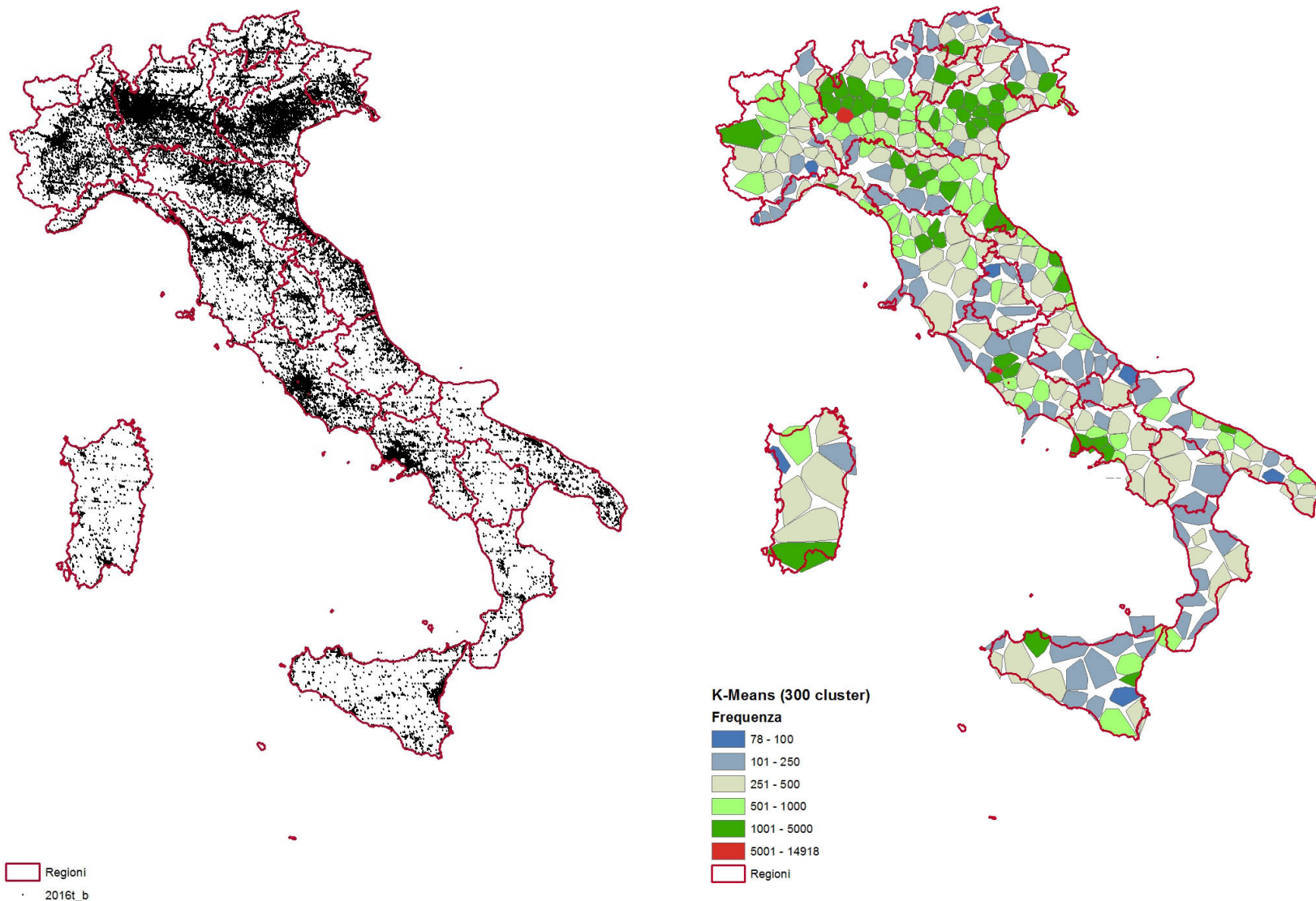
La tabella sintetizza i risultati ottenuti su un totale di **422.480**, l'universo delle imprese con almeno 10 addetti (anni 2012 e 2016).

n. indirizzi	%	Livello precisione	Descrizione livello precisione
242.032	57	0	Indirizzi correttamente geo localizzati attraverso il servizio LocationIQ
135.242	32	1	Indirizzi geo localizzati approssimando a livello di indirizzo (stessa via)
41.746	9	2	Indirizzi geo localizzati approssimando con il centroide calcolato tra cap e comune
1.610	0.3	3	Indirizzi geo localizzati approssimando con il centroide del comune
1.850	0.4	-	Indirizzi non geo localizzati

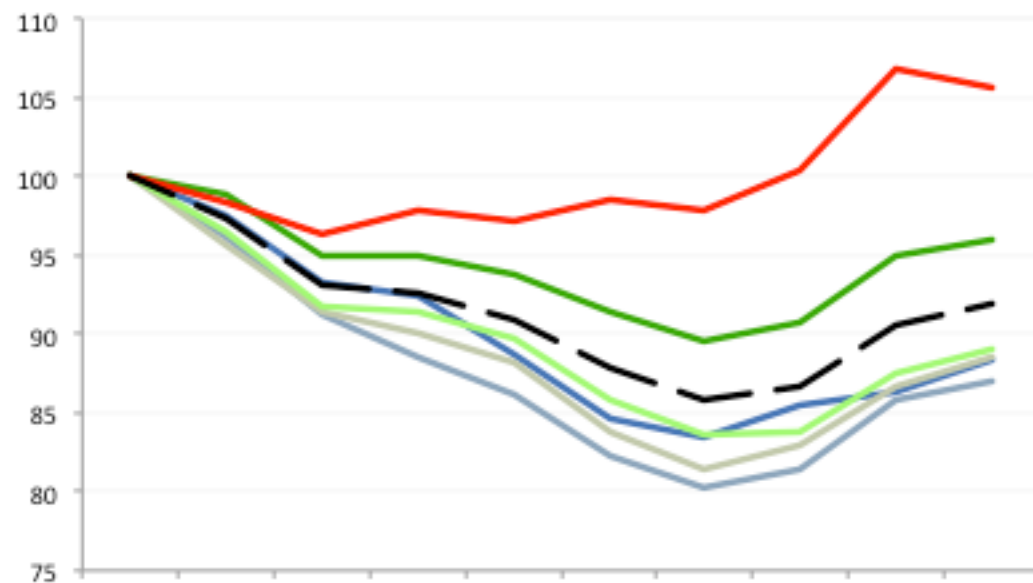
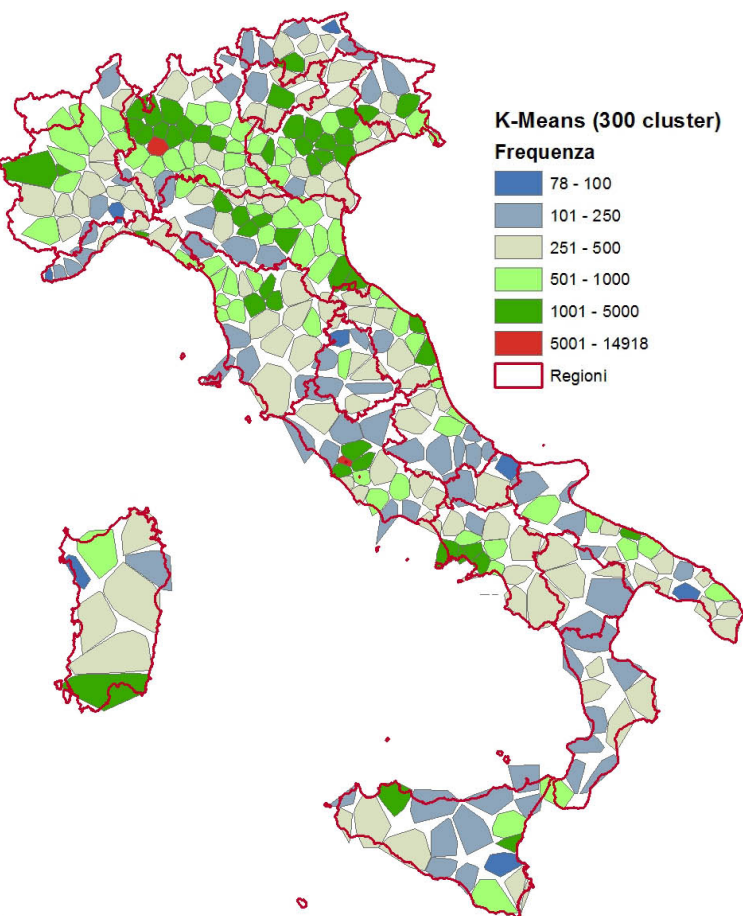
Ricorso a una molteplicità di servizi di geocoding, tramite chiamate server; maggiore capacità di referenziazione, ma qualità del risultato variabile

Necessità di uno specifico algoritmo per bonificare gli indirizzi errati

Analisi dei risultati: cluster di imprese georeferenziate

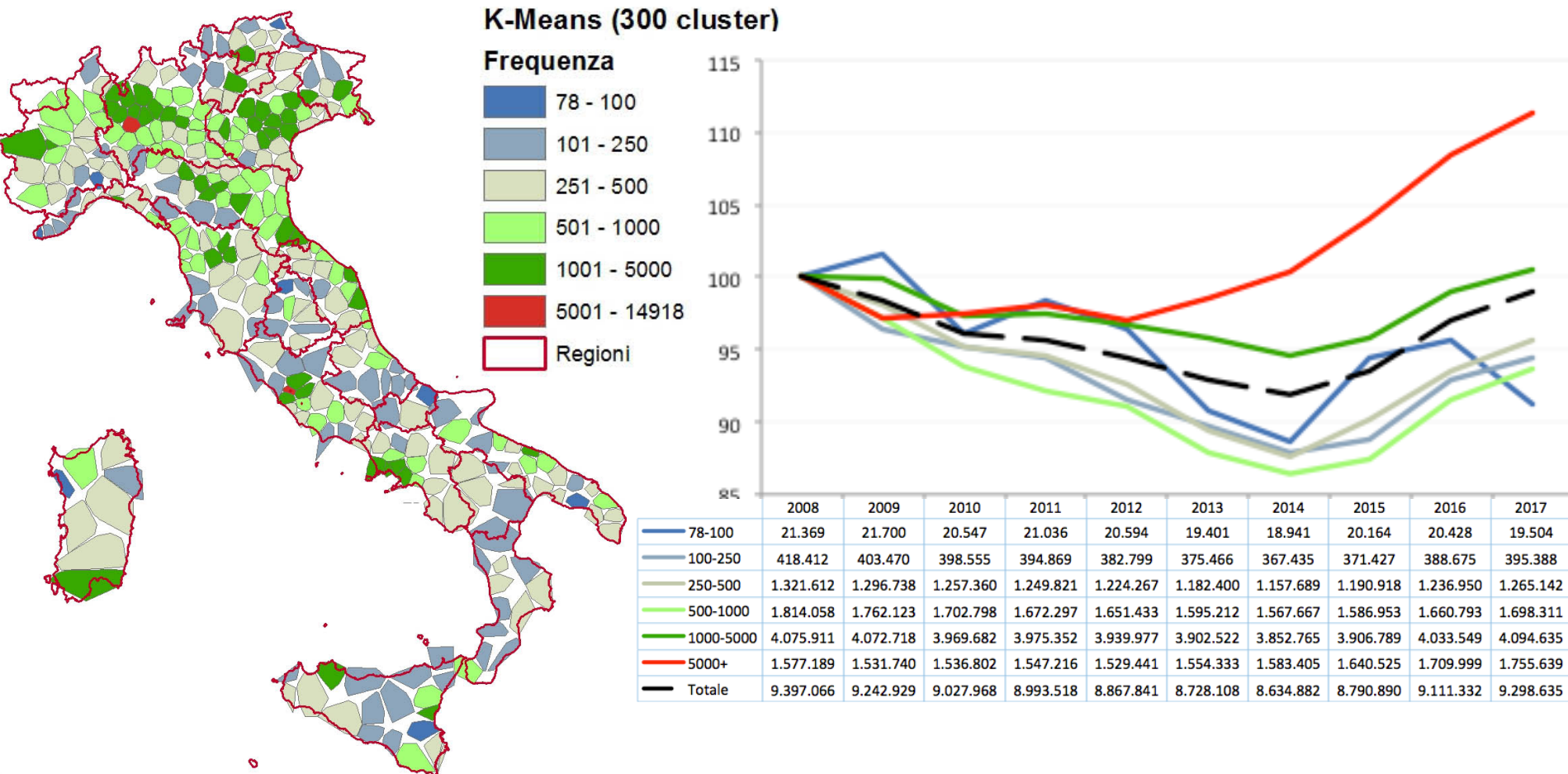


Analisi dei risultati: imprese



	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017
78-100	813	792	758	751	721	688	678	695	702	719
100-250	14.370	13.818	13.121	12.713	12.385	11.819	11.533	11.694	12.325	12494
250-500	41.830	40.011	38.210	37.661	36.874	35.078	34.039	34.717	36.237	37044
500-1000	53.829	51.929	49.403	49.159	48.274	46.238	44.971	45.138	47.140	47900
1000-5000	94.855	93.696	90.014	90.070	88.979	86.757	84.915	86.067	90.054	90964
5000+	20.317	19.968	19.572	19.891	19.732	20.013	19.893	20.378	21.697	21443
Totale	231.818	225.521	215.779	214.490	210.801	203.809	198.780	200.958	209.857	212936

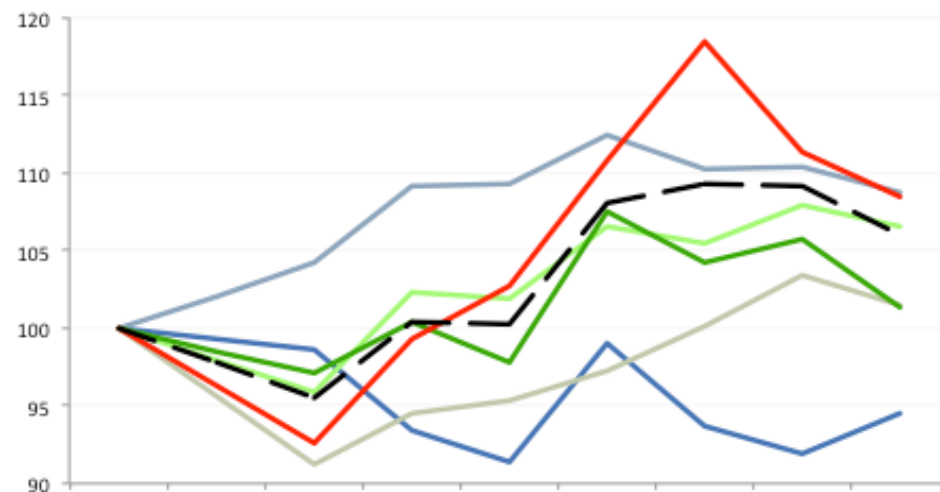
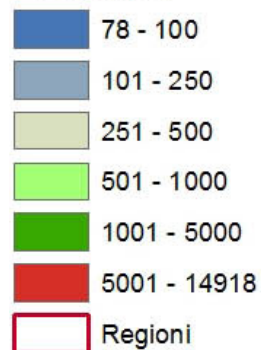
Analisi dei risultati: addetti



Analisi dei risultati: fatturato per addetto

K-Means (300 cluster)

Frequenza

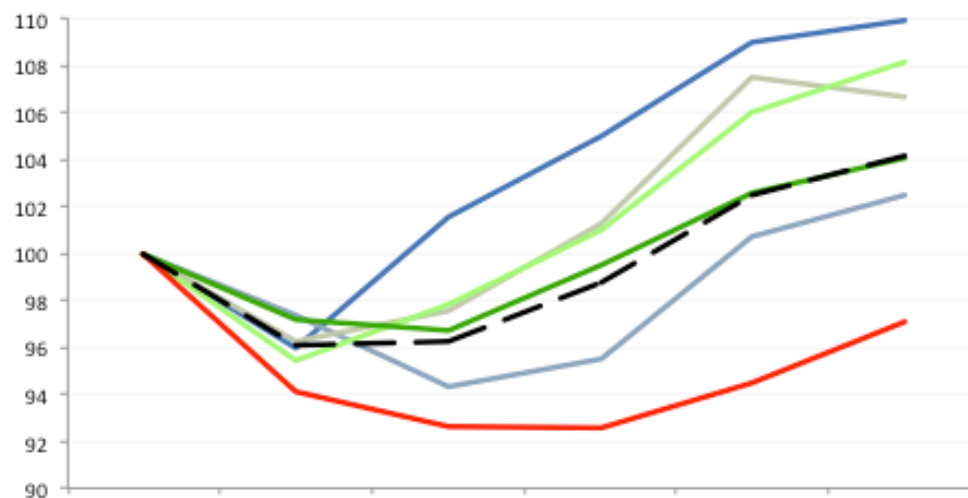
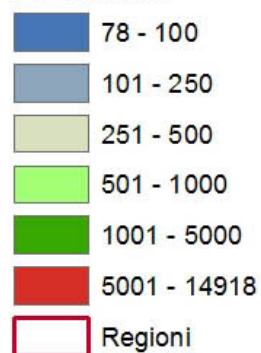


	2008	2009	2010	2011	2012	2013	2014	2015	2016
78-100	155.994	154.948	153.902	145.681	142.559	154.383	146.115	143.367	147.421
100-250	184.998	188.878	192.757	201.960	202.287	207.893	203.838	204.333	201.115
250-500	204.721	195.720	186.718	193.407	195.263	199.141	204.861	211.809	207.873
500-1000	229.208	224.415	219.622	234.527	233.669	244.089	241.858	247.439	244.080
1000-5000	246.266	242.692	239.118	247.161	240.897	264.733	256.584	260.458	249.677
5000+	380.332	366.248	352.164	377.927	390.473	421.411	450.321	423.496	412.678
Totale	255.796	250.067	244.337	256.893	256.444	276.525	279.658	279.045	271.258

Analisi dei risultati: valore aggiunto per addetto

K-Means (300 cluster)

Frequenza



	2011	2012	2013	2014	2015	2016
78-100	41.414	39.732	42.057	43.487	45.139	45.511
100-250	44.577	43.392	42.047	42.569	44.885	45.686
250-500	46.593	44.825	45.437	47.184	50.096	49.704
500-1000	51.294	48.954	50.166	51.798	54.396	55.459
1000-5000	54.460	52.932	52.649	54.209	55.884	56.654
5000+	75.426	71.001	69.833	69.825	71.249	73.242
Totale	55.805	53.623	53.717	55.120	57.188	58.123

Analisi dei risultati: valore aggiunto per addetto

K-Means (300 cluster)

Frequenza

78 - 100

101 - 250

251 - 500

501 - 1000

1001 - 5000

5001 - 14918

Regioni

FATTURATO (Coefficiente di variazione)

CLUSTER	2011	2012	2013	2014	2015	2016
78-100	132	141	136	138	140	191
100-250	704	797	765	822	662	1.044
250-500	446	519	513	612	532	715
500-1000	690	692	678	660	577	828
1000-5000	1.298	1.472	1.934	1.618	1.548	1.675
5000+	1.037	1.146	1.207	1.710	1.486	1.487
Totale	1.064	1.199	1.473	1.638	1.376	1.645

MOL su FATTURATO

CLUSTER	2011	2012	2013	2014	2015	2016
78-100	9,0%	7,2%	8,1%	9,4%	10,6%	9,4%
100-250	6,7%	6,3%	5,4%	5,5%	6,3%	6,8%
250-500	7,9%	6,8%	6,9%	7,4%	8,2%	8,1%
500-1000	7,8%	6,8%	6,8%	7,3%	7,9%	8,4%
1000-5000	7,7%	7,3%	6,4%	7,0%	7,4%	8,1%
5000+	9,4%	7,9%	7,0%	6,4%	6,9%	7,5%
Totale	8,1%	7,3%	6,6%	6,9%	7,4%	8,0%

COSTO LAVORO PER ADDETTO

CLUSTER	2011	2012	2013	2014	2015	2016
78-100	28.359	29.464	29.510	29.697	29.976	31.695
100-250	30.994	30.720	30.920	31.276	31.948	31.988
250-500	31.355	31.561	31.600	32.058	32.758	32.802
500-1000	32.976	32.973	33.513	34.182	34.878	34.902
1000-5000	35.448	35.376	35.807	36.167	36.570	36.335
5000+	39.969	40.149	40.406	41.013	41.817	42.241
Totale	34.936	34.970	35.367	35.890	36.499	36.507

Conclusioni “tecniche”

- Possibilità di costruzione di basi dati utilizzabili per molteplici scopi
- Necessità di competenze tecniche (informatiche, metodologiche, di analisi)
- Capacità di inquadrare problemi e criticità in forma insolita e innovativa
- In negativo possibilità di sforzi notevoli, senza sostanziali miglioramenti conoscitivi

Conclusioni “tematiche”

La realizzazione di diversi piani di lettura del territorio (partizioni funzionali del territorio di natura socio-economica) attraverso metodologie data driven e la loro sovrapposizione consentirebbe l'individuazione di aree omogenee (“compatte”), da scegliere e definirsi rispetto ai molteplici scopi conoscitivi (analisi/policy) possibili:

- Realizzare una serie di report tematici sul territorio, volti a coprire varie aree di interesse, con uno speciale focus su tematiche di importanza (es.: distretti industriali)
- Realizzazione di basi dati e metodologie per l'analisi tematica, l'implementazione delle politiche territoriali e le valutazioni controfattuali dei Fondi Strutturali Europei (es.: valutazioni controfattuali sull'occupabilità dei beneficiari del Fondo Sociale, modelli econometrici demografici/economici per la stima dei fabbisogni occupazionali ecc.)

GRAZIE DELL'ATTENZIONE!