

SPATIAL MODELLING E IMAGE PROCESSING PER L'ANALISI SPAZIALE DELLE AREE RURALI: OPEN DATA E BIG DATA A SUPPORTO DELLA STATISTICA UFFICIALE

Bianchino A.<sup>1</sup>, Cozzi M.<sup>2</sup>, Fusco D.<sup>3</sup>, Romano S.<sup>4</sup>, Viccaro M.<sup>5</sup>

**SOMMARIO**

L'uso dei dati amministrativi, di concerto con le indagini statistiche, ha ridotto, e in taluni casi eliminato, i problemi di sotto copertura delle unità statistiche rispetto all'universo di riferimento. Tuttavia per alcuni fenomeni si continua ad avere difficoltà nell'individuare l'insieme delle unità oggetto di rilevazione poiché esse sfuggono sia alle fonti primarie che a quelle secondarie. È il caso di alcuni segmenti dell'agricoltura per i quali, a causa dell'elevata frammentazione delle proprietà e la variabilità delle coltivazioni, non risulta agevole ottenere coperture di informazioni complete.

Allo stato attuale gli strumenti maggiormente impiegati per la identificazione delle coltivazioni agronomiche possono essere classificati secondo procedimenti di tipo diretto ed indiretto. I primi fanno essenzialmente riferimento alle rilevazioni on-site, mentre per i secondi risulta necessario un procedimento di interpretazione di immagini georiferite (aeree e da satellite), che può essere in alcuni casi condotta anche in modalità semi o automatiche. Grazie all'esponenziale aumento di basi spaziali che si sta avendo negli ultimi decenni, il secondo procedimento offre maggiori possibilità di sviluppo, anche se ad oggi, sono presenti alcune limitazioni determinate essenzialmente dalla risoluzione delle immagini e dagli algoritmi di classificazione.

Con lo scopo di ridurre il gap conoscitivo, con il presente lavoro è proposto un modello di indagine sperimentale, volto all'utilizzo delle immagini satellitari per l'individuazione delle tipologie colturali e delle relative superfici attraverso l'interpretazione delle **bande spettrali**. Tali superfici, una volta attribuite le informazioni catastali, sono identificabili all'interno degli archivi amministrativi e statistici, consentendo di consolidare le serie di dati già prodotti e, al contempo, di costituire nuove serie di informazioni, desumibili a partire dall'impiego dei Big Data e dall'interpretazione spaziale del territorio.

---

<sup>1</sup> Istat – Dipartimento per la raccolta dati, Via Pretoria 342, 85100, Potenza, e-mail: bianchin@istat.it.

<sup>2</sup> Università degli Studi della Basilicata, Via Nazario Sauro 85, 85100, Potenza, e-mail: mario.cozzi@unibas.it.

<sup>3</sup> Istat – Dipartimento per la raccolta dati, Via Chiaia 199/D, 80121, Napoli, e-mail: dafusco@istat.it.

(corresponding author).

<sup>4</sup> Università degli Studi della Basilicata, Via Nazario Sauro 85, 85100, Potenza, e-mail: severino.romano@unibas.it.

<sup>5</sup> Università degli Studi della Basilicata, Via Nazario Sauro 85, 85100, Potenza, e-mail: mauro.viccaro@unibas.it

## 1. Introduzione

Negli ultimi anni, la quantità di dati digitali creati, archiviati ed elaborati nel mondo è cresciuta esponenzialmente. Si stima che ogni minuto l'umanità generi dati per 1,7 milioni di miliardi di byte, pari a 360.000 DVD: più di 6 megabyte di dati a testa ogni giorno. Di contro, la richiesta di informazioni statistiche non è mai stata così evidente (Radermacher, 2018). Pertanto i Big Data rappresentano una grossa opportunità per ampliare le opportunità di analisi, aumentare la tempestività di informazioni e contribuire ad aumentare la qualità delle stime.

Le tecniche tradizionali di raccolta dei dati e i dati amministrativi non sempre riescono a ricoprire in maniera esaustiva tutti gli universi di riferimento. Alcuni fenomeni sfuggono a queste fonti perché non è sempre possibile eseguire operazioni di collegamento tra i dati e il territorio (Belin e Rubin, 1995) o tra due basi di dati diversi. Per questo motivo, gli istituti nazionali di statistica (NSI) hanno avviato la sperimentazione dell'uso di nuove fonti di dati, più ricche di informazioni e meglio distribuite in termini geografici.

Tra i settori che presentano questa tipologia di problema vi sono alcuni segmenti dell'agricoltura. È il caso ad esempio delle colture ortofrutticole. Tale settore, che si caratterizza per un'elevata frammentazione, sia spaziale che della proprietà fondiaria, rappresenta un importante segmento per l'agricoltura italiana, avendo un peso rilevante per l'economia e l'occupazione locale. Si pone, a questo punto, il problema di trovare soluzioni capaci di superare questo gap informativo.

L'agricoltura occupa vaste porzioni di terra con un'elevata variabilità temporale in termini di tipo di colture e superfici utilizzate. Recenti applicazioni innovative in questo campo di ricerca includono il telerilevamento per costruire mappe di copertura del suolo, come ad esempio effettuato in Europa con il programma Corine Land Cover<sup>6</sup> e negli Stati Uniti con il programma Cropland Data Layer<sup>7</sup>. Queste tecniche non solo hanno avuto un impatto sul processo, sulla qualità e sulla tempestività delle statistiche agricole, ma hanno anche aperto finestre statistiche per altri settori (Miller et al., 2009).

I possibili utilizzi di questi dati possono essere molteplici, tra cui:

1. Posizionamento e dimensionamento delle colture;
2. Individuazione del tipo esatto di coltura per un determinato appezzamento di terreno;
3. Stima delle produzioni ottenibili nel rispetto delle condizioni climatiche (precipitazioni, temperature, ecc.).

Obiettivo di questo studio è identificare un modello che utilizzi i Big Data, e nello specifico immagini satellitari ad alta risoluzione, come nuove fonti di dati a supporto della produzione statistica ufficiale, al fine di identificare i fenomeni che sfuggono alle fonti tradizionali. In particolare la procedura proposta può essere applicata ai registri statistici al fine di ottenere un quadro completo di unità statistiche, recuperando anche eventuali dati mancanti.

## 2. Il quadro di riferimento

Per i produttori e gli utilizzatori di dati statistici risulta sempre più interessante il potenziale legato al riutilizzo delle fonti di dati esterne alla statistica ufficiale per la produzione dei dati. Inoltre, l'uso di fonti di dati esterne ed esistenti, determina una riduzione dell'onere per gli intervistati e spesso riduce il costo degli istituti di statistica nazionali (Wallgren & Wallgren, 2014). Oggi, i dati amministrativi sono ampiamente utilizzati a supporto o sostituzione delle indagini. Si definiscono dati amministrativi le informazioni raccolte e conservate da istituzioni pubbliche o enti titolari di processi in ragione dei loro fini istituzionali. L'attività degli NSI è di elaborazione di tali dati non statistici che costituiscono patrimonio dell'ente titolare del processo, per renderli adatti all'uso statistico (ad es. anagrafi comunale).

---

<sup>6</sup> <https://land.copernicus.eu/pan-european/corine-land-cover>

<sup>7</sup> <https://data.nal.usda.gov/dataset/cropscape-cropland-data-layer>

Generalmente questo tipo di dati coprono l'universo delle unità analizzate (persone, imprese, aziende agricole, ecc.), ma a volte è impossibile garantire la copertura totale. Per questo motivo, gli NSI sono alla ricerca di nuove tecniche e nuovi strumenti, anche sperimentali, di raccolta ed elaborazione dei dati. In particolare, un filone di ricerca promettente appare connesso all'utilizzo integrato di Big Data e Open Data a fini statistici.

La funzione centrale dei Big Data è quella di fornire la migliore rappresentazione possibile della realtà attraverso dati che coprono tutti i tipi di informazioni, ad es. sensori e dispositivi di localizzazione (GPS, traffico), registrazione online di ricerche e visite a pagine e registrazioni di social media e opinioni. Ma questa è anche un'area molto complessa, non solo da un punto di vista statistico, ma anche in termini di molti altri aspetti che sono cruciali per la produzione statistica come politica, legislazione e comunicazione.

Gli Open Data sono dati che possono essere liberamente utilizzati, condivisi e integrati da chiunque, ovunque e per qualsiasi scopo. Sono disponibili gratuitamente, o almeno a un costo di riproduzione ragionevole; sono inoltre digitali, preferibilmente disponibili tramite download da Internet e facilmente elaborabili anche da un computer. Gli Open Data devono consentire alle persone di usarli, riutilizzarli e ridistribuirli. Quando gli Open Data corrispondono per dimensione e tempestività con la definizione fornita per i Big Data, Open Data e Big Data coincidono (Cavanillas J. M. et al., 2016).

In molti progetti lanciati da EUROSTAT, l'utilizzo di Big Data in combinazione con dati statistici ufficiali esistenti è pensato per migliorare le statistiche attuali e crearne nuove in vari settori, come l'agricoltura. Tra questi, il progetto ESSNET BIG DATA (I e II)<sup>8</sup>, sviluppato nell'ambito del sistema statistico europeo (ESS) e attuato da 28 partner (tutti enti nazionali di statistica, incluso l'ISTAT), mira a utilizzare i Big Data come strumento per la raccolta tradizionale e non, da integrare nella produzione di statistiche ufficiali. Ciò consentirebbe di arricchire la conoscenza rispetto a fenomeni specifici.

### **3. L'analisi spaziale delle aree rurali**

#### *4.1. Il contesto*

Il settore ortofrutticolo rappresenta un segmento fondamentale per l'agricoltura italiana, sia per la diffusione territoriale delle aziende agricole che per i valori produttivi ed economici che caratterizzano questo settore.

Nel 2017 le esportazioni italiane di frutta e verdura sono salite ad un valore record di 5,1 miliardi di euro, in aumento del 2,5% rispetto al 2016 (fonte Istat), diventando il comparto agroalimentare "made in Italy" più amato all'estero immediatamente dopo il vino. L'importanza del segmento riguarda in particolare alcuni contesti territoriali, in cui il settore ortofrutticolo svolge un ruolo rilevante per l'economia locale e l'occupazione. Nel sud Italia c'è una forte concentrazione in alcune aree piuttosto che in altre. A questo scopo si è deciso di approfondire questa trattazione nella Piana del Sele (regione Campania) e nel Metapontino (regione Basilicata).

Il settore è caratterizzato da una forte frammentazione delle aziende agricole e per questo motivo è difficile identificarle e la copertura è insufficiente anche nei dati amministrativi. È stato quindi deciso di avviare una sperimentazione sull'uso di fonti di dati alternative per identificare terre e aziende agricole.

#### *4.2. Obiettivi*

L'obiettivo principale del progetto è testare nuove fonti e metodi statistici per l'identificazione delle unità statistiche e delle loro caratteristiche a supporto dei registri statistici economici. La sperimentazione riguarda aziende agricole e terreni ortofrutticoli. Tradizionalmente, i dati agricoli vengono acquisiti attraverso

---

<sup>8</sup> matching administrative data i.e. LPIS with big data i.e. satellite imagery in agriculture statistics; [https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet\\_Big\\_Data](https://webgate.ec.europa.eu/fpfis/mwikis/essnetbigdata/index.php/ESSnet_Big_Data)

censimenti e indagini sul campo che sono molto costosi e non forniscono informazioni sufficientemente dettagliate per determinare l'estensione o la distribuzione geografica delle principali colture.

Le osservazioni satellitari e aeree, grazie alla loro natura sinottica e ripetitiva, hanno il vantaggio di fornire informazioni tempestive e spazialmente contigue sulla crescita delle colture su scala regionale e globale (Cai et al., 2018; King et al. 2017; Song et al., 2017).

L'obiettivo principale può dunque essere dettagliato come segue:

- Ottenere un frame completo di unità statistiche;
- Integrare i valori mancanti nei registri;
- Validare le informazioni presenti sui registri.

#### *4.3. Spatial modelling e image processing: la proposta metodologica*

L'uso delle immagini satellitari può essere particolarmente efficace nel migliorare la strategia globale di costruzione del sistema informativo statistico dell'agricoltura (GSARS, 2017). Tale uso è altamente versatile, potendo variare dalla ottimizzazione della progettazione del campionamento, per facilitare il lavoro sul campo dei rilevatori, alla garanzia della qualità e produzione uniforme dei dati.

La possibilità di avere a disposizione immagini con una risoluzione alta o altissima (da pixel di 5 metri a pixel di 5 centimetri di lato) terra può garantire l'interpretazione di dettaglio sulla tipologia colturale coltivata; inoltre i brevi intervalli temporali della ripresa consentono di rilevare anche l'avvicendamento delle colture.

In termini di caratteristiche dei sensori satellitari, una prima suddivisione può essere operata in base al funzionamento del sensore nel dominio spettrale visibile e infrarosso (multispettrale) o nello spettro delle microonde (SAR). Questa distinzione determina se le immagini possono essere acquisite indipendentemente dall'illuminazione solare e dalla copertura nuvolosa. I sensori multispettrali misurano la luce solare riflessa e come tale dipendono ovviamente dal fatto che vi sia sufficiente illuminazione solare e assenza di copertura nuvolosa. Questo è importante non solo per le acquisizioni istantanee ma anche per la creazione di serie temporali coerenti, ad esempio per confrontare i dati in diverse stagioni delle colture.

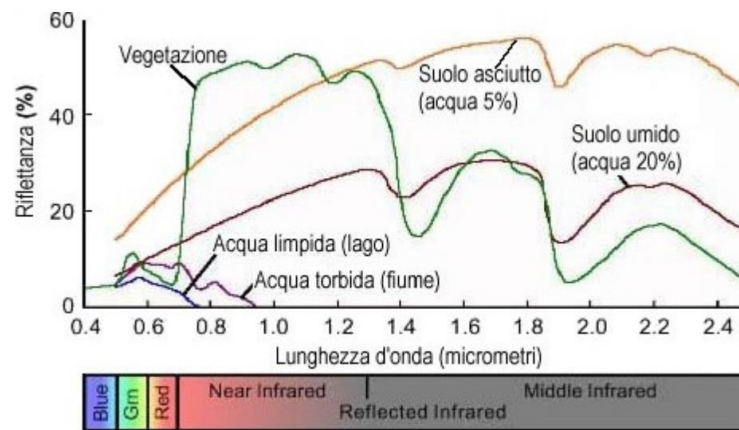
La seconda suddivisione in gruppi di sensori dipende dalla risoluzione spaziale. La denominazione dell'European Space Agency (ESA) fa affidamento alle gamme di risoluzione <4 m, 4–10 m, 10–30 m, 30–300 m e > 300 m per raggruppare le famiglie di sensori. Oltre alla differenza di risoluzione, la gamma di 10–30 m delimita anche la divisione tra i dati disponibili gratuitamente e quelle categorie (<10 m) che rientrano nelle immagini commerciali.

Da un punto di vista metodologico, scelta la modalità di acquisizione delle immagini (Open data o Big Data), il passaggio successivo comporta l'interpretazione delle immagini attraverso la risposta multispettrale (Figura 1).

È importante che le immagini rispettino quindi determinati requisiti tecnici di base, quali registrazione coerente tra diverse bande di dati multibanda, ridimensionamento radiometrico e l'assenza di eccessiva saturazione o compressione dell'intervallo.

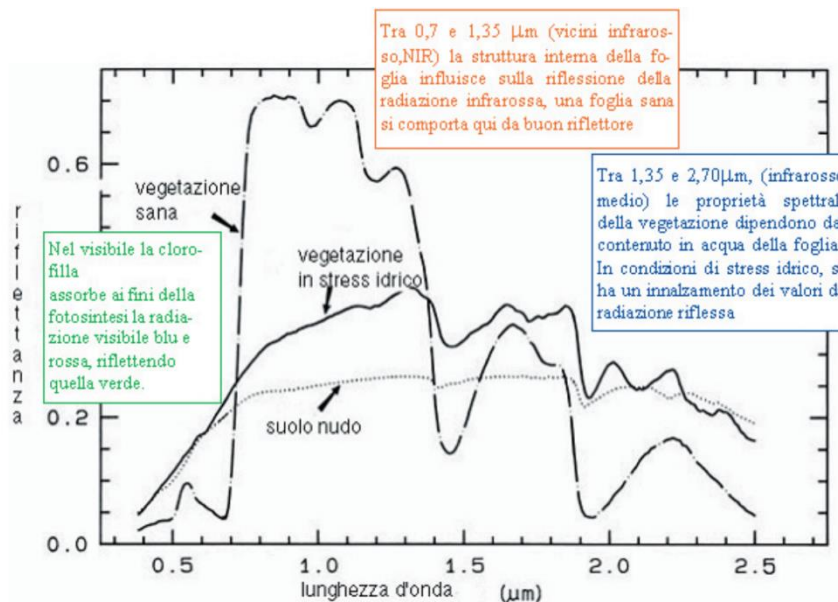
L'interpretazione parte dall'assunto che il comportamento tipico della vegetazione in campo spettrale è determinato dall'apparato fogliare, la cui struttura e vitalità condiziona la riflettanza della radiazione solare (Figura 2).

Figura 1 – Firme spettrali di alcuni elementi naturali



Fonte: <http://dipsa.unibo.it/catgis/index.html>

Figura 2 – Curve di riflettanza a differenti lunghezze d'onda



Fonte: <http://dipsa.unibo.it/catgis/index.html>

Il riconoscimento delle colture presenti è effettuato sfruttando la multispettralità e la multitemporalità dei dati telerilevati. La prima operazione da effettuare è quella di individuare e comparare le tonalità cromatiche e successivamente di correlare le date di acquisizione delle immagini con i cicli fenologici delle colture. La disponibilità dei passaggi temporali consente di monitorare gran parte del ciclo di sviluppo delle colture agrarie di indirizzo<sup>9</sup>.

Con lo scopo di fornire una base conoscitiva solida, la prima fase dell'attività prevede una ricognizione del territorio da fotointerpretare, sfruttando le informazioni contenute anche nei dati ancillari esistenti e a disposizione.

A questa fase segue una fase di apprendimento della coltura, ottenuto mediante la correlazione tra la coltura rilevata on-site e la specifica firma spettrale prodotta. Nel caso in cui le condizioni di contorno

<sup>9</sup> Ad esempio, nel caso delle coltivazioni autunno-primaverili, in genere la prima immagine invernale è acquisita quando la coltura si trova nella fase di accestimento; la seconda in primavera, durante la fase di levata e la terza nella fase di maturazione.

(elementi del clima, periodo di impianto e/o semina) sono simili risulta possibile replicare in automatico la interpretazione della coltura rispetto ad un territorio maggiormente esteso.

Con un crescente background di dati e relazioni funzionali (ricognizione delle colture presenti, e delle loro fasi fenologiche nel corso delle stagioni) risulta successivamente possibile predisporre un manuale operativo di fotointerpretazione.

#### 4.3.1. La scelta delle immagini.

Le immagini utilizzabili nelle statistiche agricole possono essere fornite da un'ampia gamma di satelliti. L'ESA ha creato un'utile classificazione dei set di dati dei sensori disponibili. I sensori sono classificati in gruppi di "mission" che li caratterizzano e da essi dipende la risoluzione e la ripetitività dei passaggi sulla stessa area. Ovviamente per ciascuna tipologia di immagine ci può essere un utilizzo diverso per le statistiche agricole (tabella 1).

Tabella 1 – Classificazione dei sensori satellitari in categorie, basata sulla nomenclatura ESA

<b>Mission Group</b>	<b>Risoluzione</b>	<b>Esempi</b>	<b>Utilità nel contesto delle statistiche agricole</b>
SAR	Alta, 4-30 m	Sentinel-1, Radarsat-2	Da 6 a 24 giorni di passaggio sull'area; tutte le condizioni climatiche, notte e giorno; gratuito; Adatto a delineare e identificare le colture.
	Molto alta, 1-4 m	TerraSAR-X, Radarsat-2	Sensori multipli e passaggi variabili (anche più volte al giorno); Alta risoluzione, ma su piccole aree; Poco adatto alla identificazione delle colture; Costoso.
Multispettrale	Alta, 4-30 m	Landsat, Sentinel, SPOT	Da 5 a 16 giorni di passaggi sull'area su ampie zone; gratuite (Landsat e Sentinel fino a 10 m); SPOT combina 6 m multispettrali con 1,5 m pancromatici; Adatto a delineare e identificare le colture e la tipologia di coltura (4-10 m); A pagamento <10 m.
	Molto alta, 0,3-4 m	WorldView-3, Pleiades, DMC-III	Altissima risoluzione, numero variabile di band spettrali; adatto alla misurazione delle particelle (<1 m)

Fonte: nostra elaborazione su dati GSARS.

La risoluzione spaziale o geometrica fa riferimento alla dimensione dell'area minima rilevata al suolo dal sensore e quindi esprime la capacità dello strumento ad identificare due oggetti separati, definendo il grado di dettaglio dell'immagine. Nei dati digitali tale risoluzione coincide con la dimensione del pixel. Le immagini disponibili gratuitamente come Open Data (Landsat, Sentinel) consentono di avere almeno una disponibilità stagionale e una risoluzione in termini di pixel fino a 10 metri, mentre le immagini a pagamento (SPOT, Compsat, Ikonos, Quickbird) consentono di personalizzare la frequenza delle immagini e di avere una risoluzione fino a 5 centimetri.

*Figura 3 – Esempio di immagine SPOT*



Fonte: <https://earth.esa.int/web/eoportal/satellite-missions/s/spot-6-7>

Una limitazione pratica all'uso delle immagini satellitari commerciali riguarda i costi associati. Diversi fattori contribuiscono a determinare il costo totale dell'uso dei dati. Va notato che, nella maggior parte dei casi, l'uso dei dati è concesso in licenza: il cliente non diventa il proprietario dei dati, ma paga piuttosto il diritto di utilizzare i dati e ricavarne informazioni. Il tipo di licenza stessa può determinare i costi in misura considerevole. Ad esempio, la licenza monouso è generalmente il più economico, ma la condivisione dei dati può essere limitata a un singolo utente o a un gruppo di utenti pre-identificato.

Un caso ottimale potrebbe essere rappresentato dall'impiego congiunto delle immagini a media ed alta risoluzione, impiegando queste ultime su piccole aree al fine di utilizzare i risultati della interpretazione come validazione.

Va da sé che a seconda della qualità delle immagini utilizzate i risultati consentiranno di ottenere dei tipi di dettaglio diverso. Le immagini ad alta risoluzione consentiranno di ottenere informazioni sui macrousi (es. cereali per la produzione di granella, piante industriali, ortive in piena aria o in serra, alberi da frutto, ecc.) e su alcuni usi (es. vite e olivo), mentre le immagini ad altissima risoluzione consentiranno di ottenere informazioni su tutti gli usi (es. frumento, segale, orzo, melo, pero, ecc.) e su i dettagli (es. carciofo, pomodoro, zucchina, ecc.).

#### *4.3.3. Sovrapposizione dei poligoni e aggancio con il registro delle aziende agricole.*

Allo scopo di attribuire l'appartenenza delle superfici interpretate si procede ad un overlay mapping con le mappe catastali (Figura 3).

Figura 3 – Overlay mapping



Fonte: portale statistico regione Basilicata

In alcuni casi si potrà verificare nella singola particella catastale siano presenti più colture. Allo scopo, il Sistema Informativo Geografico implementato sarà in grado di determinare le percentuali colturali per ciascuna proprietà.

A questo punto, si avrà a disposizione l'esatta coltivazione presente su ciascuna particella catastale.

Di fatto, questo tipo di informazione è presente nel Registro Statistico delle aziende agricole o Farm register, il registro di base della popolazione delle aziende agricole orientato alla produzione di statistiche agricole che lo utilizzano come Frame di riferimento. Esso nasce dall'integrazione tra fonti amministrative specifiche del settore e non e fonti statistiche.

La configurazione di registro tematico deriva dalla caratteristica della popolazione di unità statistiche che esso deve contenere e quindi dalla definizione dell'unità statistica di riferimento: l'azienda agricola (UNECE, 2015).

La definizione utilizzata è la stessa del Censimento generale dell'agricoltura 2010, ossia "Unità tecnico-economica, costituita da terreni, anche in appezzamenti non contigui, ed eventualmente da impianti e attrezzature varie, in cui si attua, in via principale o secondaria, l'attività agricola e zootecnica ad opera di un conduttore – persona fisica, società, ente - che ne sopporta il rischio sia da solo, come conduttore coltivatore o conduttore con salariati e/o compartecipanti, sia in forma associata".

Il registro si compone sia di una parte anagrafica che di una specifica sulle colture, dove, per ciascuna particella di terreno, sono indicati i macrousi dei terreni.

I terreni sono identificati attraverso le informazioni catastali (Foglio, sezione, particella), pertanto, utilizzando tecniche di record linkage (Tuoto T., 2014), sarà possibile ricondurre le informazioni relative alle coltivazioni, ricavate dalla fotointerpretazione, alle unità statistiche presenti nel registro.

In tal modo sarà possibile:

- Validare le informazioni presenti nel registro;
- Integrare le informazioni presenti con i dettagli colturali;
- Integrare il database con le unità non presenti nel registro a causa della assenza nelle fonti amministrative (es. parte del comparto ortofrutticolo)

Le particelle che non trovano corrispondenza nel registro delle aziende agricole potranno essere riallineate attraverso un procedimento di filtraggio del database. Le soluzioni proposte per consolidare i dati mancanti sono le seguenti:



1. Interrogazione proprietà e conduzione attraverso il catasto e l’Agenzia del Territorio;
2. Attraverso il web mediante tecniche di parsing, per verificare la paternità aziendale della particella mancante (es. API Place Search di google maps).

#### 4. Conclusioni

L’argomento proposto si colloca in un contesto di ricerca ampio e complesso. Numerose risultano le variabili che entrano in gioco in questo tipo di applicazione e che possono inficiare la qualità dei risultati. È stato già discusso di come variazioni climatiche e colturali tra ambiti territoriali possano produrre effetti sulla pianta e sulla sua firma spettrale.

Dall’altra parte è in crescendo la possibilità di accesso ai Big Data ed ai dati spaziali, espressi sia nella quantità temporale che nella qualità della risoluzione; a questo si vanno ad aggiungere le crescenti capacità computazionali dei computer e la possibilità di impiego di Intelligenza Artificiale. Questi elementi accentuano la spinta di sempre più sofisticati e precisi algoritmi di classificazione.

Allo stato attuale la ricerca è in una fase di costruzione del modello e primi test relazionali tra dato reale e firma spettrale.

#### 5. Bibliografia

- Cai, Y., Guan, K., Peng, J., Wang, S., Seifert, C., Wardlow, B., & Li, Z. (2018). *A high-performance and in-season classification system of field-level crop types using time-series Landsat data and a machine learning approach*. *Remote Sensing of Environment*, 210, 35-47.
- Radermacher W. J., (2018). *Official statistics in the era of big data opportunities and threats*. *International Journal of Data Science and Analytics* n.6, pg. 225-231.
- Song, X. P., et al. (2017). National-scale soybean mapping and area estimation in the United States using medium resolution satellite imagery and field survey. *Remote sensing of environment*, 190, 383-395.
- Global Strategy to improve Agricultural and Rural Statistics (GSARS). 2017. *Handbook on Remote Sensing for Agricultural Statistics*. GSARS Handbook: Rome.
- King, L., et al. (2017). *A multi-resolution approach to national-scale cultivated area estimation of soybean*. *Remote Sensing of Environment*, 195, 13-29.
- Cavanillas J. M., Curry E, Wahlster W. (eds) (2016) *New Horizons for a Data-Driven Economy. A Roadmap for Usage and Exploitation of Big Data in Europe*. Springer Open.
- UNECE (2015). *Guidelines on statistical business registers*. United Nations: New York.
- Tuoto T. (2014). *Method: Fellegi-Sunter and Jaro Approach to Record Linkage*. *Memobust Handbook on Methodology of Modern Business Statistics*, Eurostat, 1-18.
- Wallgren A., Wallgren B. (2014). “Register-based Statistics Statistical Methods for Administrative Data”, John Wiley & Sons, Ltd Ed.
- Miller D., McCarthy J., Zakzeski A., (2009). *A Fresh Approach to Agricultural Statistics: Data Mining and Remote Sensing*. Section on Government Statistics – JSM 2009. <http://www.asasrms.org/Proceedings/y2009/Files/304336.pdf>
- Belin T.R., Rubin D.B. (1995). *A method for calibrating false-match rates in record linkage*. *Journal of the American Statistical Association*, 90, 694-707.

## ABSTRACT

In recent years, the quantity of digital data created, stored and processed in the world has grown exponentially. The demand for statistical information has never been so apparent.

The role of the official statistics is delivering high quality data concerning broad variety of topics. A very important domain of the official statistics is agriculture as a part of wider classification of land cover or use. Agriculture occupies large portions of land with a high temporal variability in terms of crops type and surfaces used. For this reason, it is necessary to develop statistical techniques able to collect and process spatial data at intervals close enough (for example each month). Recent innovative applications in this research field include remote sensing to construct land cover maps with the Cropland Data Layer program and data mining techniques to identify patterns and subgroups in large datasets to use in data collection and analytic operations. These techniques have not only made an impact on the process, quality, and timeliness of agricultural statistics but also opened statistical windows for other sectors.

The traditional data collection techniques give back the so-called primary data. The official statistics use also secondary data, data produced for purposes different from statistical but adapted at statistical purposes. Among them, the best known are the administrative data. Some phenomena escape this type of observation because it is not always possible to carry out operations of record linkage between the data and the territory or between two bases of different data. For this reason, National Statistics Institutes (NSI) are started the experimentation of the use of new data sources, richer in information and better distributed in geographical terms.

Among the secondary data, increasing interest derives from the use in agriculture of aerial and / or satellite images with different spatial resolution (high, medium or low), free or paid (public or private). The possible uses of this data can be multiple, including:

1. Positioning and sizing of crops;
2. Identifying the exact type of crop for a given parcel of land;
3. Estimate of the productions obtainable respect to climate conditions (rainfall, temperatures, etc.).

Therefore, the objective of this project is to test the use of aerial and/or satellite (as Big Data), found on the web, as new data sources in support to the official statistical production, in order to identify phenomena that escape traditional sources. Particularly the methodology proposed can be applied to the statistical registers in order to:

- Obtain a complete frame of statistical units;
- Integrate missing values;
- Validate the statistical registers information and/or units.

The experimentation of this proposal will be carried out within the fruit and vegetable productions to make as complete as possible the frame used to identify the universe of agricultural farms.