Giambattista Amati, Simone Angelini, Giuseppe Marcone
(Fondazione Ugo Bordoni)

Anna Caterina Carli, Giuseppe Pierri
(Istituto Superiore delle Comunicazioni e delle Tecnologie dell'Informazione)

Analisi di correlazione tra dati geo-localizzati e temporali da sorgenti informative diverse - Verso l'analisi dei dati IoT

Correlation analysis between geo-localized and temporal data from different information sources - Towards the analysis of IoT data

Sommario: *Si descrivono alcuni risultati scientifici che sono stati condotti nel progetto "Big Data & Open: metodologie e Tecnologie abilitanti" (BigDOT), dedicato alla definizione e alla validazione di una piattaforma di tipo Big Data per l'analisi di dati provenienti da sorgenti informative eterogenee tra loro.*

Gli obiettivi del progetto sono stati:

- *L'acquisizione e analisi di basi di dati disponibili in rete in modalità Open, oltre a quelle proprietarie, cioè quelle acquisite durante progetti FUB-ISCOM pregressi, al fine di testare e validare la piattaforma di tipo Big Data.*
- *L'individuazione delle tecnologie abilitanti di Big Data in ambiente di programmazione di tipo MapReduce, quali SparkR [1, 2] per le analisi statistiche di dati massivi e GraphX per l'analisi delle reti sociali.*
- *La valutazione della capacità elaborativa della piattaforma rispetto all'infrastruttura attualmente in dotazione del laboratorio Big Data e relativamente a flussi informativi geo-localizzati e temporali.*

Nel seguito si descrivono solo le attività di analisi statistiche su grandi Dataset e si riportano le tecniche e i risultati ottenuti mediante l'infrastruttura del laboratorio Big Data.

Abstract: *We describe some scientific results that have been carried out in the "Big Data & Open: methodologies and enabling technologies" (BigDOT) project, dedicated to the definition and validation of a Big Data platform for the analysis of data coming from heterogeneous information sources.*

The objectives of the project were:

- *Acquisition and analysis of datasets available on the Internet in Open mode, in addition to the proprietary ones, i.e. those acquired during previous FUB-ISCOM projects, in order to test and validate the Big Data platform.*
- *The identification of the enabling technologies of Big Data in a MapReduce programming environment, such as SparkR [1, 2] for statistical analysis of massive data and GraphX for the analysis of social networks.*

- *Evaluation of the processing capacity of the platform compared to the infrastructure currently provided by the Big Data laboratory and relative to geo-localized and temporal information flows.*

Below we describe only the statistical analysis activities on large Datasets and report the techniques and results obtained through the Big Data laboratory infrastructure.

1. Introduzione

L'*Internet of Things* (*IoT*) è un concetto che prevede di dotare il nostro mondo di sensori e di rispondere ai dati ricevuti da questi sensori in modo significativo e tempestivo. Si tratta di aggiungere dei sensori a tutte le cose in modo che sia possibile misurare, analizzare, visualizzare, prevedere e reagire all'ambiente intorno a queste cose. In altre parole, il concetto dell'*IoT* è quello di raccogliere da sensori dati geo-referenziati e distribuiti temporalmente con l'intento di produrre in tempo reale un'analisi e fornire in modo tempestivo qualche risposta. In questo articolo dimostriamo come definire una piattaforma di tipo *Big Data* adattabile al nuovo concetto dell'*IoT* e, un secondo obiettivo di questo lavoro è quello di presentare una infrastruttura e una metodologia per studiare la dinamica di fenomeni complessi. La piattaforma infatti è adattabile a trattare dinamiche complesse in diversi settori sociali o economici, quali quelli di tipo energetico, urbanistico, sociale, dei trasporti, meteorologico, sanitario, ingegneristico, della sicurezza informatica. In particolare la piattaforma comprende due funzionalità importanti:

- funzionalità di *Data Analytics* o *Data Mining* per l'analisi di grandi moli di dati geo-temporali;
- funzionalità per l'analisi della correlazione tra dati eterogenei.
- A tal scopo si sono acquisiti i seguenti dati e studiati i seguenti problemi:
- 745GB dei Big Data Challenge di Telecom, descritto in [3], contenenti informazioni relative al periodo novembre e dicembre 2013 della grid di Milano e Trento (dati mobili, elettrici, qualità dell'aria, meteo e dei social network).
- correlazione tra consumo energetico e traffico mobile.

Analisi predittive del consumo di energia elettrica si trovano in [4], dove viene applicato l'algoritmo Fast Fourier Transform alle serie storiche giornaliere del consumo energetico: vengono estratte le tre componenti (stagionale, rumore e tendenza) e diverse misure statistiche (medie, kurtosis, mediane, varianze di diverse statistiche) per caratterizzarne il consumo in aree diverse. Rimandiamo al lavoro di [5] per una rassegna esaustiva sullo stato dell'arte dell'analisi del traffico mobile, che si concentra su grafici sociali estratti da *dataset* di traffico mobile, e al lavoro di [6] per lo stato dell'arte sugli aspetti e i modelli

spazio-temporali. Più vicino al nostro approccio è il lavoro di [7] che correla tre tipologie di dataset spazio temporali: la popolazione, i dati Twitter e i dati della rete mobile, e il lavoro di [8] che contiene un'analisi della mobilità mediante pattern temporali estratti dai dati di Twitter.

Il progetto BigDOT, nell'ambito del quale è stato svolto questo studio, ha anche riguardato tutte le attività di aggiornamento del laboratorio Big Data di ISCTI realizzato nel corso degli anni di diversi progetti bilaterali tra FUB e ISCTI.

2. Gli Open Big Data raccolti

Nome File	Dimensione	Contenuto
MILANO		
milano-grid.zip	324.4 KB	Grid Milano
december/full.zip	2.5 GB	SMS, Call, Internet - MI
november/full.zip	2.5 GB	SMS, Call, Internet - MI
december/full.zip	1.1 GB	MI - Province
november/full.zip	1.2 GB	MI - Province
december/full.zip	45.6 GB	MI - MI
november/full.zip	47.6 GB	MI - MI
Milano_WeatherPhenomena.zip	153.6 KB	Milano Meteo
mi_meteo_legend.csv	2.2 KB	Milano Meteo
precipitation-milano.zip	96.1 KB	Precipitazioni
pollution-legend-mi.csv	3 KB	Qualità Aria - MI
pollution-mi.zip	153.3 KB	Qualità Aria - MI
TRENTO		
december/full.zip	1.6 GB	SMS, Call, Internet - TN
november/full.zip	1.3 GB	SMS, Call, Internet - TN
december/full.zip	596.4 MB	TN - Province TELCO
november/full.zip	442.3 MB	TN - Province TELCO
december/full.zip	43.1 GB	TN - TN TELCO
november/full.zip	36.8 GB	TN - TN TELCO
precipitation-trentino-data-availability.zip	20.8 KB	Precipitazioni TN
precipitation-trentino.zip	13.1 MB	Precipitazioni TN
air-2013.zip	344.9 KB	Qualità Aria - TN
line.zip	12.2 KB	SET, Electricity ENERGY
SET-dec-2013.zip	3.5 MB	SET, Electricity ENERGY
SET-nov-2013.zip	3.4 MB	SET, Electricity ENERGY

Figura 1. I dati del Big Data Challenge di telecom utilizzati dal progetto BigDot. Contengono circa 83 GB di dati compressi in formato zip, che corrispondono a circa 745 GB di dati complessivi non compressi.

All'inizio del 2014, Telecom Italia ha lanciato la prima edizione del Big Data Challenge, un concorso destinato a stimolare la creazione e lo sviluppo di idee tecnologiche innovative nel campo dei Big Data.

I *dataset* sono stati rilasciati solo per essere utilizzati dai partecipanti, ma dopo la fine del concorso sono liberamente disponibili e pertanto costituis cono un'ottima risorsa per effettuare analisi di *benchmarking*. In particolare è possibile effettuare analisi di correlazione per volumi di dati dell'ordine di diversi *gigabyte* che sono anche fortemente

eterogenei. Tipicamente le analisi di correlazione sono difficilmente trattabili se non in ambiente distribuito. Un tipico problema utile a misurare la capacità di elaborazione delle piattaforme di Data Analytics è ad esempio quello di correlare il traffico mobile e il consumo elettrico per fasce orarie e per aree geografiche.

Il progetto ha analizzato circa 83 GB di dati compressi in formato zip che corrispondono a circa 745GB di dati. Questi dati sono accessibili sul sito https://dandelion.eu/datamine/open-big-data/.

2.1. Dati CDR di Telecom

Tra il set dei dati c'è un surrogato dei Call Detail Record (CDR), che sono generati dalla rete cellulare di Telecom Italia nelle città di Milano e Trento. I dati CDR registrano le attività degli utenti ai fini della fatturazione e della gestione della rete. Ci sono molti tipi di CDR, ma nel *dataset* esistono solo i dati relativi alle seguenti attività:

- *SMS ricevuti*: un CDR viene generato ogni volta che un utente riceve un SMS
- *SMS inviato*: un CDR viene generato ogni volta che un utente invia un SMS
- *Chiamate in arrivo*: un CDR viene generato ogni volta che un utente riceve una chiamata
- *Chiamate in uscita*: CDR viene generato ogni volta che un utente invia una chiamata
- *Internet*: un CDR è generato ogni volta un utente avvia una connessione a internet, un utente termina una connessione ad internet o durante la stessa connessione uno dei seguenti limiti viene raggiunto: 15 minuti o 5 MB prodotti dall'ultimo CDR generato.

Aggregando questi dati CDR è stato creato l'insieme dei dati finali che fornisce il volume di SMS, di chiamate e attività di traffico Internet per intervalli di tempo. La misura è il livello di interazione tra utenti con la rete di telefonia mobile; per esempio maggiore è il numero di SMS inviati dagli utenti, maggiore è l'attività degli SMS inviati. Le misure di chiamata e di attività di SMS hanno la stessa scala e quindi sono comparabili tra loro; quelli relativi al traffico Internet invece hanno una scala indipendente.

2.2. Telecomunicazioni - SMS, chiamate, Internet - Milano e Trento

I dati contengono i seguenti campi:

- *id Square*: l'identificativo numerico del quadrato che fa parte della GRID di Milano; TIPO: numerico.
- *Intervallo di tempo*: l'inizio dell'intervallo di tempo espresso come il numero di millisecondi trascorsi dalla Unix Epoch dal 1 gennaio 1970 UTC. Il termine dell'intervallo di tempo può

essere ottenuto aggiungendo 600.000 millisecondi (10 minuti) a questo valore. TIPO: numerico.

- *Codice del paese*: il codice telefonico del paese di una nazione. TIPO: numerico
- *Attività di SMS-in*: l'attività in termini di SMS ricevuti all'interno della id Square, durante l'intervallo di tempo e inviato dalla nazione identificata dal codice del paese. TIPO: numerico
- *Attività di SMS-out*: l'attività in termini di SMS inviato all'interno della id Square, durante l'intervallo di tempo e ricevuto dalla nazione identificata dal codice del paese. TIPO: numerico
- *Attività Call-in*: l'attività in termini di chiamate ricevute all'interno del id Square, durante l'intervallo di tempo e rilasciato dalla nazione identificata dal codice del paese. TIPO: numerico
- *Attività Call-out* : l'attività in termini di chiamate emesse all'interno del id Square, durante l'intervallo di tempo e ricevuto dalla nazione identificata dal codice del paese. TIPO: numerico
- *Attività di traffico Internet*: l'attività in termini di traffico Internet effettuato all'interno della id Square, durante l'intervallo di tempo e dalla nazione degli utenti che effettuano il collegamento identificato dal codice del paese. TIPO: numerico

I file sono in formato TSV. Se nessuna attività è stata registrata per un campo specificato nello schema di cui sopra, allora il valore corrispondente è assente nel file.

Inoltre, se per una data combinazione di id e Square s, intervallo di tempo e codice paese non viene registrata nessuna attività, allora il record corrispondente non è presente nel set di dati.

2.3. Telecomunicazioni - Da Milano (Trento) alla Provincia

- *id Square*: l'id del quadrato della GRID di Milano (Trento); TIPO: numerico
- *Provincia*: il nome della provincia italiana; Tipo: STRING
- *Intervallo di tempo*: l'inizio dell'intervallo di tempo espresso come il numero di millisecondi trascorsi dalla Unix Epoch dal 1 gennaio 1970 UTC. Il termine dell'intervallo di tempo può essere ottenuto aggiungendo 600.000 millisecondi (10 minuti) a questo valore. TIPO: numerico
- *Interazione da id Square alla Provincia*: un valore che rappresenta l'interazione tra l'id Square e la Provincia. Questo valore è proporzionale al numero di chiamate scambiate tra chiamanti ubicati nella piazza id e ricevitori ubicati nella provincia. TIPO: numerico

- *Interazione da Provincia a id Square* : un valore che rappresenta l'interazione tra l'id piazza e la Provincia. Questo valore è proporzionale al numero di chiamate scambiate tra i chiamanti ubicati nella Provincia e ricevitori situati nella id Square. TIPO: numerico

I file sono in formato TSV. Se nessuna attività è stata registrata per un campo specificato nello schema di cui sopra, allora il valore corrispondente è assente nel file.

2.4. Telecommunications - Milano su Milano (Trento su Trento)

- *id1 Square*: l'id del quadrato della griglia di Milano (Trento) che è l'origine dell'interazione; TIPO: numerico
- *id2 Square*: l'id del quadrato della griglia di Milano (Trento) che è la destinazione dell'interazione; TIPO: numerico
- *Intervallo di tempo*: l'inizio dell'intervallo di tempo espresso come il numero di millisecondi trascorsi dalla Unix Epoch dal 1 gennaio 1970 UTC. Il termine dell'intervallo di tempo può essere ottenuto aggiungendo 600000 millisecondi (10 minuti) a questo valore. TIPO: numerico
- *Forza di Interazione Direzionale*: il valore che rappresenta la forza di interazione direzionale tra id1 Square e id2 Square. Questo valore è proporzionale al numero di chiamate scambiate tra chiamanti ubicati in id1 Square e ricevitori ubicati in id2 Square. TIPO: numerico

2.5. Stazione Meteo Dati

Legenda dei dati:

- *ID sensore*: l'id del sensore. TIPO: numerico
- *Il nome della strada del Sensore*: il nome della via in cui si trova il sensore identificato da l'ID del sensore. TIPO: alfanumerico
- *Latitudine del sensore* : la latitudine geografica specifica della posizione del sensore identificato dall'ID sensore. TIPO: numerico
- *Longitudine del sensore* : la longitudine geografica specifica della posizione del sensore identificato dall'ID sensore. TIPO: numerico
- *Tipo di sensore*: il tipo di sensore identificato dall'ID sensore. TIPO: alfanumerico
- *UOM*: l'unità di misura del valore registrato dal sensore identificato dall'ID sensore. TIPO: alfanumerico

2.6. Fenomeni Meteo

Questo insieme di dati contiene un file per ogni sensore. Il nome dei file ha il seguente formato MI_Meteo_<sensorID>.csv.

- *ID sensore*: l'id del sensore. TIPO: alfanumerico
- *Tempo istantaneo*: l'istante temporale della misura espressa come data e ora con i seguenti formati
- AAAA/MM/DD HH24: MI. TIPO: data.
- *Misura*: il valore dell'intensità dei fenomeni meteorologici misurata nell'istante tempo dal sensore ID. L'unità di misura (UM) del valore registrato dal sensore proposta è specificata nella Legenda dell'insieme di dati. TIPO: numerico
- La direzione del vento viene misurata in gradi aventi il nord, come piano di riferimento (il Nord viene specificato come valore di 0 o 360 gradi). Inoltre, i valori di misurazione direzione del vento possono assumere i seguenti valori speciali:
- 777: calma
- 7777: calma
- 888: variabile
- 8888: variabile

2.7. Precipitazioni

- *Timestamp*: il timestamp con il seguente formato: yyyymmddHHMM. TIPO: numerico
- *ID quadrante*: l'id del quadrante. TIPO: numerico (un valore compreso tra 1 e 4)
- *Intensità*: l'intensità della precipitazione. TIPO: numerico (un valore compreso tra 0 e 3)
- *Copertura*: la percentuale del quadrante che è coperta dalla precipitazione. TIPO: numerico (un valore compreso tra 0 e 100)
- *Tipo*: il tipo di precipitazione. TIPO: numerico (un valore compreso tra 0 e 2)

2.8. Qualità dell'aria - Milano (Trento)

Legenda dei dati:

- *ID sensore*: l'id del sensore. TIPO: numerico
- nome della strada del Sensore: il nome della via in cui si trova il sensore identificato dall'ID del sensore. TIPO: alfanumerico
- *latitudine del Sensore*: la latitudine geografica specifica della posizione del sensore identificato dall'ID sensore. TIPO: numerico

- *longitudine del Sensore*: la longitudine geografica specifica della posizione del sensore identificato dall'ID sensore. TIPO: numerico
- *Tipo di sensore*: il tipo di sensore identificato dall'ID sensore. TIPO: alfanumerico
- *UOM*: l'unità di misura del valore registrato dal sensore identificato dall'ID sensore. TIPO: alfanumerico
- *Formato ora istantaneo*: il formato che rappresenta il tipo di aggregazione temporale, in altre parole la granularità dei dati. Esso varia da sensore a sensore e può essere:
- *aggregazione al giorno* (AAAA/MM/GG). TIPO: alfanumerico.
- *aggregazione all'ora* (AAAA/MM/DD HH24: MI). TIPO: alfanumerico.

2.9. Set dei dati di Inquinamento

Questo insieme di dati contiene un file per ogni sensore. Il nome dei file ha il seguente formato MI_pollution_<sensorID>.csv.

- *ID sensore*: l'id del sensore. TIPO: alfanumerico
- *Istante di Tempo*: l'istante di tempo della misurazione espresso come data o data con l'ora secondo i seguenti formati:
- AAAA/MM/DD. TIPO: data.
- AAAA/MM/DD HH24: MI. TIPO: data.
- *Misura*: il valore dell'intensità dell'inquinamento misurata nell'istante tempo dal sensore ID. L'unità di misura (UM) del valore registrato dal sensore proposta è specificata nella Legenda dei dati . TIPO: numerico

2.10. Pulizia dei dati (cleansing)

I dati in generale possiedono molti campi non definiti o mancanti, alcune misure non sono conformi tra database eterogenei e vanno rese coerenti tra loro. Ad esempio si deve trasformare la colonna TimeStamp dei dataset dal formato data 'YYYY-MM-DD HH24:MI' al valore in milli-secondi tramite la funzione:

*as.numeric(as.POSIXct(Energia$TimeStamp))*1000*

Occorre effettuare una prima fase di pulizia e armonizzazione dei dati, che se non effettuata potrebbe produrre risultati non attendibili. Le operazioni da effettuare sono molto semplici e veloci essendo operazioni riconducibili a due primitive, *subset* e *join*.

Per la città di Trento da un database di 69.877.653 milioni di record, si ottiene un *dataset* della dimensione 33.596.149 milioni di record. Una volta che questo file viene letto come file distribuito (HDFS) è poi possibile utilizzare *SparkR* per elaborarne i dati.

2.11. Analisi dei Dati

L'analisi tra traffico mobile e consumo elettrico è stata effettuata utilizzando i dati del Big Data challenge di Telecom. Questi dati sono rilasciati in modalità open e sono accessibili sul sito https://dandelion.eu/datamine/open-big-data/. Lo studio presente è stato realizzato utilizzando i dati della Grid della provincia di Trento (https://dandelion.eu/datagems/SpazioDati/trentino-grid/description/).

2.12. Infrastruttura di calcolo

Al fine di integrare la piattaforma esistente con ecosistemi software di Big Data Analytics e di valutare l'efficacia dell'infrastruttura tecnologica hardware e software adottata in termini di volume di dati trattati e di scalabilità delle soluzioni, il progetto ha previsto l'acquisizione di un nuovo cluster di macchine. Le analisi sono state effettuate nel laboratorio Big Data mediante la seguente infrastruttura:

Il primo cluster è costituito da 8 macchine con:

- Processore Intel(R) Xeon(R) CPU E3-1225 v3 QuadCore@3.20Ghz
- un hard disk Seagate della capienza di 1TB e velocità di lettura a 5400rpm
- 8 GB di RAM (banco unico) e cache multi-livello L3 di 8MByte.

Il secondo cluster è costituito da 8 macchine Server CX 2550 Fujitsu, (CX 400 M1). Ogni server CX 2550 è configurato con:

- Processore : 2xXeon E5-2630v3 8C/16T 2.40 GHz
- Memoria: 2x 16 GB (1x16GB) 2 Rx4 DDR4-2133 R ECC
- Hard Disk: 1x HD SATA 6G 1TB 7,2K HOT PL 2,5” BC
- Solid State Drivers: 1x SSD SATA 6G 480GB Read Intensive 2,5” H-P

Dal punto di vista software, ogni macchina è equipaggiata con:

- sistema operativo Ubuntu 16.04 LTS Server.
- Java Virtual Machine Oracle v1.8.0

Le macchine sono collegate ad una LAN con schede di rete a 1 Gb/s.

2.13. Analisi del traffico mobile

I dati si riferiscono ai mesi di novembre e dicembre del 2013 relativamente all'area di Trento. Ci sono 166.571.878 record (TNMobile) relativi a tutte le comunicazioni mobili in entrata e uscita effettuate nella provincia di Trento, e al consumo effettuato mediante dispositivo mobile su internet.

- *Square id*: l'identificativo del pixel che fa parte del grid del Trentino; TIPO: numerico
- *Intervallo di tempo*: l'inizio dell'intervallo di tempo espresso come il numero di millisecondi trascorsi dalla Unix Epoch dal primo gennaio 1970 UTC. Il termine dell'intervallo di tempo può essere ottenuto aggiungendo 600.000 millisecondi (10 minuti) a questo valore. TIPO: numerico
- *Codice del paese*: il codice internazionale telefonico del paese di una nazione. A seconda dell'attività misurata, questo valore assume diversi significati. TIPO: numerico
- *Attività SMS-in*: l'attività in termini di SMS ricevuti all'interno del pixel, durante l'intervallo di tempo, e inviato dalla nazione identificata dal codice paese. TIPO: numerico
- *Attività di SMS-out*: l'attività in termini di SMS inviato dall'interno del pixel, durante l'intervallo di tempo, e ricevuto dalla nazione identificata dal codice paese. TIPO: numerico
- *Attività di Call-in*: l'attività in termini di chiamate ricevute all'interno del pixel, durante l'intervallo di tempo, e rilasciato dalla nazione identificata dal codice paese. TIPO: numerico
- *Attività di Call-out*: l'attività in termini di chiamate emesse all'interno del pixel, durante l'intervallo di tempo e ricevuto dalla nazione identificata dal codice paese. TIPO: numerico
- *Attività di traffico Internet*: l'attività in termini di traffico internet effettuato all'interno del pixel, durante l'intervallo di tempo e dalla nazione identificata dal codice paese. TIPO: numerico

Le misure di chiamata e di attività di SMS hanno la stessa scala (quindi sono comparabili); quelle relative al traffico internet invece non lo sono.

Il Traffico da Trento e su Trento ci forniscono informazioni importanti sulle abitudini e sulla stessa possibile composizione della popolazione residente. Comparando il traffico festivo e feriale su 8 giorni ciascuno, la Tavola di Figura 2 mostra un uso minore delle chiamate sia in uscita sia in entrata, compensato da un uso maggiore degli SMS sia in uscita sia in entrata, e da un uso maggiore di internet.

	TRAFFICO FERIALE		TRAFFICO FESTIVO
SMS IN	8,955,804	SMS IN	9,835,057
SMS OUT	5,414,090	SMS OUT	7,035,961
CALL IN	5,118,852	CALL IN	3,498,481
CALL OUT	5,559,843	CALL OUT	3,939,161
INTERNET	77,659,454	INTERNET	93,474,845

Figura 2.. Il traffico mobile durante i festivi a confronto con i feriali. Si usano di più gli SMS e internet per comunicare durante le feste.

La Tavola di Figura 3 invece mostra una asimmetria marcata (*skewness*) nella distribuzione dei dati secondo una power-law dovuta

alla concentrazione della popolazione in alcuni pixel. Da notare che questa asimmetria si attenua durante i giorni festivi probabilmente proprio per un'attenuazione della concentrazione della popolazione su alcuni pixel della griglia.

	TRAFFICO FERIALE			TRAFFICO FESTIVO		
	STAND.DEV.	MEDIA	MODA	STAND.DEV.	MEDIA	MODA
SMS IN	4.348662	0.949	0.156	2.827757	0.875	0.002
SMS OUT	4.594695	1.007	0.003	3.210749	1.017	0.002
CALL IN	3.964936	0.982	0.003	2.013244	0.632	0.002
CALL OUT	4.244486	0.964	0.003	2.151294	0.630	0.003
INTERNET	39.77267	9.789	0.198	29.78781	10.048	0.002

FIgura 3. Il traffico mobile durante i festivi a confronto con i feriali. Si usano di più gli SMS e internet per comunicare durante le feste.

Il traffico da e verso l'estero aumenta in modo significativo durante i festivi anche in modo assoluto e non solo in termini di valori in percentuale come evidenzia la Tavola di Figura 4. Lo scarto tra traffico in entrata e uscita potrebbe essere spiegato in parte da una popolazione residente straniera.

	FERIALE	FESTIVO
Italia	35.8%	30.1%
Fisso	24.1%	21.0%
Germania	5.3%	5.9%
Polonia	3.7%	4.7%
Romania	4.1%	4.3%
Rep Ceca	2.8%	3.6%
UK	1.4%	3.1%
Svizzera	1.0%	1.9%
Belgio	1.2%	1.8%
Francia	2.0%	1.8%
Olanda	0.7%	1.7%
Russia	0.3%	1.5%
USA	0.6%	1.5%

Figura 4. Traffico mobile feriale e festivo a confronto. Il campione è ottenuto selezionando un ugual numero di giorni feriali e festivi (gli otto giorni festivi di dicembre). Il traffico da e verso l'estero aumenta durante le festività. Il traffico su fisso e da e verso l'Italia scende dal 60% al 51%.

2.14. Traffico mobile Trento su Trento

Esiste inoltre un secondo DB ancora più corposo contenente una media di circa 67 milioni di rilevazioni giornaliere relative al dettaglio del traffico da Trento a Trento (TNtoTN). Il DB

TNtoTN infatti contiene un valore aggregato per intervallo di tempo (10 minuti) che definisce un indicatore di intensità di traffico da un pixel a pixel.

Ad esempio la Figura 5 rappresenta il traffico complessivo effettuato nella Grid di Trento.

Questi ultimi dati sono molto preziosi in quanto con altissima probabilità si riferiscono quasi al numero di comunicazioni effettuate da utente a utente per intervallo di tempo. Infatti, considerando le comunicazioni possibili da pixel a pixel (19.584.411) e gli intervalli temporali possibili in una giornata (144) si hanno circa 2.820.155.184 combinazioni possibili. Quelle rilevate sono state invece "solo" 67.770.922 e dunque una comunicazione ha la probabilità del 2.40% di essere effettuata in un dato intervallo e tra due dati pixel. Dunque la probabilità che due utenti possano effettuare nello stesso intervallo di tempo tra due stesi pixel è estremamente bassa. Quindi possiamo assumere che queste informazioni ci possano fornire il numero medio giornaliero di comunicazioni in uscita (tra Sms, chiamate) che è vicino al centinaio ipotizzando che sia presente la sola popolazione residente che è vicina alle 600,000 unità. Questa stima potrebbe essere molto più bassa se potessimo considerare la popolazione non residente presente sul territorio.

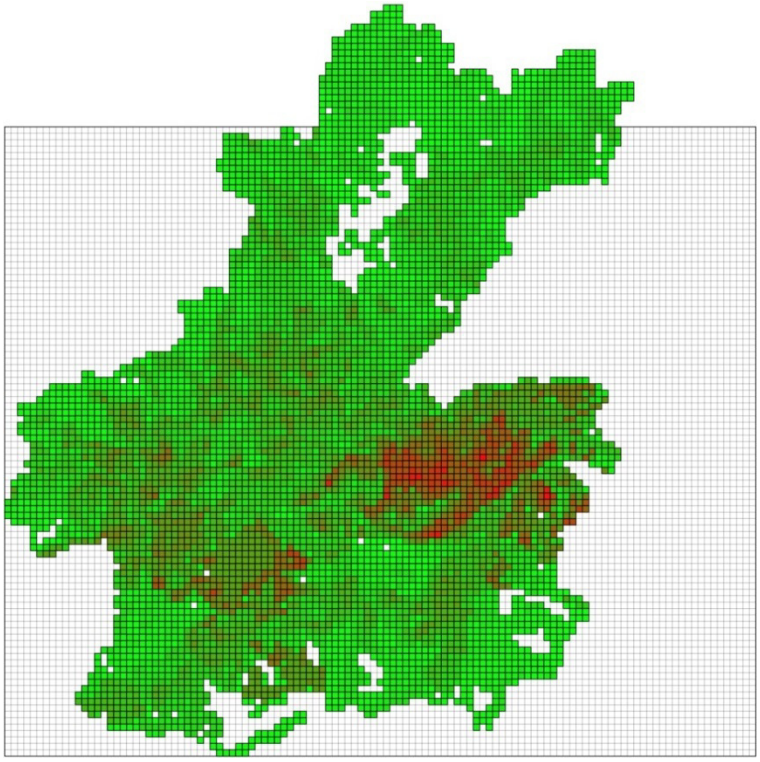


Figura 5: Il traffico mobile Trento su Trento. 67.770.992 record analizzati. La griglia è di 117×98 pixel per un totale di 6.247 pixel attivi

2.15. Analisi del consumo energetico

Purtroppo i dati rilasciati in modalità open forniscono una mappatura molto parziale tra traffico mobile e consumo elettrico (vedi Figura 6).

Si sono potuti utilizzare per la correlazione solo il 31% dei dati mobili.

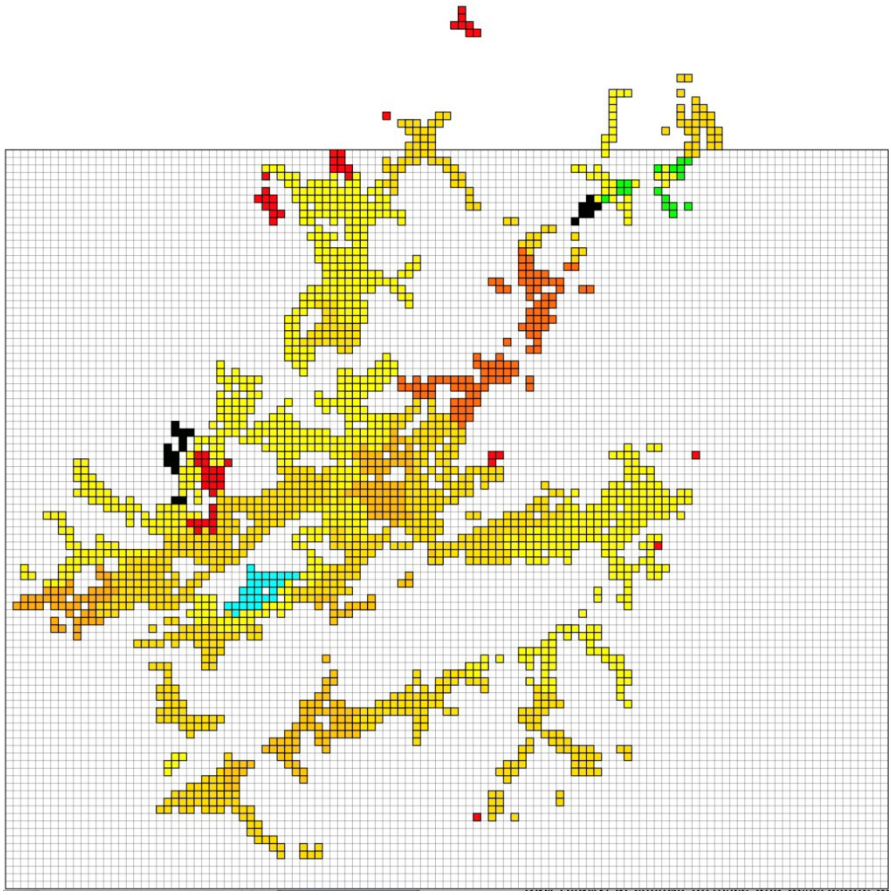


Figura 6: Traffico mobile feriale e festivo a confronto. Il campione è ottenuto selezionando un ugual numero di giorni feriali e festivi (gli otto giorni festivi di dicembre). Il traffico da e verso l'estero aumenta durante le festività. Il traffico su fisso e da e verso l'Italia scende dal 60% al 51%.

La SET gestisce quasi tutta la rete elettrica sul territorio trentino. La SET utilizza circa 180 linee di distribuzione primaria (linee di media tensione) per portare l'energia dalla rete nazionale e distribuirla tra gli utenti del Trentino. Le informazioni nel *dataset* riguardano il flusso di fornitura di corrente elettrica dalle linee di distribuzione, e contengono dettagli su come le linee di distribuzione sono distribuite sul territorio trentino.

Il set di dati è composto da sotto-insiemi di dati:

- *Dati relativi al cliente*: fornisce una descrizione delle linee di distribuzione primaria che servono il territorio trentino. La geometria delle linee non è esplicitamente esposta. La descrizione fisica delle linee è data in termini di sedi di clienti collegate alle linee di distribuzione primaria. Si noti che le sedi dei clienti spesso fornire energia a più di un cliente. In altre parole, possono fornire energia elettrica a un cliente (case unifamiliari), a molti clienti (condomini), alle attività commerciali e a strutture pubbliche.

- *Aggregazione spaziale*: i siti dei clienti per ogni linea sono raggruppati per pixel della GRID del Trentino. Ciò significa che, dato un quadrato della GRID del Trentino e una linea di distribuzione specifica viene solo registrato il numero di siti di clienti che rientrano in tale gruppo.
- *Aggregazione temporale*: la topologia della rete è considerata statica, pertanto l'aggregazione temporale è disponibile.
- *Dati di misurazione delle linee*: questo insieme di dati fornisce la quantità di corrente che fluisce attraverso le linee in istanti specifici.
- *Aggregazione spaziale*: non vi è alcuna aggregazione spaziale per questo insieme di dati.
- *Aggregazione temporale*: la corrente che fluisce attraverso le linee di distribuzione è stata registrata ogni 10 minuti.

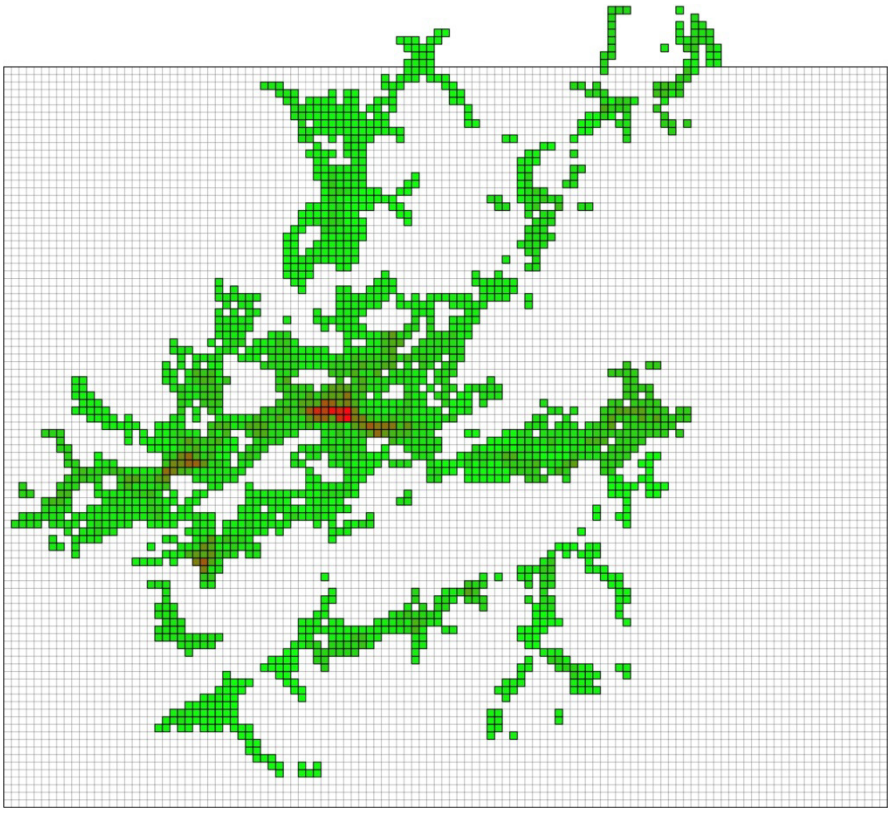


Figura 7: Consumo energia SET nel mese di novembre. La mappatura tra pixel e centraline è però solo parziale. Sono solo 200 le centraline associate a 2.575 pixel e quindi si sono potuti analizzare solo 10.928.880 record relativi al consumo energetico. Da una confronto qualitativo con la Figura 5 i circa 4.000 pixel mancanti sembrano riferirsi proprio alle zone dove esiste un maggior traffico mobile.

3. Esempio di Correlazione per Big Data: consumo energetico e traffico mobile

Abbiamo verificato il modello di regressione lineare e la correlazione di *Pearson* tra consumo elettrico e le cinque voci di consumo dal cartellino CDR del traffico mobile, ovvero consumo Internet da mobile, traffico Sms e chiamata sia in entrata sia in uscita (vedi Tavola 8).

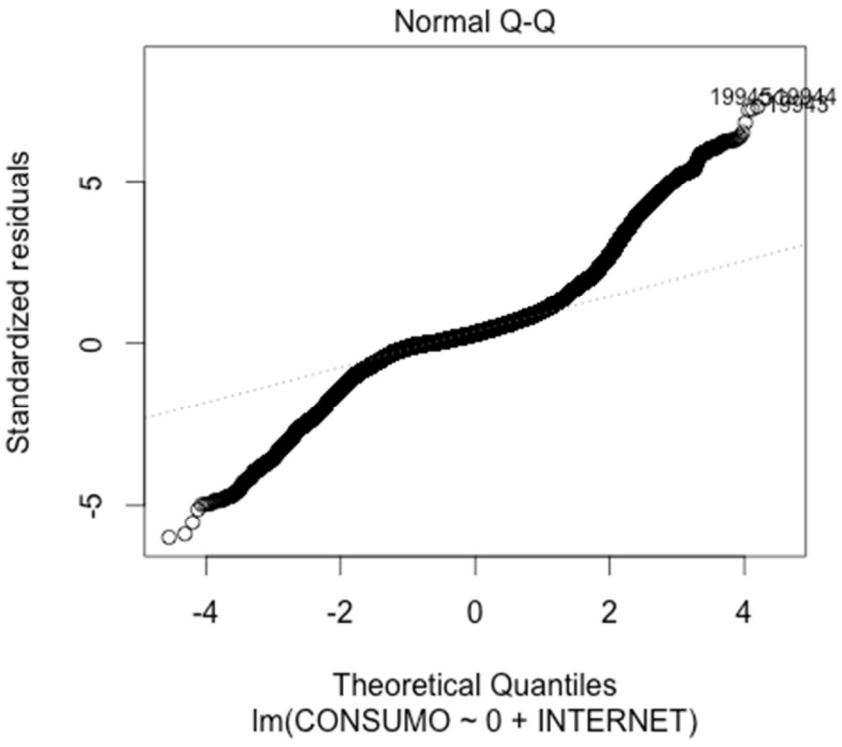
Il consumo Internet (insieme al traffico degli Sms in entrata) è il più correlato linearmente al consumo elettrico. Le chiamate sono correlate in modo meno significativo con il consumo elettrico. Probabilmente l'accesso alla rete da mobile presuppone una minore mobilità nel territorio dell'utente al momento dell'accesso alla rete.

Il valore R-quadratico di regressione è buono tra consumo elettrico e *Internet* è 0,433 con 185.471 gradi di libertà. Pertanto grazie a questo numero elevato di gradi di libertà, sebbene se la correlazione di *Pearson* sia moderata (0,360% Pearson) rimane comunque significativa. L'analisi dei residui mostra però una distribuzione asimmetrica dello scarto tra modello predittivo e dati osservati. Ciò significa che è necessario trovare modelli predittivi più adeguati al fitting dei dati rispetto al modello lineare.

*Figura 8: Correlazione di Pearson e R-squared della regressione lineare per BigData per la provincia di Trento. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*

	R-square	Pearson
FESTIVO	CONSUMO ELETT.	CONSUMO ELETT.
SMS IN	0.334***	0.323
SMS OUT	0.308***	0.291
CALL IN	0.307***	0.332
CALL OUT	0.301***	0.321
INTERNET	0.433***	0.360

Figura 9: Modello di regressione lineare tra consumo elettrico e consumo Internet da mobile. Il consumo Internet (insieme al traffico degli Sms in entrata) è il più correlato linearmente con il consumo elettrico. Il valore R-quadratico di regressione è 0,433 con 185.471 gradi di libertà. Pertanto grazie a questo numero elevato di gradi di libertà la correlazione è moderata ma significativa (0,38% Pearson). L'analisi dei residui mostra però una distribuzione asimmetrica dello scarto tra modello predittivo e dati osservati. Ciò significa che è necessario trovare modelli predittivi più adeguati al fitting dei dati rispetto al modello lineare.



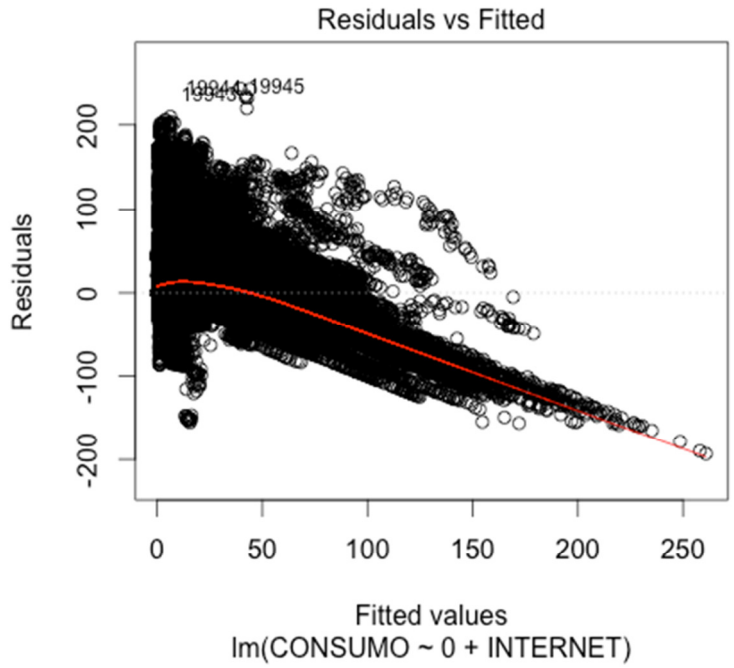


Figura 10: Distribuzione dei residui con il modello di regressione lineare tra consumo elettrico e consumo Internet da mobile. La distribuzione dei residui non è casuale ma ha una tendenza.

Esiste una maggiore correlazione trai dati se si considerano le fasce orarie come si vede dalla Figura 11.

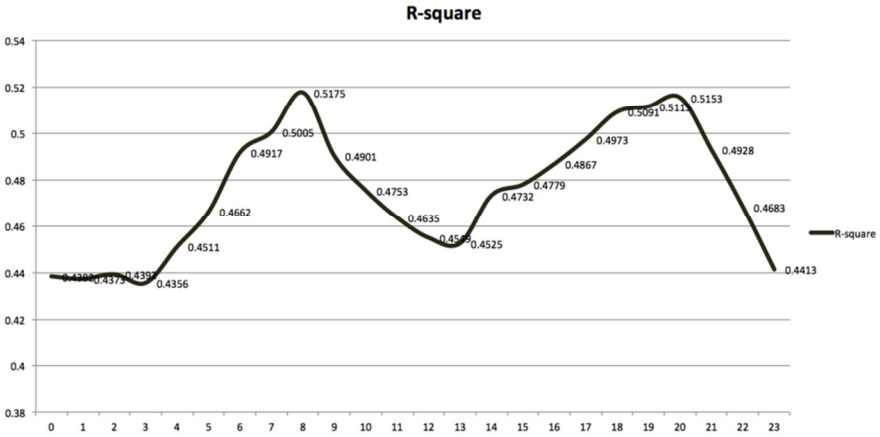


Figura 11: Se si suddividono i dati per fasce orarie si aumenta la correlazione lineare (52% circa) tra consumi elettrici e traffico mobile. In particolare il picco si ha alle ore 8:00 e 20:00 (la correlazione di Pearson è 0,38), cioè nelle ore di minore mobilità delle persone.

4. Conclusioni

I dati open del *Big Data Challenge*, sebbene in forma molto aggregata e con informazioni sulla mappatura molto sparsa tra dati elettrici e mobili e poco granulare rispetto al numero di centraline elettriche e pixel sui dati mobili, forniscono già moltissime informazioni relativamente alla popolazione residente e sulle proprie abitudini. Inoltre i dati dimostrano una correlazione moderata tra consumi elettrici e consumi mobili. Restano ancora da effettuare le seguenti analisi,

previo uno studio di stress test sulla scalabilità sulla nostra piattaforma di calcolo distribuito:

Il traffico da Trento su Trento fornisce un incredibile fonte di informazione sulla mobilità sul territorio della popolazione residente. Si potrebbero analizzare un insieme di circa 70 milioni di record circa al giorno (per una popolazione residente di circa 600.000 persone) relativamente a un grafo contenente 10^10 archi potenziali.

A ogni arco si può associare un peso dovuto al traffico mobile in entrata e in uscita (o meglio solo al numero di interazioni effettuate) che se studiato per fasce orarie può fornire un analisi più precisa della mobilità interna alla *grid* interessata e fornire una stima più precisa della popolazione *stanziale* in ciascun pixel indipendentemente dalla fascia di tempo o relativamente alla fascia di tempo.

In base a questa stima aggregata si potrebbe migliorare la correlazione tra consumo elettrico e mobile.

In effetti una possibile congettura sulla migliore capacità predittiva del traffico internet rispetto agli altri tipi di traffico mobile è che sia un'attività maggiormente di tipo stanziale rispetto a quella delle altre.

Bibliografia

[1] Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., and Stoica, I. (2010). Spark: Cluster computing with working sets. HotCloud, 10(10- 10):95.

[2] Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M. J., Shenker, S., and Stoica, I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In Proceedings of the 9th USENIX Conference on Networked Systems Design and Implementation, NSDI'12, pages 2–2, Berkeley, CA, USA. USENIX Association.

[3] Barlacchi, G., De Nadai, M., Larcher, R., Casella, A., Chitic, C., Torrisi, G., Antonelli, F., Vespignani, A., Pentland, A., and Lepri, B. (2015). A multi-source dataset of urban life in the city of Milan and the province of Trentino. Scientific Data, 2:150055 EP -.

[4] Bogomolov, A., Lepri, B., Larcher, R., Antonelli, F., Pianesi, F., and Pentland, A. (2016). Energy consumption prediction using people dynamics derived from cellular network data. EPJ Data Science, 5(1):13.

[5] Naboulsi, D., Fiore, M., Ribot, S., and Stanica, R. (2016). Large-scale mobile traffic analysis: A survey. IEEE Communications Surveys & Tutorials, 18(1):124–161.

[6] Blondel, V. D., Decuyper, A., and Krings, G. (2015). A survey of results on mobile phone datasets analysis. EPJ Data Science, 4(1):10.

[7] Lenormand, M., Picornell, M., Cantù-Ros, O. G., Tugores, A., Louail, T., Herranz, R., Barthelemy, M., Frìas-Martìnez, E., and Ramasco, J. J. (2014). Cross-checking different sources of mobility information. PLOS ONE, 9(8):1–10.

[8] Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. (2014). Geo-located Twitter as proxy for global mobility patterns. Cartography and Geographic Information Science, 41(3):260–271. PMID: 27019645.

[9] Deville, P., Linard, C., Martin, S., Gilbert, M., Stevens, F. R., Gaughan, A. E., Blondel, V. D., and Tatem, A. J. (2014). Dynamic population mapping using mobile phone data. Proceedings of the National Academy of Sciences, 111(45):15888–15893.