1° Congresso Nazionale — AISAM Associazione Italiana di Scienze dell'Atmosfera e Meteorologia — 10 -13 settembre 2018

"TIPI di TESTO per TIPI di TEMPO"

Analisi testuale dei bollettini meteo a supporto della classificazione dei tipi di circolazione

*Valentina Grasso[1,2] *, Alfonso Crisci[1], Giulio Betti[1,2], Bernardo Gozzini[2], Gianni Messeri[1,2], Roberto Vallorani[1,2], Federica Zabini[1,2]*

1) CNR IBIMET, Via G. Caproni 8, Firenze, 50138 - 2) Consorzio LaMMA, Via Madonna del Piano 10, Sesto Fiorentino (FI), 50019

** Valentina Grasso, Tel: +055 4483 068, Email: grasso@lamma.rete.toscana.it*

L'obiettivo di questo lavoro è quello di esplorare nuovi strumenti di indagine nel campo del trattamento dell'informazione meteo-climatica, partendo da diverse tipologie di dati come quelli di natura testuale. Il Consorzio LaMMA (Regione Toscana-CNR), servizio meteorologico della Regione Toscana, ha elaborato due classificazioni in tipi di circolazione atmosferica per l'Italia, una ottimizzata per la temperatura [SAN09] e una per la precipitazione [PCT09]. La classificazione qui esaminata è la PCT09, che è stata realizzata applicando il metodo di analisi delle componenti principali ai dati di pressione media a livello del mare (Reanalysis 2 NCEP/NCAR 1981-2010) [1]. In questa analisi sono state analizzate le informazioni relative alla precipitazione, mettendo in relazione i 9 tipi di circolazione con le frequenze dei termini appartenenti al dominio semantico della precipitazione contenuti nel bollettini meteo giornalieri. La disponibilità di poter associare un inventario giornaliero del pattern di circolazione dell'atmosfera ed uno storico di prodotti testuali generati dall'attività dei previsori rappresenta un'opportunità per valutare il grado di coerenza fra prodotti meteo di natura diversa ma orientati verso l'unico scopo di descrivere in modo efficace lo stato del tempo. Questa prima analisi ha mostrato interessanti potenzialità evidenziando bene come tipi di tempo opposti, ciclonici/anticiclonici, diano effettivamente luogo a corpus testuali sensibilmente diversi.

[1] Vallorani et al (2017) ,Circulation type classifications for temperature and precipitation stratification in Italy,International journal of climatology https://doi.org/10.1002/joc.5219

Dati e metodi

In questo studio sono state analizzate le informazioni relative alla precipitazione usando la classificazione PCT09 per i pattern pluviometrici. L'analisi testuale ha riguardato la descrizione relativa allo "stato del cielo" contenuta nei bollettini meteo emessi dal LaMMA nel periodo 01/01/2011 - 31/01/2016.
I testi relativi a 1854 bollettini sono stati suddivisi in 9 *corpora* raggruppando i bollettini emessi nei giorni aventi lo stesso tipo di tempo (PCT09). Tra le variabili testuali esaminate si presentano quelle relative a: lunghezza dei testi; ricchezza di vocabolario; frequenza dei termini connessi al dominio semantico della precipitazione; caratteri di unicità di ogni corpus. Sono state poi calcolate le frequenze di occorrenza di due pattern testuali polari rispetto al carattere spaziale delle precipitazioni [diffus*/isolat*] e della loro intensità [deboli/intense] e sui termini legati alle precipitazioni solide [grandin*,nev*]. Le frequenze relative sono state rappresentate sia su base complessiva che su base stagionale (consultabile online tramite i relativi QRcode). Le analisi sono state condotte sulla piattaforma https://voyant-tools.org/.

Evidenze

Il corpus totale dei testi analizzati contiene 59712 parole totali, con 845 forme di parole uniche. Le parole più frequenti sono: nuvoloso (2152); zone (1234); poco (1011); pomeriggio (845); sereno (719); mattinata (665); precipitazioni (635); addensamenti (573); nuvolosità (558); serata (547). Come mostra la tabella, il volume testuale relativo allo stato del cielo varia in funzione del tipo di circolazione, così come la ricchezza di termini usati, fornendo una misura indiretta della complessità della situazione meteo da prevedere. I grafici delle frequenze lessicali mostrano abbastanza chiaramente come esista una coerenza fra la natura dei pattern pluviometrici associati alle tipologie di circolazione e la frequenza di termini specifici utilizzati nei testi del bollettino. I tipi di circolazione "ciclonica/anticiclonica" confermano e spiegano le distribuzioni delle frequenze terminologiche sia a livello complessivo (annuale), che, con ulteriore accuratezza, a livello della singola stagione.

Conclusioni

L'analisi dei termini connessi alla precipitazione mostra come la classificazione in tipi di circolazione delinei frequenze molto diverse dei termini all'interno dei bollettini. Molto frequenti nei corpus relativi a circolazioni cicloniche, i termini connessi alla precipitazione riducono drasticamente la loro frequenza negli altri corpus. Così come varia la frequenza di caratteri polari quali "deboli/intense" o "isolate/diffuse" in circolazioni cicloniche e anticicloniche. L'analisi testuale si dimostra un metodo capace di fornire ulteriore elementi a supporto della classificazione dei tipi di circolazione finalizzata alla descrizione climatica di un dominio spaziale limitato come quello della Toscana.

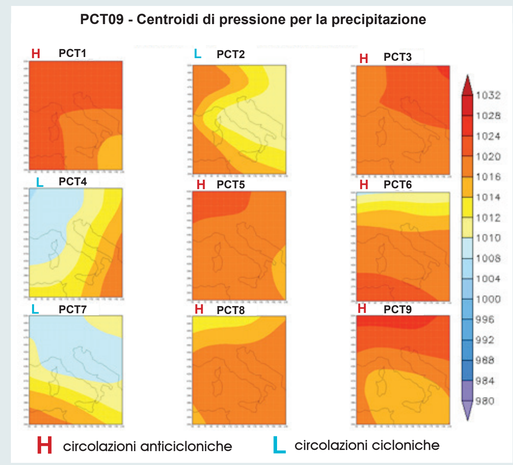
dati testuali

1854 bollettini	9 corpora
59712 parole	845 parole uniche

Tipi di circolazione	N. bollettini corpus	Parole del corpus	Densità del vocabolario	Parole connesse alla precipitazione *
PCT1	284	6702	0.067	133
PCT2	249	9112	0.057	478
PCT3	183	5073	0.078	152
PCT4	162	6338	0.069	469
PCT5	177	5471	0.080	217
PCT6	245	8666	0.057	346
PCT7	90	3540	0.110	198
PCT8	158	4545	0.081	130
PCT9	306	10265	0.053	516

Sono state analizzate le parole relative alle precipitazone con i seguenti prefissi: piogg; rovesci*; temporal*; precipitazion*; nev*; grandin*

dati climatici



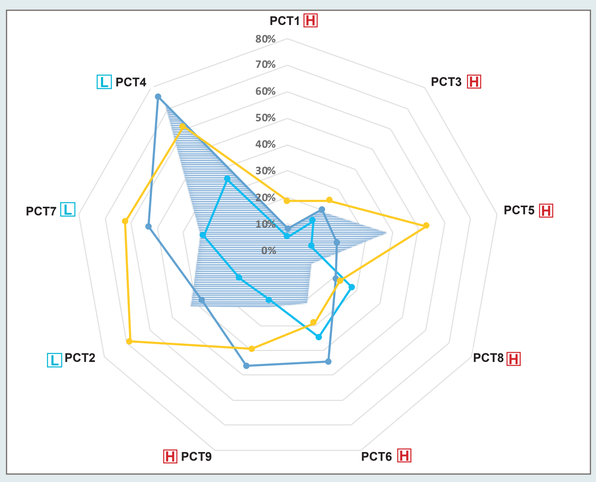
I 9 tipi di circolazione sono stati etichettati in "cicloniche" e "anticicloniche" considerando il dominio spaziale toscano.

Scan me

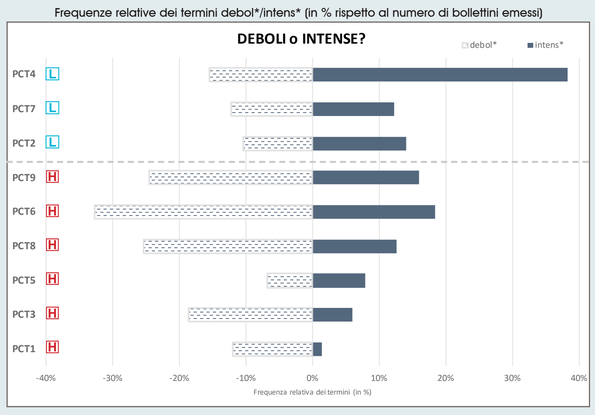
Pattern di pioggia e circolazioni

Analisi della frequenza dei termini relativi alla pioggia per ciascun tipo di circolazione

temporal* // piogg* // precipitazion* // rovesci*

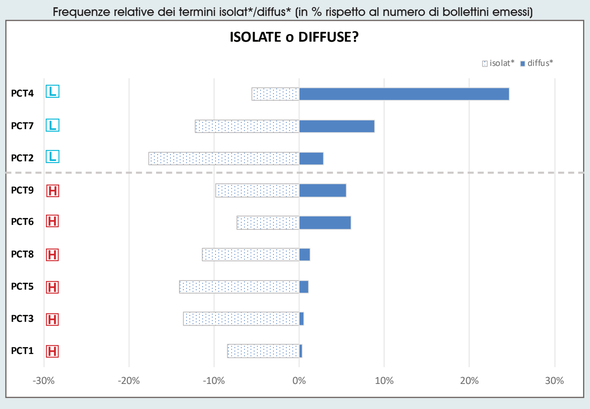


Frequenze relative dei termini debol*/intens* (in % rispetto al numero di bollettini emessi)



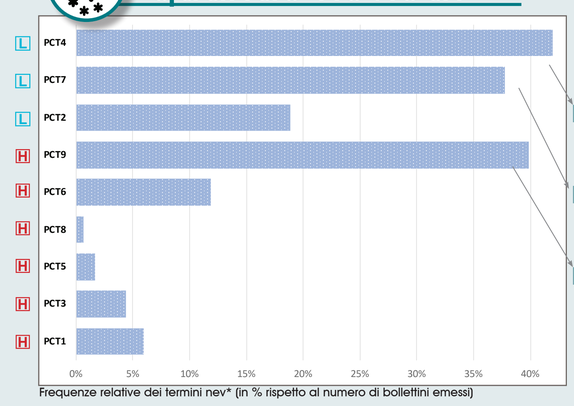
deboli/intense

Frequenze relative dei termini isolat*/diffus* (in % rispetto al numero di bollettini emessi)



isolate/diffuse

quando nevica?



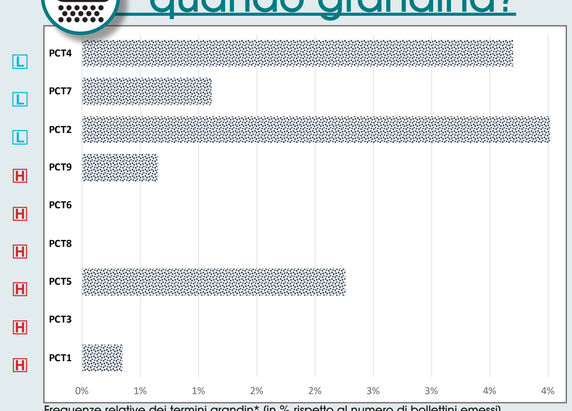
Frequenze relative dei termini nev* (in % rispetto al numero di bollettini emessi)

L'analisi testuale dei bollettini fa emergere ulteriori specificità: le parole correlate (r) al termine "nevicate" variano in funzione del tipo di circolazione:

PCT 4	
700 m	r= 0.79*
1200 m	r= 0.61
Appe.ino	r= 0.59
PCT 7	
1100 m	r= 0.76
1200 m	r= 0.72
intensità	r= 0.59
PCT 9	
Arezzo	r= 0.78
Est	r= 0.77
deboli	r= 0.76

dove "r" è il valore di term correlation all'interno del corpus (analisi fatta con Voyant Tools).

quando grandina?



Frequenze relative dei termini grandin* (in % rispetto al numero di bollettini emessi)

WEB REPOSITORY e DATI

I dati e le risorse del lavoro sono disponibili in repository sull'account GitHub ConsorzioLamma.

Scan me

Consulta tutti i diversi grafici di frequenza suddivisi per stagione.

Scan me

PER CITARE IL LAVORO: *Grasso, Valentina; Crisci, Alfonso; Betti, Giulio; Gozzini, Bernardo; Messeri, Gianni; Vallorani, Roberto; & Zabini, Federica. (2018). Analisi testuale dei bollettini meteo a supporto della classificazione dei tipi di circolazione (Version 0.1). Zenodo. http://doi.org/10.5281/zenodo.1409505*