Università degli Studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Triennale in

Statistica per le Tecnologie e le Scienze

RELAZIONE FINALE

Due fonti per le stesse informazioni:
confronti e integrazione di reti di stazioni meteorologiche

Relatore: Prof. Livio Finos
Dipartimento di Psicologia dello Sviluppo e della Socializzazione

Correlatori:
Prof. Bruno Scarpa
Dipartimento di Scienze Statistiche

Dott. Luca Menini
Servizio Informatica e Reti - ARPA Veneto

Laureando: Sara Ceschin
Matricola N.: 1101119

Anno Accademico 2016/2017

Ringrazio il Servizio Informatica e Reti di ARPA Veneto, in particolare il Dott. Luca Menini e Giovanna Ziroldo per avermi dato l'opportunità di svolgere lo stage all'interno dei loro uffici. Ringrazio nuovamente ARPAV e MeteoNetwork per i dati messi gentilmente a disposizione. Ringrazio inoltre i professori Livio Finos e Bruno Scarpa per avermi aiutato in questo progetto di tesi. Ringrazio infine tutti coloro che mi sono sempre stati affianco nel mio percorso universitario.

Indice

Introduzione

Lo studio qui riportato si presenta come un'analisi dei dati riguardanti la temperatura dell'aria a due metri dal suolo. Essa è uno dei principali parametri meteorologici superficiali che vengono rilevati e monitorati dalle Agenzie Regionali per la Prevenzione e la Protezione Ambientale. Nel territorio del Veneto se ne occupa ARPAV, per la quale ho avuto l'opportunità di svolgere uno stage che ha portato allo sviluppo di questo progetto di tesi.
Nell'ambito meteorologico, però, non vengono coinvolti solamente gli enti pubblici regionali. Esiste infatti un'associazione che dal 2002 sta sviluppando una rete di rilevazione a livello nazionale. Si tratta di MeteoNetwork. Essa si appoggia ad utenti privati che possiedono una centralina che rispetti le normative dell'Organizzazione Meteorologica Mondiale per quanto riguarda il posizionamento della stazione e dei sensori. [Wikb] L'associazione, senza scopo di lucro, è aperta a tutti. [MNW]
Scopo dello studio è dunque cercare di capire, per quanto riguarda il territorio della regione Veneto, se il circuito di stazioni di MeteoNetwork abbia innanzitutto una qualità comparabile a quella della rete ufficiale ARPAV e, in caso affermativo, vedere se esso possa costituire informazione aggiuntiva per l'Agenzia Regionale. Il mio lavoro si propone, attraverso attente analisi descrittive e la stima di alcuni modelli lineari e non, di giungere ad una prima risposta per questi quesiti
Il progetto qui proposto prende spunto da uno studio precedentemente effettuato da ARPAE per la Regione Emilia-Romagna. [SAP16]

Il software utilizzato per tutte le analisi effettuate è R. [R C13]

I dati

I dati sono stati forniti da ARPAV e da MeteoNetwork. Essi riguardano la temperatura dell'aria a due metri dal suolo misurata in gradi Celsius. Il periodo di rilevazione è compreso tra gennaio 2013 e dicembre 2015.
I data set sono composti da otto variabili quali il codice ed il nome della stazione, le coordinate spaziali longitudine e latitudine, le informazioni temporali data e ora, il tipo di sensore (temperatura) ed il valore della misurazione. Le osservazioni infatti sono a cadenza oraria.
Come si può vedere in Figura 1, le stazioni sono collocate in tutto il territorio del Veneto, con la differenza che le stazioni MeteoNetwork non sono distribuite uniformemente nella regione in quanto sono centraline posizionate da privati e si concentrano maggiormente dove la densità di popolazione è più elevata. Inoltre le stazioni ARPAV sono 169 mentre quelle di MeteoNetwork sono rispettivamente 114, 120, 123 nel 2013, 2014 e 2015, a sostegno del fatto che è una rete in espansione.

Inizialmente i dati sono stati mantenuti divisi tra ARPAV e MeteoNetwork ed inoltre sono stati separati per anno per questioni sia computazionali, in quanto si tratta di data set di dimensioni notevoli, sia logiche per poter vedere se ci fossero stati cambiamenti nel corso del tempo.
Dopo le prime analisi di controllo dei data set, si evidenziava una forte asimmetria negativa nella distribuzione della temperatura rilevata da ARPAV, cosa che non trovava appoggio nei dati di MeteoNetwork. Controllando meglio il data set si è riscontrato che i dati che portavano all'asimmetria provenivano dalla stazione s501. Si tratta della Dolina Campoluzzo (Monte Lozze) che è una stazione particolare in quanto posizionata in maniera strategica su una particolare conformazione del ter-

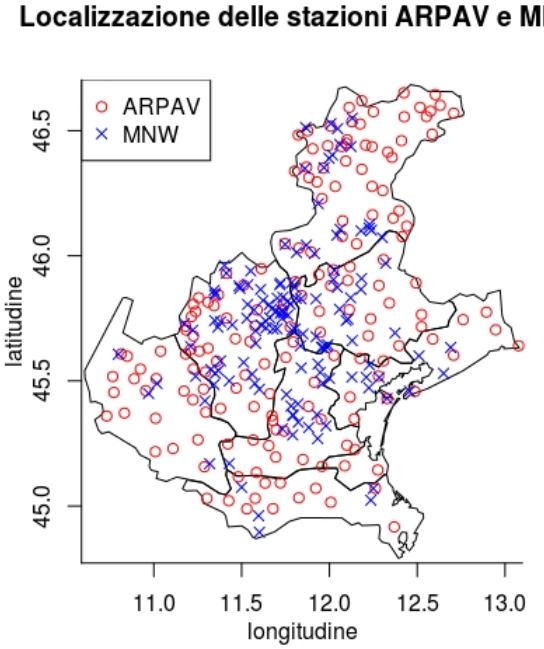


Figura 1: Mappa del Veneto con le posizioni delle stazioni ARPAV e MeteoNetwork con il sensore per la temperatura.

reno in modo tale da rilevare le temperature più basse del Veneto. Per questo motivo non può essere confrontata con altre stazioni meteo e di conseguenza è stata rimossa definitivamente dai data set utilizzati per le analisi successive.

Come si nota dai grafici in Figura 2, nelle due reti di rilevazione la temperatura presenta lo stesso andamento nel tempo. Ciò costituisce una base solida da cui far partire un confronto tra l'Agenzia Regionale del Veneto e l'associazione MeteoNetwork.

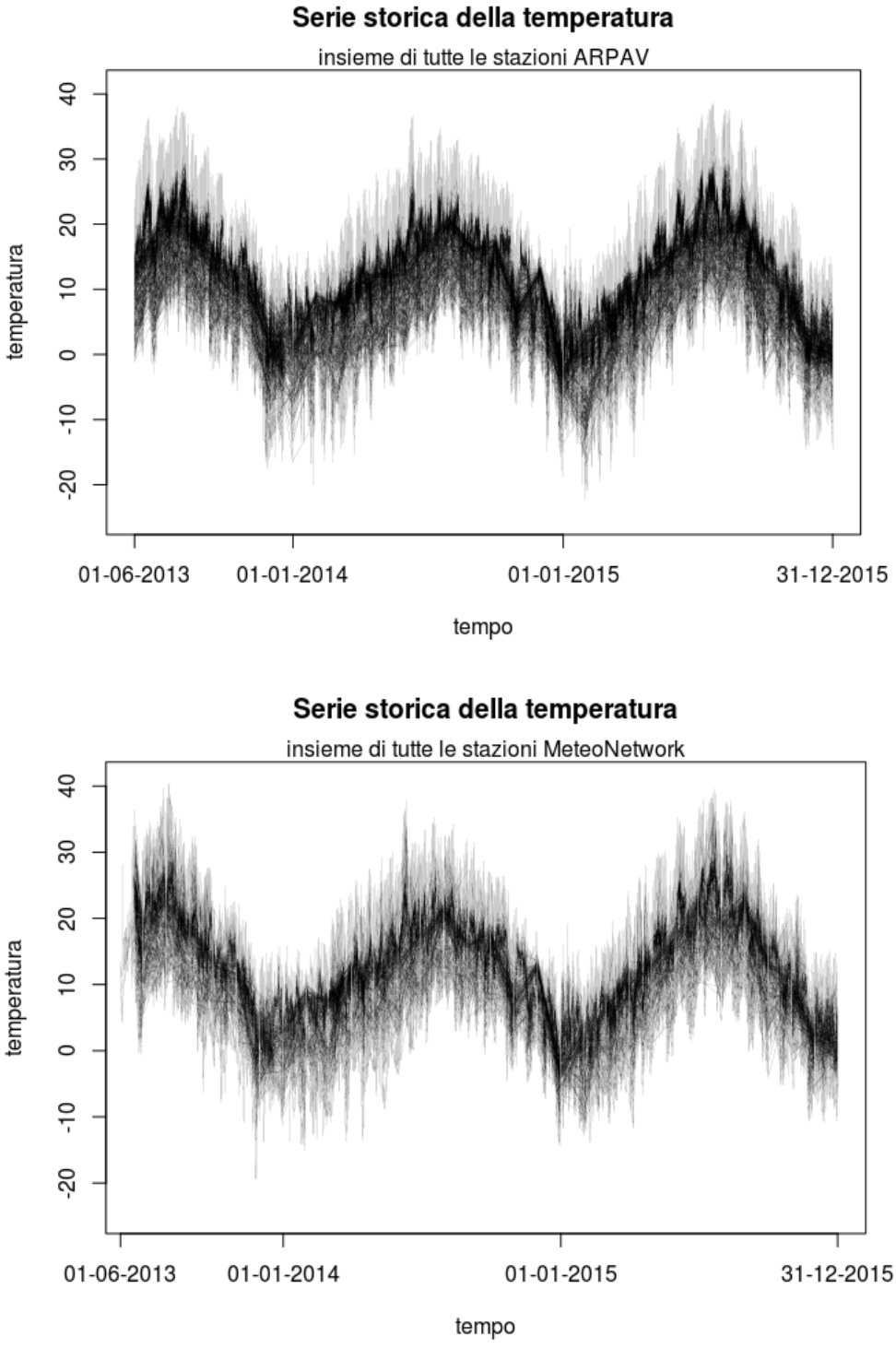


Figura 2: Serie storica della temperatura per le stazioni ARPAV e MeteoNetwork.

Esplorazione dei dati

Per iniziare a lavorare sui dati si è subito proceduto a controllare se ci fossero delle anomalie che avrebbero potuto distorcere i risultati finali. La procedura è stata semiautomatica in quanto la dimensione dei data set, nonostante fossero stati divisi, è grande. Si va da un minimo di circa 400 mila osservazioni ad un massimo di circa 800 mila all'anno per MeteoNetwork. Per quanto riguarda i data set ARPAV, infatti, non è stato effettuato alcun controllo in quanto i dati trasmessi dalle stazioni vengono sottoposti a severi controlli e validati prima di essere inseriti nel sistema, mentre i dati invalidati vengono segnati come dati mancanti. [ARP16] È stato scelto perciò di usarli come base per effettuare un controllo sui data set di MeteoNetwork.
Ad una prima vista si nota subito che sono presenti parecchi dati mancanti, non quantificabili, in quanto si intuisce che mancano delle rilevazioni in alcune ore della giornata.
Si è deciso di procedere per mese, a partire da giugno 2013 poiché gennaio e febbraio non erano presenti nel data set MeteoNetwork mentre marzo, aprile e maggio presentavano ben poche osservazioni.

Valori anomali

Sono stati presi i dati per ogni mese sia dai data set MeteoNetwork sia da quelli ARPAV e sono stati dapprima calcolati i range delle osservazioni,

ciò per vedere se MeteoNetwork presentasse dei valori molto più bassi o alti rispetto ad ARPAV. Si è supportata l'analisi descrittiva con vari istogrammi, come quelli in Figura 3, per cercare di individuare la presenza di valori anomali. In alcuni casi non sono state notate osservazioni particolari. In altri sono apparse, focalizzando l'istogramma sulle osservazioni laterali meno frequenti, degli outliers, a volte di un singolo valore, a volte di gruppi di valori isolati dalla distribuzione principale. Queste osservazioni sono state prese in esame e sono state formulate varie ipotesi, tra le quali il fatto che possa trattarsi di un'osservazione anomala non compatibile con le temperature delle ore limitrofe, oppure di un malfunzionamento della stazione o di temperature che non si sono convertite in gradi Celsius in quanto nell'originario data set di MeteoNetwork erano espresse in gradi Fahrenheit. Dove erano presenti osservazioni valide vicine nel tempo si è proceduto a modificare il dato con una media o convertendolo in Celsius. Nelle altre situazioni i valori ritenuti anomali sono stati eliminati.

Valori zero

Oltre a possibili outliers sono state analizzate le osservazioni che presentavano il valore 0, in quanto spesso risultava molto più frequente nei data set di MeteoNetwork rispetto ai data set di ARPAV, come si nota nell'esempio in Figura 4. Ad un primo controllo la maggior parte delle osservazioni sembravano indicare un malfunzionamento nel sistema di raccolta dati di MeteoNetwork dunque si è pensato che fossero stati utilizzati per evidenziare i dati mancanti. Si è deciso di imputare una media per alcuni valori mancanti qualora ci fossero abbastanza osservazioni vicine nel tempo. Nei casi in cui mancavano blocchi di dati invece, sono state rimosse le osservazioni anomale o, nei casi peggiori, è stata rimossa per quel mese l'intera stazione. Infatti sarebbe stata da considerare inaffidabile per le future analisi. L'eliminazione di molti dati non ha creato problemi in quanto si avevano a disposizione numerose osservazioni.

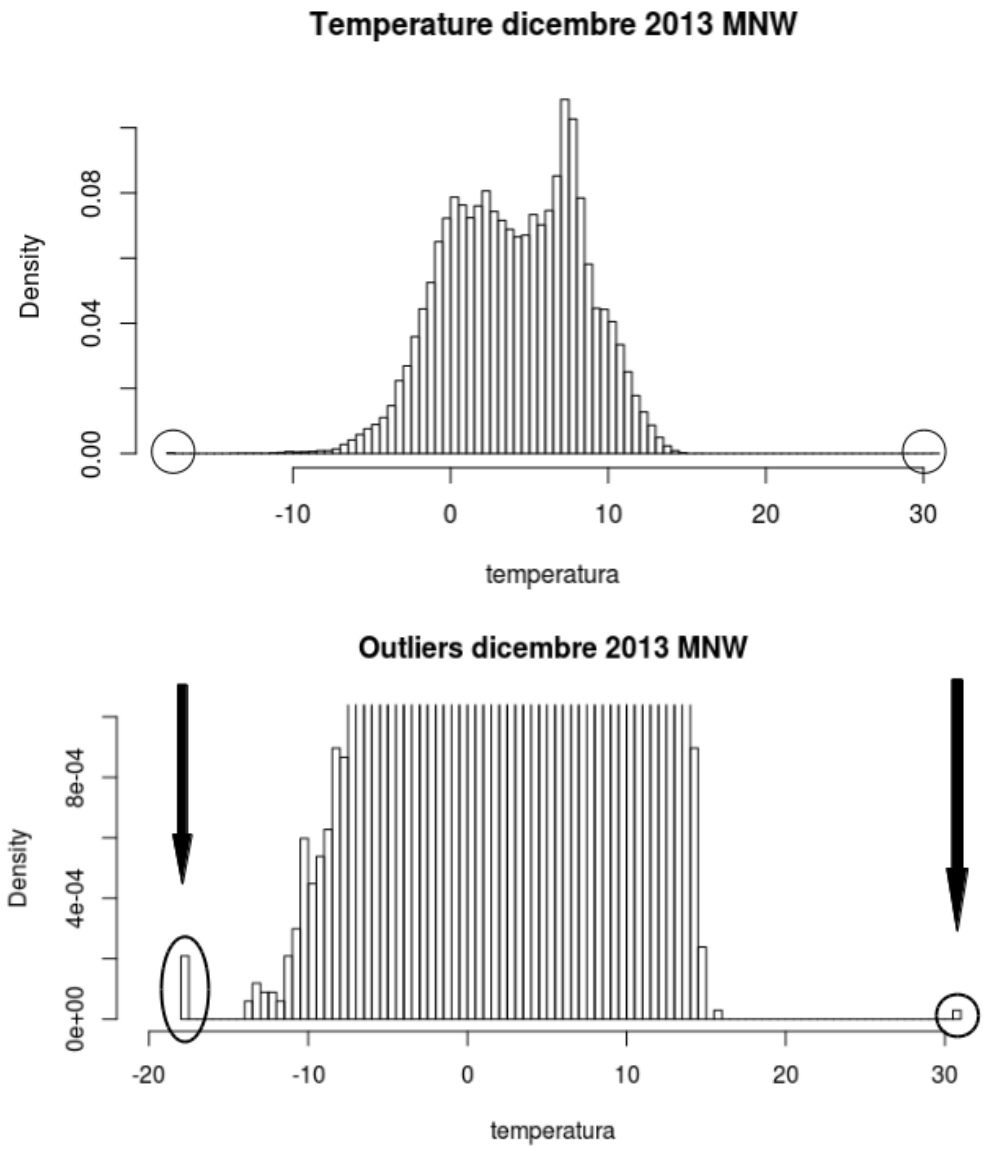


Figura 3: Esempio di outliers. In basso un ritaglio del grafico superiore, nell'estremo destro e sinistro si possono notare delle osservazioni isolate.

Per controllare i valori pari a zero ho sviluppato una funzione che potesse capire in automatico se uno zero fosse anomalo o veritiero, ciò è risultato molto utile nei mesi più freddi dell'anno, come quelli invernali in cui è molto probabile rilevare una temperatura pari proprio a zero. La funzione trova tutti gli zeri presenti in un certo mese e guarda se è un'osservazione isolata oppure un blocco di zeri. Nel secondo caso li segnala subito, mentre nel primo caso controlla le osservazioni precedenti e successive disponibili. Se vi sono osservazioni vicine nel tempo allora guarda se nelle ore appena prima e appena dopo ci sono dei valori che giustificano la rilevazione dello 0. In caso affermativo esso viene mantenuto nel data set, altrimenti viene segnalato. La funzione fornisce in output una matrice con tutte le anomalie ritrovate e anche quei valori per i quali non erano presenti sufficienti osservazioni precedenti e/o successive per poter fare un controllo automatico. La funzione tende dunque a trovare più elementi problematici del necessario e serve quindi poi procedere ad un controllo manuale. Alcuni valori anomali isolati sono stati sostituiti da una media dei valori rilevati nelle ore precedenti e successive se presenti.

A seguito di tutti i controlli le stazioni peggiori si sono rivelate la vnt201 nel periodo da giugno 2013 fino a gennaio 2014 e la stazione vnt162 da dicembre 2013 a maggio 2015. Ci sono state anche altre stazioni a cui è stata prestata particolare attenzione e che sono state eliminate per uno o più mesi (vnt91, vnt232, vnt110, vnt222, vnt181, vnt363, vnt54 e vnt282).
A seguito dei controlli effettuati si può affermare che la qualità dei data set MeteoNetwork è migliorata con il passare del tempo, in quanto si sono registrate via via meno anomalie e si sono conservate più osservazioni.

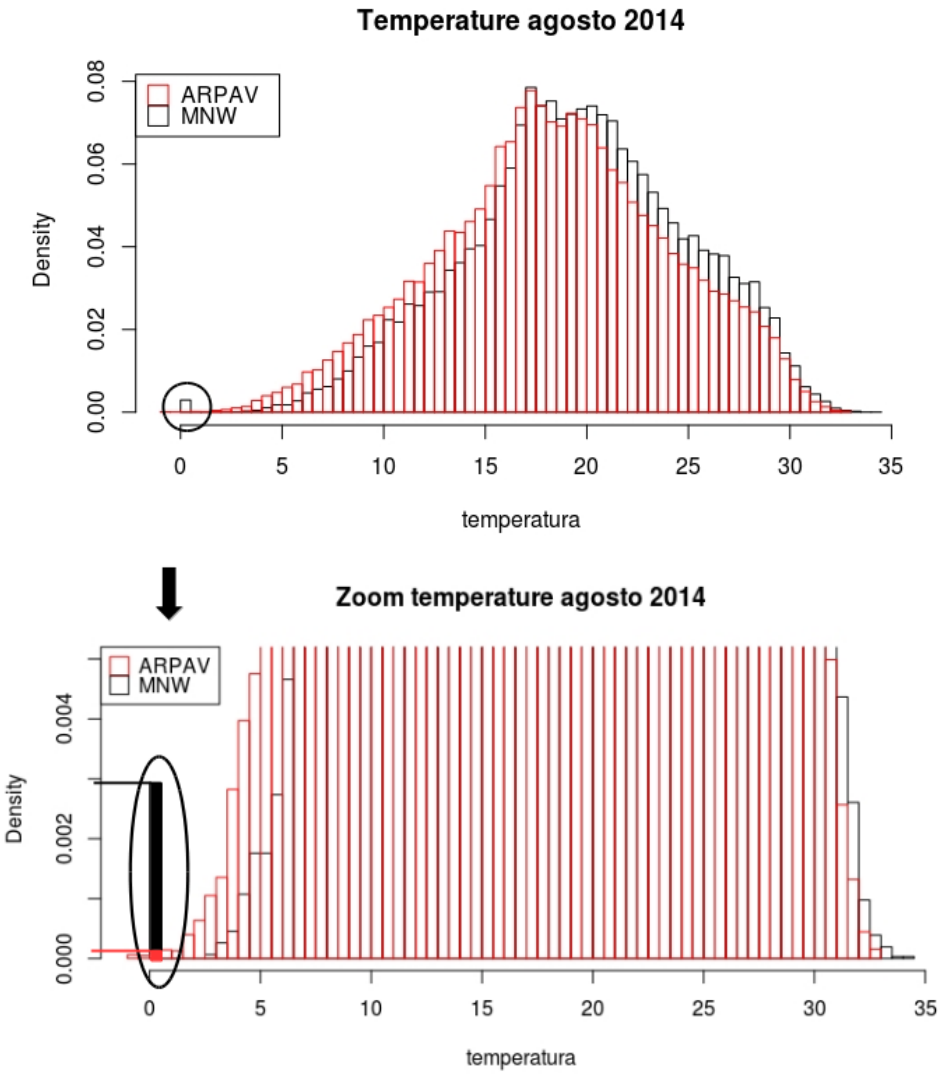


Figura 4: Esempio per evidenziare l'anomalia nel valore 0, il grafico in basso è un ritaglio del grafico superiore.

Modelli di regressione lineare

Un'attenta analisi dei data set separati è necessaria per capire quali possono essere le basi per strutturare il confronto tra le due fonti di rilevazione dei valori meteorologici. Per ogni anno si è stimato un modello di regressione lineare multipla, sia per i dati di ARPAV sia per i dati di MeteoNetwork, per poter vedere le relazioni lineari presenti tra la temperatura e le altre variabili considerate.

Analisi esplorativa

Per prima cosa si è effettuata un'analisi esplorativa di entrambi i data set, calcolando prima la correlazione tra le variabili e producendo poi dei grafici per le variabili più correlate con la risposta. Si nota in Tabella 1 che latitudine, data e mese risultano sempre le più influenti. Nei grafici in Figura 6 e 7 si vuole evidenziare la presenza di una doppia stagionalità nei dati che sarà modellata tramite delle variabili indicatrici.
Si è provato a vedere se le stagionalità mensili e orarie fossero abbastanza forti e si è notato che, infatti, le mediane dei gruppi creati in base al mese o in base alle ore della giornata differiscono tra di loro con un andamento che rispecchia la logica stagionale annuale e giornaliera, ma non si è stabilito se le differenze fossero significative o meno. Si è notato però che la stagionalità mensile è più forte nel data set di MeteoNetwork mentre quella oraria sembra più presente nel data set ARPAV.
Inoltre si è visualizzato graficamente l'andamento della temperatura ri-

spetto alla latitudine e dal diagramma di dispersione in Figura 5 si intuisce che la relazione tra le due variabile è decrescente e nel modello il coefficiente della latitudine risulterà probabilmente negativo.

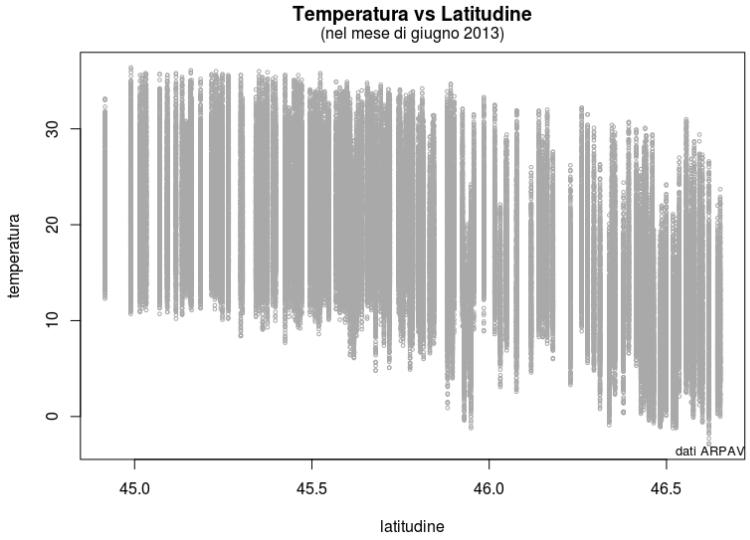


Figura 5: Diagramma di dispersione tra latitudine e temperatura (2013).

Tabella 1: Correlazioni tra la temperatura e latitudine, data e mese divise per anno e fonte dei dati. La variabile mese è stata considerata quantitativa per calcolare la correlazione.

Variabile	2013		2014		2015	
	ARPAV	MNW	ARPAV	MNW	ARPAV	MNW
Latitudine	-0.3408	-0.251	-0.39	-0.3338	-0.3099	-0.225
Data	-0.6876	-0.7576	0.1448	0.1686	0.1517	0.1329
Mese	-0.6861	-0.7468	0.1379	0.1768	0.1584	0.1402

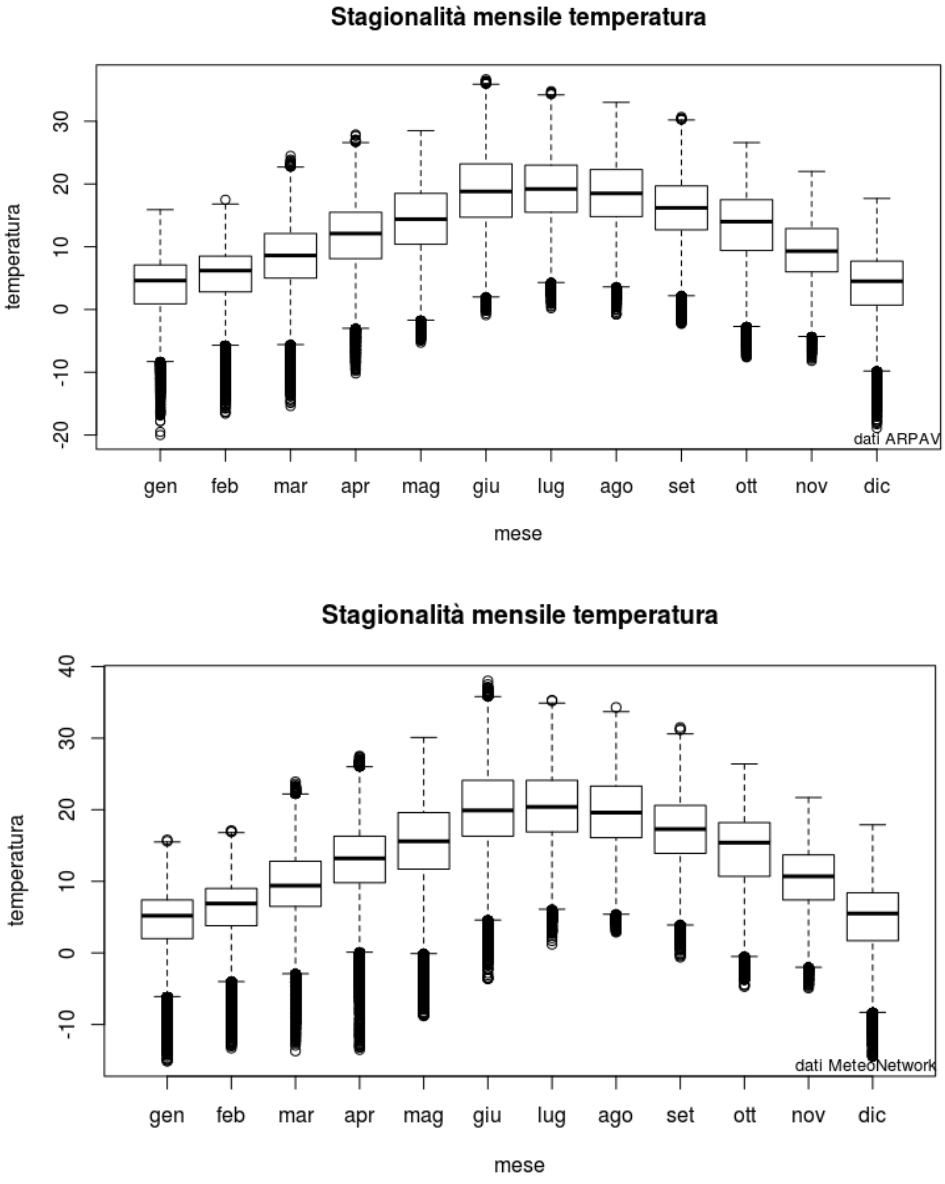


Figura 6: Boxplot per la stagionalità mensile nel 2014.

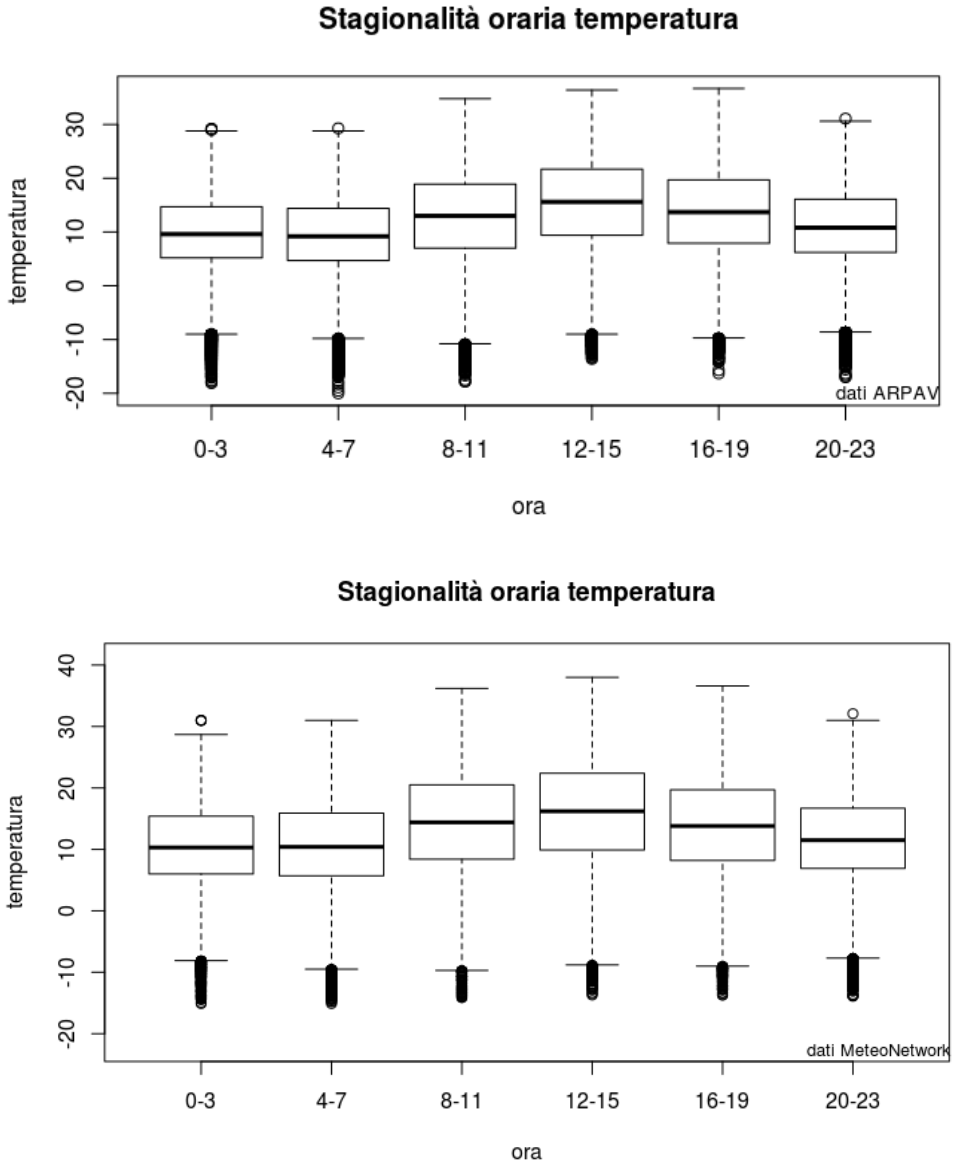


Figura 7: Boxplot per la stagionalità oraria nel 2014.

Stima del modello

Per iniziare si è stimato un modello di regressione lineare multipla.
Nell'analisi si è posta come variabile risposta la temperatura, che presenta però una dipendenza tra le osservazioni. Perciò si è cercato di modellarla inserendo delle variabili che descrivono le dimensioni tempo e spazio. Per il tempo si sono considerate una variabile quantitativa continua ad indicare lo scorrere dei giorni e delle variabili indicatrici per individuare il mese e la fascia oraria di appartenenza dell'osservazione. La giornata è stata divisa in 6 fasce orarie da 4 ore ciascuna a partire dalla mezzanotte. Per lo spazio, invece, si è fatto uso di longitudine e latitudine come variabili quantitative.
Il modello da stimare risulta dunque

$$T_i = \beta_1+\beta_2 y_{i1}+\beta_3 y_{i2}+\alpha_1 x_{i1,2}+....+\alpha_{11} x_{i1,12}+\alpha_{12} x_{i2,2}+...+\alpha_{16} x_{i2,6}+\beta_{17} x_{i3}+\epsilon_i$$

$$\begin{aligned} \underline{T} &= \underline{\beta} X + \underline{\epsilon} \\ \underline{\epsilon} &\sim N(\underline{0}, \sigma^2 I_n) \end{aligned} \tag{1}$$

con T=temperatura; y_1=longitudine; y_2=latitudine; $x_{1,j}$=1 se il mese è j, j=2,..,12; $x_{2,j}$=1 se la fascia oraria è j, j=2,..,6; x_3=data.
Da quanto emerso nell'analisi esplorativa, ci si aspetta un andamento nel modello che evidenzi la stagionalità rispetto ai mesi e rispetto alle parti del giorno.

Il modello stimato risulta negli anni assumere i coefficienti e i valori di R^2 presenti nella Tabella 2. È stato adattato un modello molto semplice nel quale i coefficienti di regressione sono facilmente interpretabili. Tutti i parametri inseriti sono risultati fortemente significativi nonostante la correlazione con le variabili concomitanti non fosse elevata. Si nota dalla Tabella 2 che essi seguono lo stesso andamento in tutti gli anni, sia per dati ARPAV sia per i dati MeteoNetwork. Si evidenzia che la temperatura tende a crescere con la longitudine mentre si abbassa all'aumentare della latitudine e dello scorrere dei giorni all'interno di un mese. Inoltre la

Tabella 2: Coefficienti di regressione e indice R^2 per i modelli di regressione lineare multipla stimati, divisi per anno e fonte dei dati.

Variabile	2013		2014		2015	
	ARPAV	MNW	ARPAV	MNW	ARPAV	MNW
Intercetta	1313	1702	1256	1264	1119	1280
Longitudine	1.089	1.897	1.195	1.754	0.8268	1.59
Latitudine	-6.981	-7.384	-7151	-7.928	-6.470	-7.099
Data	-0.0624	-0.08631	-0.05707	-0.05571	-0.04915	-0.05757
Gennaio			-23.54	-23.67	-24.53	-26.07
Febbraio			-20.41	-20.51	-22.19	-23.42
Marzo			-15.66	-15.57	-17.07	-18.15
Aprile			-10.62	-10.5	-11.73	-12.53
Maggio			-6.334	-6.329	-5.41	-5.816
Luglio	5.647	6.282	2.103	2.061	5.767	5.908
Agosto	6.491	7.799	3.157	2.967	5.05	5.593
Settembre	4.082	5.851	2.504	2.231	1.241	1.908
Ottobre	1.762	4.247	1.456	1.213	-2.118	-1.326
Novembre	-1.654	1.584	-0.8381	-1.168	-5.068	-4.237
Dicembre	-3.542	0.431	-4.371	-4.757	-7.218	-6.203
Ora 4-7	-0.3007	0.1601	-0.3363	0.1825	-0.3694	0.1765
Ora 8-11	3.700	4.211	3.152	3.947	3.461	4.325
Ora 12-15	6.501	6.196	5.754	5.712	6.552	6.413
Ora 16-19	4.627	3.675	4.093	3.505	4.592	3.826
Ora 20-23	1.262	1.039	1.193	1.116	1.3	1.179
R^2	0.7775	0.7735	0.7599	0.7315	0.8042	0.7932

N.B.: Vengono presi a riferimento il mese di giugno e la fascia oraria 0-3.

stagionalità segue i boxplot presentati prima di stimare il modello. Infatti la temperatura presenta un minimo nella fascia oraria 4-7 per ARPAV e nella fascia 0-3 per MeteoNetwork, aumenta fino alle 12-15 poi diminuisce. Lo stesso accade per i mesi, dove il massimo si registra ad agosto nel 2013 e 2014 e a luglio nel 2015 e poi la temperatura scende con un minimo registrato a gennaio per poi riprendere ad aumentare da febbraio.

Vengono dunque confermate le conoscenze di base del fenomeno, ossia che la temperatura è più bassa a latitudini più elevate e segue una stagionalità che la porta ad essere più alta d'estate e nel primo pomeriggio e più bassa d'inverno e tra la notte e la mattina. Ciò sembra banale, ma è molto utile in quanto indica che il modello utilizzato non sta conducendo verso una direzione lontana dalla realtà, ma anzi si adatta a raccontarla nonostante sia così semplice.

L'indice di bontà del modello, R^2, varia negli anni assumendo sempre un valore molto alto che indica che il modello spiega circa tra il 73% e l'80% della variabilità totale della variabile risposta. È un buon risultato e per quanto riguarda MeteoNetwork è importante il fatto che migliori nel tempo in quanto sottolinea che la qualità dei dati rilevati cresce come preannunciato a seguito del controllo dei data set. Tutto ciò però non basta per capire se il modello si adatta bene ai dati che descrive.

Validazione del modello

Per controllare il modello sono stati utilizzati i residui standardizzati, ovvero trasformati in modo tale che avessero media nulla e varianza unitaria.
Dal diagramma quantile contro quantile (qq-plot) in Figura 8 si nota che i residui non seguono proprio del tutto una distribuzione normale ma ci si avvicinano molto. Nel 2013 e 2015 si notano delle code più pesanti

rispetto alla distribuzione Normale standard ma per quanto riguarda il 2013 ciò potrebbe essere dovuto al fatto che i dati partono da giugno anziché da gennaio come negli altri anni.
L'ipotesi di normalità sembra essere quindi quasi confermata anche dall'istogramma dei residui che si posiziona perfettamente sotto la curva della Normale Standard.
Dal diagramma di dispersione con la temperatura stimata, presente in Figura 9, si verifica che non c'è correlazione tra i residui e i valori previsti dal modello in quanto viene rappresentata una nube informe di punti che non lascia spazio ad andamenti sistematici evidenti e conferma l'ipotesi di omoschedasticità. La correlazione risulta prossima a 0, di conseguenza temperatura stimata e residui non sono correlati tra loro. L'assenza di correlazione viene evidenziata dalla linea rossa, che interpola i dati, nonostante non sia perfettamente parallela all'asse delle ascisse. Essa costituisce una media dei residui calcolata per intervalli di dimensione 1 della temperatura.
I residui sembrano dunque soddisfare tutte le ipotesi di un modello di regressione lineare normale, ciò significa che il modello, per quanto semplice ed azzardato in presenza di risposte dipendenti, può essere utile per una prima analisi.

Il modello sembra buono anche se non viene rispettata l'ipotesi di indipendenza della variabile risposta. La temperatura infatti presenta una dipendenza sia spaziale che temporale e ciò influenza i p-value facendo in modo che quelli calcolati dal modello siano più piccoli dei reali. In questo caso però, i p-value risultano tutti prossimi a 0 quindi il problema non si pone in maniera rilevante.
Si potrebbe migliorare il modello inserendo una funzione quadratica delle coordinate spaziali, però sarebbe meglio provare un modello diverso che possa gestire in maniera migliore le variabili esplicative.

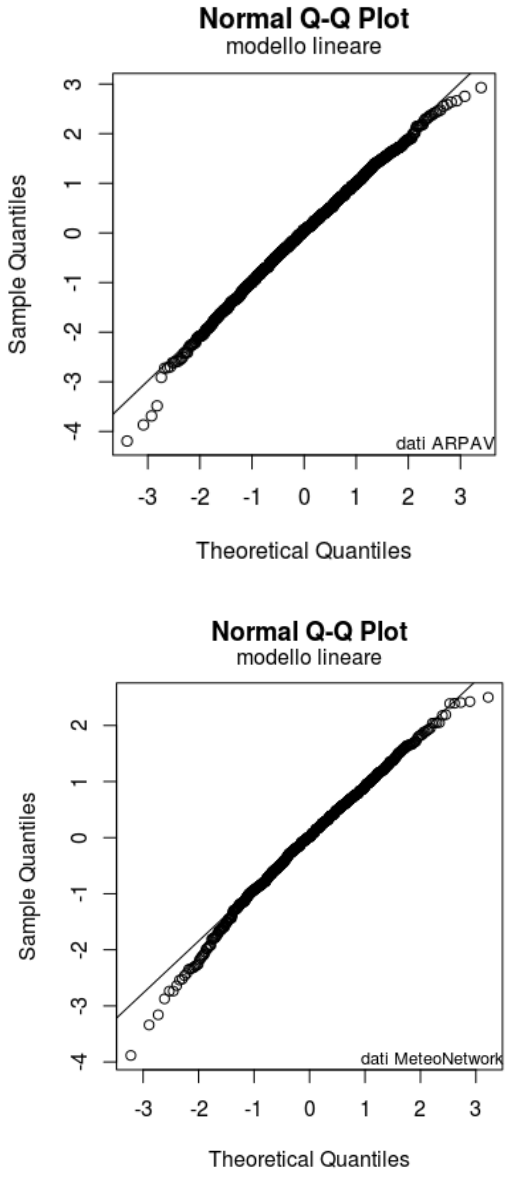


Figura 8: Diagramma quantile contro quantile dei residui dei modelli adattati ad ARPAV e MeteoNetwork nel 2014.

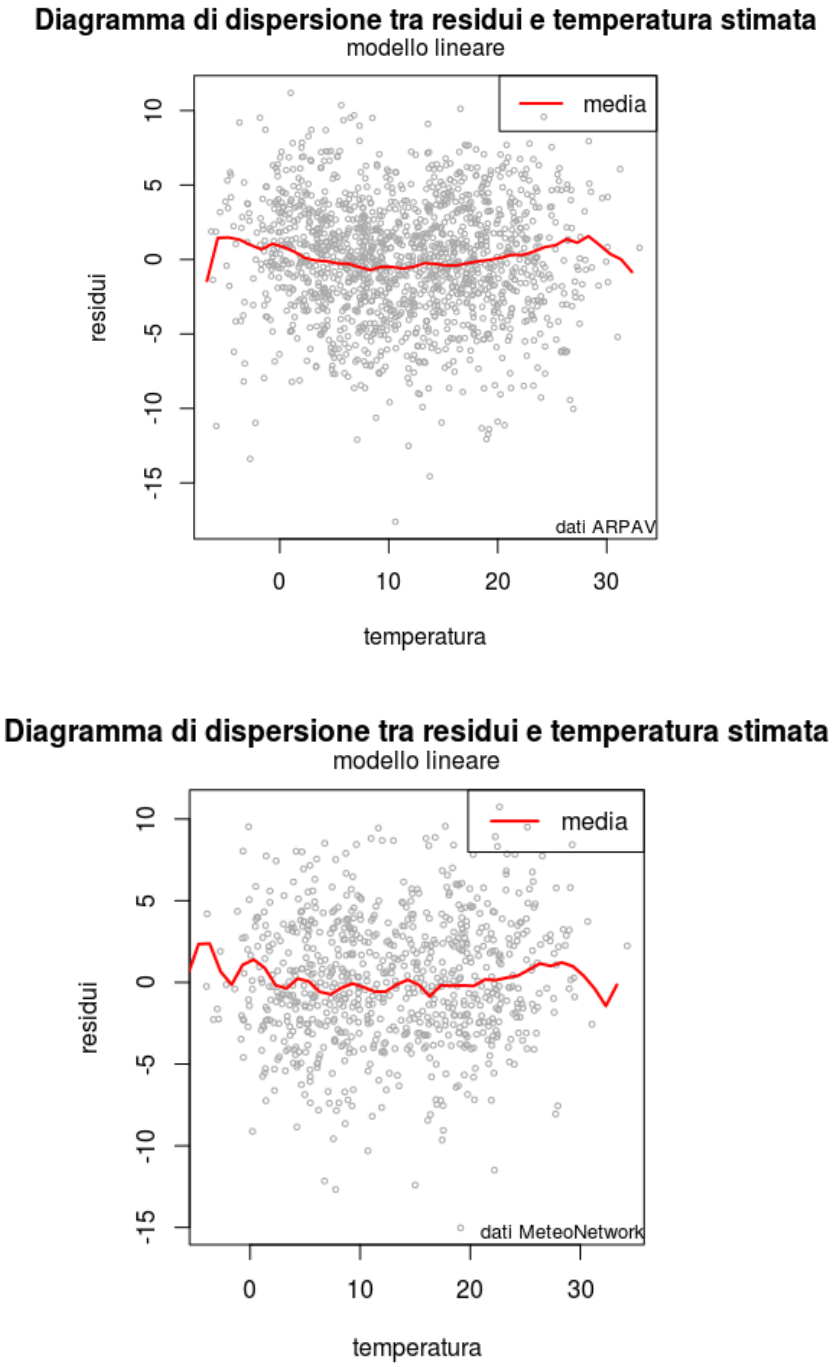


Figura 9: Diagramma di dispersione dei residui dei modelli adattati ad ARPAV e MeteoNetwork nel 2015.

Modelli additivi

Un'alternativa al modello di regressione lineare è il modello additivo. Esso è una particolare estensione dei modelli lineari che li rende più flessibili verso i dati.
Per ogni variabile esplicativa continua viene stimata una funzione non parametrica. Vengono definiti modelli additivi in quanto le funzioni vengono sommate tra di loro, come si può notare in (2). Dal punto di vista grafico, con stima della funzione si intende trovare la curva che meglio interpola i dati, proiettati su un diagramma di dispersione tra la risposta e un'altra variabile. Tutto questo scegliendo un opportuno numero di gradi di libertà in modo tale che la funzione non risulti né troppo liscia né troppo frastagliata. Ciò comporta una perdita nella facilità di interpretazione del modello ma lascia più libertà ai dati per esprimersi.
Un modello additivo viene definito come un qualsiasi modello lineare generalizzato, ossia si stabilisce una distribuzione per la variabile riposta, un predittore per la sua media e una funzione di legame. L'unica differenza si trova nel predittore che, nel caso della distribuzione Normale in cui la funzione di legame è la funzione identità, si può scrivere in questo modo

$$y_i = \alpha + \sum_{j=1}^{p} f_j(x_j) + \epsilon_i \tag{2}$$

$$\epsilon_i \sim N(0, \sigma^2)$$

Il vantaggio dei modelli additivi è che grazie ad alcuni metodi non parametrici tra cui le splines, si evita di cercare e di limitarsi ad una trasfor-

mata logaritmica o polinomiale della variabile esplicativa.
L'algoritmo su cui si basa la stima di questi modelli è una procedura iterativa chiamata backfitting, tramite la quale si stima una funzione alla volta. [AS12; HTF09]

Il modello di seguito adottato per la temperatura è in realtà semiparametrico in quanto per le variabili qualitative non è possibile stimare una curva ed esse vengono introdotte linearmente rispetto ai parametri, come nei precedenti modelli.

Stima del modello

La stima del modello additivo è stata fatta solo per i dati di MeteoNetwork in quanto esso verrà utilizzato per prevedere dei valori per la temperatura sulla base delle variabili esplicative misurate da ARPAV. Tutto ciò è propedeutico per il primo confronto che sarà effettuato tra le due reti di rilevazione.
Si è adattato quindi ai dati il seguente modello

$$Y_i = \beta_0 + f_1(longitudine, k) + f_2(latitudine, k) + f_3(data, k) + \underline{\gamma} \cdot mese + \underline{\alpha} \cdot ora + \epsilon_i$$

$$\underline{\epsilon} \sim N(\underline{0}, \sigma^2 I_n) \tag{3}$$

Si sono utilizzate le splines solo per le variabili quantitative ossia longitudine, latitudine e data in quanto non è possibile applicarle alle variabili qualitative come sono le indicatrici per i mesi e le fasce orarie. Il parametro k è riconducibile ai gradi di libertà utilizzati per stimare la funzione di un regressore. Per quanto riguarda il modello stimato si è scelto di lasciare longitudine e latitudine ai valori di default, ossia k=4, mentre per la scelta di k per la variabile data si è dato spazio ad una valutazione grafica che ha mostrato come si adattava ai dati la curva in funzione di k. Alla fine si è preso in considerazione il valore 13 per l'anno 2013 e 18 per il 2014 e 2015.

Le funzioni stimate sono risultate tutte fortemente significative e dai loro grafici in Figura 10, 11 e 12 si evidenzia un andamento simile a quello ottenuto con la regressione lineare, ma si può dire molto di più sull'effetto delle variabili quantitative sulla temperatura. Infatti si nota come la latitudine abbia un effetto inversamente proporzionale alla temperatura ma con intensità diversa nel suo dominio. Inoltre si evidenzia che la longitudine mostra un andamento fluttuante attorno ad una retta parallela all'asse delle ascisse e dunque sembrerebbe meno influente rispetto alle altre variabili. Infine la data ha una funzione decrescente e questo rispecchia il coefficiente negativo ottenuto nel modello lineare. Tuttavia questo non vuol dire che la temperatura diminuisce con il passare del tempo. Infatti i grafici iniziali in Figura 2 non mostrano un trend negativo ma evidenziano delle stagionalità. Il modello coglie completamente la stagionalità con le variabili mese ed ora mentre è plausibile che la funzione stimata per la data decresca in modo tale da compensare l'aumento progressivo del valore della variabile causato dal passare dei giorni.

Rispetto al modello lineare si può affermare che questo è migliore in quanto vede una diminuzione della varianza dei residui, della devianza residua e del criterio di informazione di Akaike (AIC) come si può leggere nella Tabella 3. La devianza residua (RD) è stata riscritta in termini di R^2 e devianza nulla (ND) come $R^2 = 1 - \frac{RD}{ND}$.
I gradi di libertà usati per la stima del modello sono 15 per il modello lineare e 33 per il modello additivo nel 2013. Salgono a 20 per i modelli lineari e 43 per i modelli additivi nel 2014 e 2015. I modelli additivi risultano infatti più costosi, anche in termini di costo computazionale.

Tabella 3: Confronto tra modello lineare ed additivo per i dati MeteoNetwork in termini di AIC, R^2 e varianza dei residui (σ^2_R).

		AIC	R^2	σ^2_R
2013	lineare	2213250	0.77	17.09
	additivo	2071114	0.84	11.87
2014	lineare	4369583	0.73	15.36
	additivo	4171668	0.79	11.93
2015	lineare	4488806	0.79	16.53
	additivo	4284599	0.84	12.79

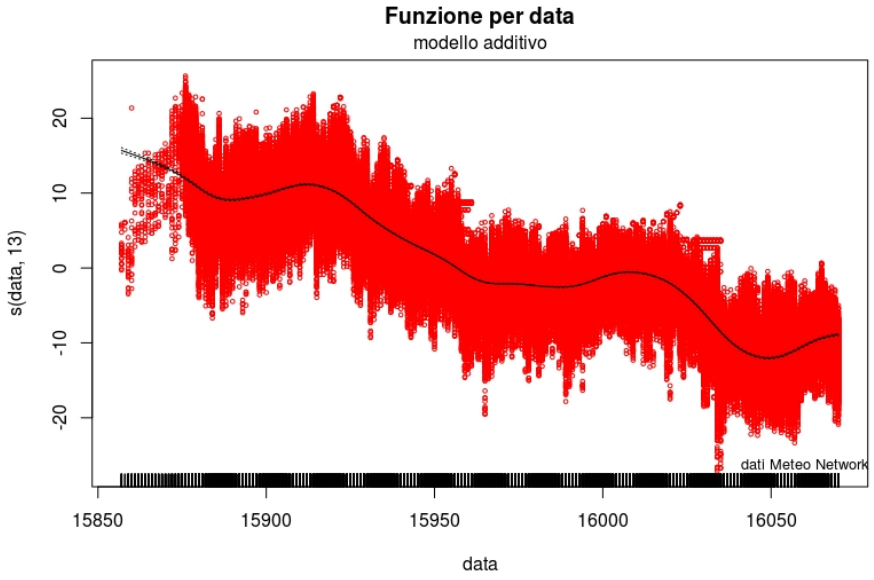


Figura 10: Funzione stimata per data nel 2013.

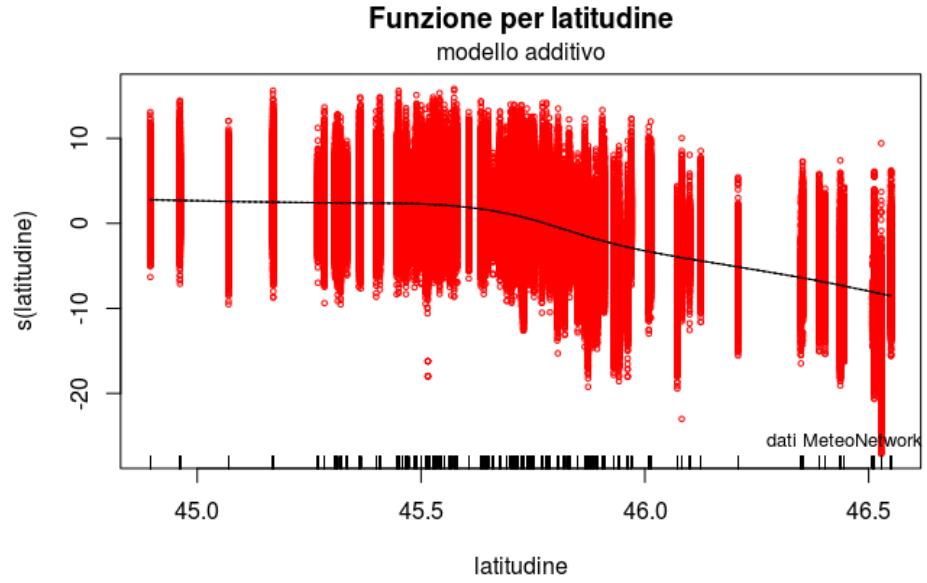


Figura 11: Funzione stimata per latitudine nel 2014.

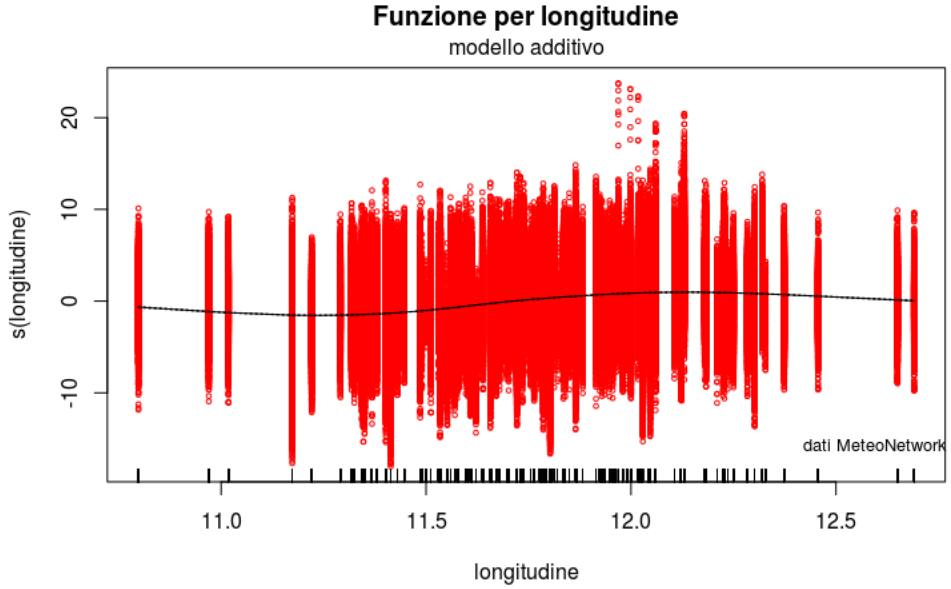
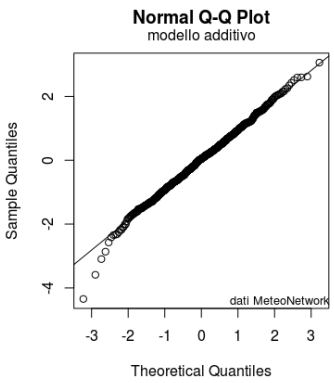


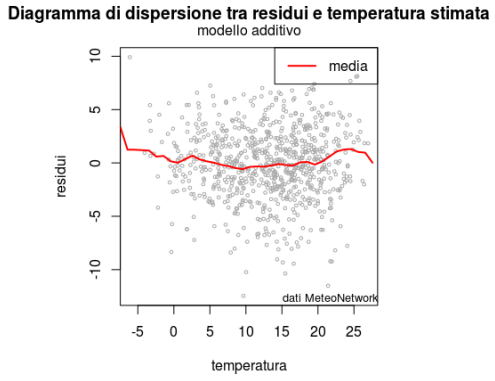
Figura 12: Funzione stimata per longitudine nel 2015.

Validazione del modello

Per verificare che il modello additivo fosse adatto a descrivere i dati si sono controllati i residui nello stesso modo in cui si è proceduto per il modello lineare. Anche in questo caso i diagrammi quantile contro quantile, ad esempio quello in Figura 13(a), mostrano che l'assunzione di normalità per la distribuzione degli errori è valida. Inoltre il diagramma di dispersione tra i valori stimati ed i residui nella Figura 13(b) mostra assenza di correlazione e di eteroschedasticità.
Tutto ciò è simile a quanto considerato anche per il modello lineare. Come si poteva pensare sin dall'inizio, il modello additivo si è dimostrato più flessibile nell'adattarsi ai dati e questo ha comportato che risultasse migliore sotto vari punti di vista rispetto al primo modello. Ciononostante risulta molto più difficile da interpretare in quanto non si ha un unico coefficiente che esprime la relazione tra la temperatura ed il regressore, ma c'è una curva che interpola i dati.



(a) *Diagramma quantile contro quantile (2015).*



(b) *Diagramma di dispersione (2014).*

Figura 13: Due dei grafici utilizzati per controllare le assunzioni dei residui dei modelli.

Confronto tra ARPAV e MeteoNetwork

Obiettivo principale di questo studio è il confronto tra le due reti di rilevazione. Per fare questo si sono pensati due metodi diversi che coinvolgono l'intero data set e una prima analisi descrittiva che divide i dati in coppie di stazioni in base alla loro distanza.

Qualità delle previsioni con MeteoNetwork

Come confronto iniziale tra i due sistemi di rilevazione dei dati si è scelto di utilizzare un modello di regressione lineare per la sua semplicità di interpretazione. Il modello additivo (3) stimato per i dati di MeteoNetwork è stato utilizzato per fare previsione. Per fare ciò sono stati presi come valori delle variabili esplicative quelli presenti nel data set di ARPAV. Queste previsioni sono state integrate in un nuovo modello di regressione lineare per i dati ARPAV.
Si è quindi stimato il seguente modello

$$T_i = \beta_1+\beta_2 y_{i1}+\beta_3 y_{i2}+\alpha_1 x_{i1,2}+....+\alpha_{11} x_{i1,12}+\alpha_{12} x_{i2,2}+...+\alpha_{16} x_{i2,6}+\beta_{17} x_{i3}+\beta_{18} p_i+\epsilon_i$$

$$\begin{aligned} \underline{T} &= \underline{\beta} X + \underline{\epsilon} \\ \underline{\epsilon} &\sim N(\underline{0}, \sigma^2 I_n) \end{aligned} \tag{4}$$

con T=temperatura; y_1=longitudine; y_2=latitudine; $x_{1,j}$=1 se il mese è j, j=2,..,12; $x_{2,j}$=1 se la fascia oraria è j, j=2,..,6; x_3=data; p=previsioni. Se il data set di MeteoNetwork cogliesse tutte le informazioni necessarie per descrivere la temperatura, il modello ampliato dovrebbe mostrare il coefficiente dei valori predetti come significativo e porre statisticamente a 0 gli altri coefficienti.
Nell'anno 2013 si nota che è la latitudine quella variabile che non risulta più utile inserendo le previsioni nel modello. Nel 2014 e nel 2015, invece, risultano statisticamente nulli i coefficienti relativi ad alcuni mesi, mentre per qualche fascia oraria e qualche mese la significatività si abbassa (p-value maggiore). Questo comportamento, osservabile nella Tabella 4, rappresenta un primo passo verso l'idea di somiglianza delle due reti di rilevazione.

Si è provato a confrontare in Figura 15 un insieme di valori osservati e predetti, cercando di visualizzare nella mappa le zone più e meno calde. Si è notato che la distribuzione delle previsioni, nonostante sia poco flessibile, riesce a rispecchiare abbastanza bene l'andamento della temperatura osservata.
I valori predetti risultano molto correlati con la temperatura osservata, come si può notare nella Figura 14, che evidenzia il valore della correlazione pari a 0.9 nel 2013 e 2014, 0.92 nel 2015. Per questo motivo si è stimato un modello di regressione lineare semplice utilizzando come variabile esplicativa solamente le previsioni

$$y = \beta_0 + \beta_1 \cdot x, \tag{5}$$

dove y rappresenta la temperatura osservata e x i valori predetti. Dal modello (5) si evince che i valori predetti approssimano la temperatura misurata da ARPAV in maniera leggermente distorta in quanto l'intercetta non è statisticamente pari a 0, mentre il coefficiente angolare si può assumere uguale a 1. Un motivo di questa distorsione si può trovare nel grafico in Figura 14, dove si nota che le previsioni tendono a sottostimare

i valori alti della temperatura osservata.
Inoltre il modello (5) è stato confrontato in termini di analisi della varianza e di indice R^2 con il modello che presenta tutte le variabili (4). L'indice R^2 risulta praticamente uguale nei due modelli.
Si è confrontata anche la capacità esplicativa del modello di partenza (1) con il modello di regressione semplice (5) e si è notato che quest'ultimo è migliore in termini di devianza spiegata ed R^2 del modello con tutte le variabili concomitanti. L'aumento del coefficiente di determinazione R^2 è tra il 4 e 5% di varianza spiegata ogni anno, come si nota in Tabella 5.

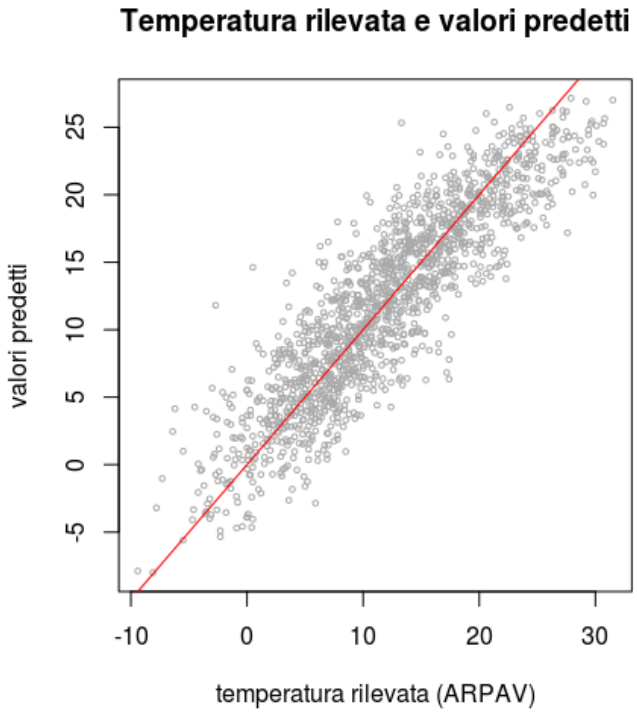


Figura 14: Diagramma di dispersione tra temperatura osservata e valori predetti nel 2014. La linea rossa indica la bisettrice del primo e terzo quadrante.

Tabella 4: Coefficienti di regressione e loro significatività, per i modelli di regressione lineare con i valori predetti, divisi per anno.

Variabile	2013		2014		2015	
	coefficiente	p-value	coefficiente	p-value	coefficiente	p-value
Intercetta	-635.700	<0.001	-73.142	<0.001	-132.700	<0.001
Longitudine	0.228	<0.001	0.278	<0.001	0.143	<0.001
Latitudine	0.011	0.521	0.729	<0.001	0.492	<0.001
Data	0.039	<0.001	0.002	<0.001	0.007	<0.001
Gennaio			0.880	<0.001	1.500	<0.001
Febbraio			0.719	<0.001	1.233	<0.001
Marzo			0.372	<0.001	1.054	<0.001
Aprile			0.230	<0.001	0.829	<0.001
Maggio			0.177	<0.001	0.378	<0.001
Luglio	0.426	<0.001	-0.050	0.003	-0.153	<0.001
Agosto	-0.877	<0.001	-0.046	0.055	-0.437	<0.001
Settembre	-1.905	<0.001	-0.029	0.375	-0.417	<0.001
Ottobre	-3.125	<0.001	-0.085	0.038	-0.437	<0.001
Novembre	-4.534	<0.001	-0.031	0.531	-0.371	<0.001
Dicembre	-5.830	<0.001	-0.015	0.799	-0.420	<0.001
Ora 4-7	-0.466	<0.001	-0.526	<0.001	-0.561	<0.001
Ora 8-11	-0.409	<0.001	-0.860	<0.001	-0.942	<0.001
Ora 12-15	0.460	<0.001	-0.050	<0.001	0.029	0.055
Ora 16-19	1.044	<0.001	0.530	<0.001	0.699	<0.001
Ora 20-23	0.246	<0.001	0.057	<0.001	0.100	<0.001
Predetti	0.974	<0.001	1.015	<0.001	1.016	<0.001

N.B.: Vengono presi a riferimento il mese di giugno e la fascia oraria 0-3.

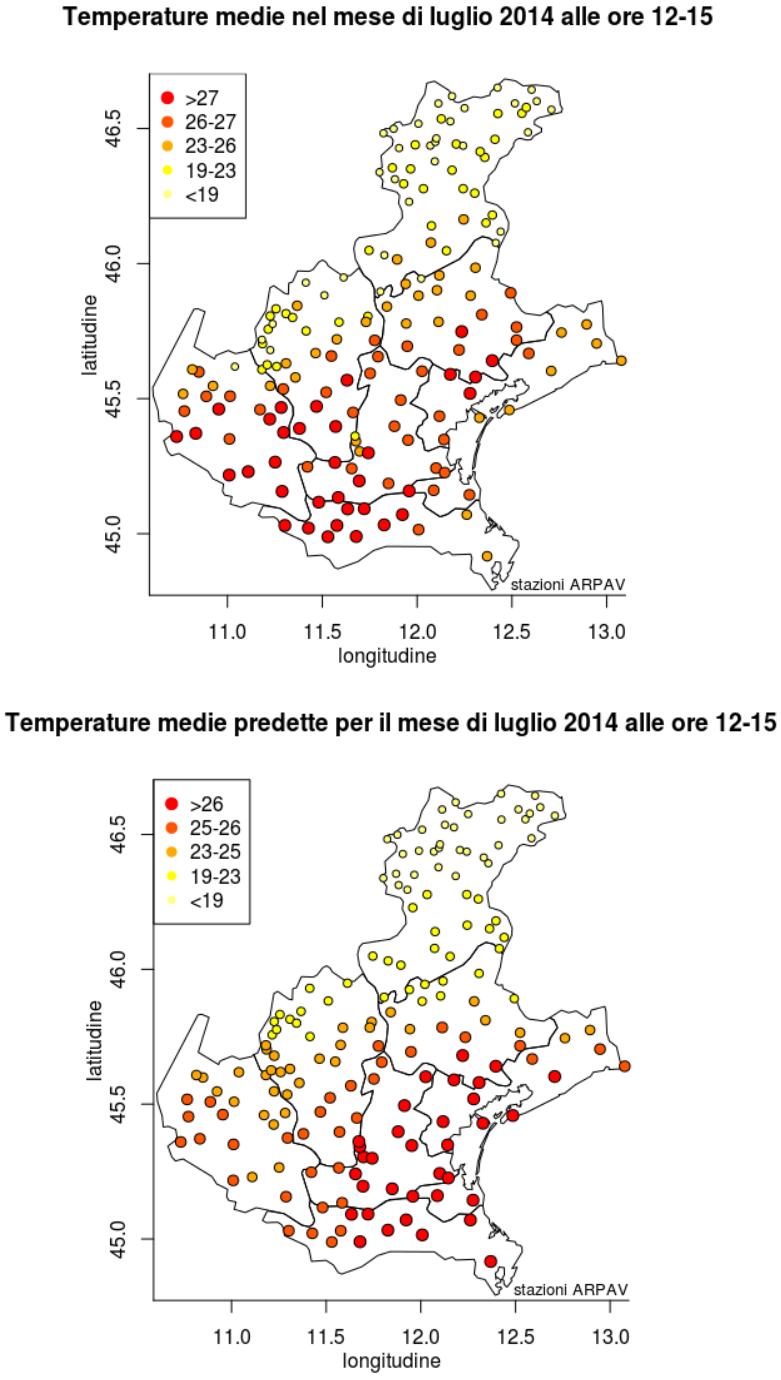


Figura 15: Mappa di confronto tra temperatura osservata e valori predetti nel 2014.

Tabella 5: Coefficienti di determinazione lineare R^2 per il modello iniziale, il modello con i valori predetti ed il modello semplice.

	Modello		
Anno	(1)	(4)	(5)
2013	0.7775	0.8271	0.8176
2014	0.7599	0.8077	0.8023
2015	0.8042	0.8421	0.8381

Secondo confronto: un modello unico

Un altro metodo usato per confrontare le due reti di rilevazione è stato stimare un unico modello di regressione lineare per tutti i dati, stando attenti ad inserire una variabile indicatrice per differenziare le due fonti di rilevazione. Sulla base del rapporto delle varianze dei residui dei modelli lineari iniziali si sono calcolati dei pesi assegnati alle osservazioni provenienti dai due data set. Questo per evitare di assumere l'omoschedasticità delle osservazioni, che provenendo da fonti diverse potrebbe non essere corretto.

$$T_i = \beta_1+\beta_2 y_{i1}+\beta_3 y_{i2}+\alpha_1 x_{i1,2}+....+\alpha_{11} x_{i1,12}+\alpha_{12} x_{i2,2}+...+\alpha_{16} x_{i2,6}+\beta_{17} x_{i3}+\beta_{18} d_i+\epsilon_i$$

$$\begin{aligned} \underline{T} &= \underline{\beta} X + \underline{\epsilon} \\ \underline{\epsilon} &\sim N(\underline{0}, \sigma^2 \underline{c}^T I_n \underline{c}) \end{aligned} \tag{6}$$

con T=temperatura; y_1=longitudine; y_2=latitudine; $x_{1,j}$=1 se il mese è j, j=2,..,12; $x_{2,j}$=1 se la fascia oraria è j, j=2,..,6; x_3=data; d=indicatrice, $d = 0$ se i dati provengono da MeteoNetwork, 1 se provengono da ARPAV; c un vettore di dimensione n con due costanti che indicano il peso dato a ciascuna fonte dei dati.

Il modello stimato mostra tutti i coefficienti fortemente significativi, an-

Tabella 6: Coefficienti di regressione della variabile indicatrice.

Anno	2013	2014	2015
Coefficiente	-0.5607	-0.6720	-0.8765

che quello della variabile indicatrice. Ciò dovrebbe presupporre che i due gruppi siano significativamente diversi in toto, ma il numero elevato di record porta il livello di significatività osservato a non essere molto preciso nel giudicare l'importanza di una variabile e di conseguenza trarre delle conclusioni valide. Si può notare, anche dai coefficienti nella Tabella 6, che il modello sottolinea ancora una volta la distorsione presente nella stima fatta con la rete privata rispetto alle temperature osservate da ARPAV in quanto sembra che i valori osservati da MeteoNetwork siano, a parità delle altre variabili, leggermente più alti. Questo conferma il risultato del modello stimato in precedenza tra la temperatura e i valori predetti (5). I residui del modello soddisfano le proprietà di normalità e omoschedasticità, come si può vedere in Figura 16.

Questo modello unico (6) è da considerare attentamente in quanto innanzitutto si basa solo sulla significatività di un coefficiente per trarre delle conclusioni ed avendo molti dati, tra l'altro dipendenti tra loro, non è facile interpretare il livello di significatività osservato (p-value). Inoltre esso implica che i coefficienti per tutte le variabili siano uguali sia per ARPAV che per MeteoNetwork, ma dall'analisi separata è emerso che è un'assunzione forte. Per migliorarlo occorrerebbe inserire un'interazione tra la variabile indicatrice e tutte le altre però ciò porterebbe all'equivalente di stimare due modelli diversi per le due reti, cosa già fatta in precedenza.

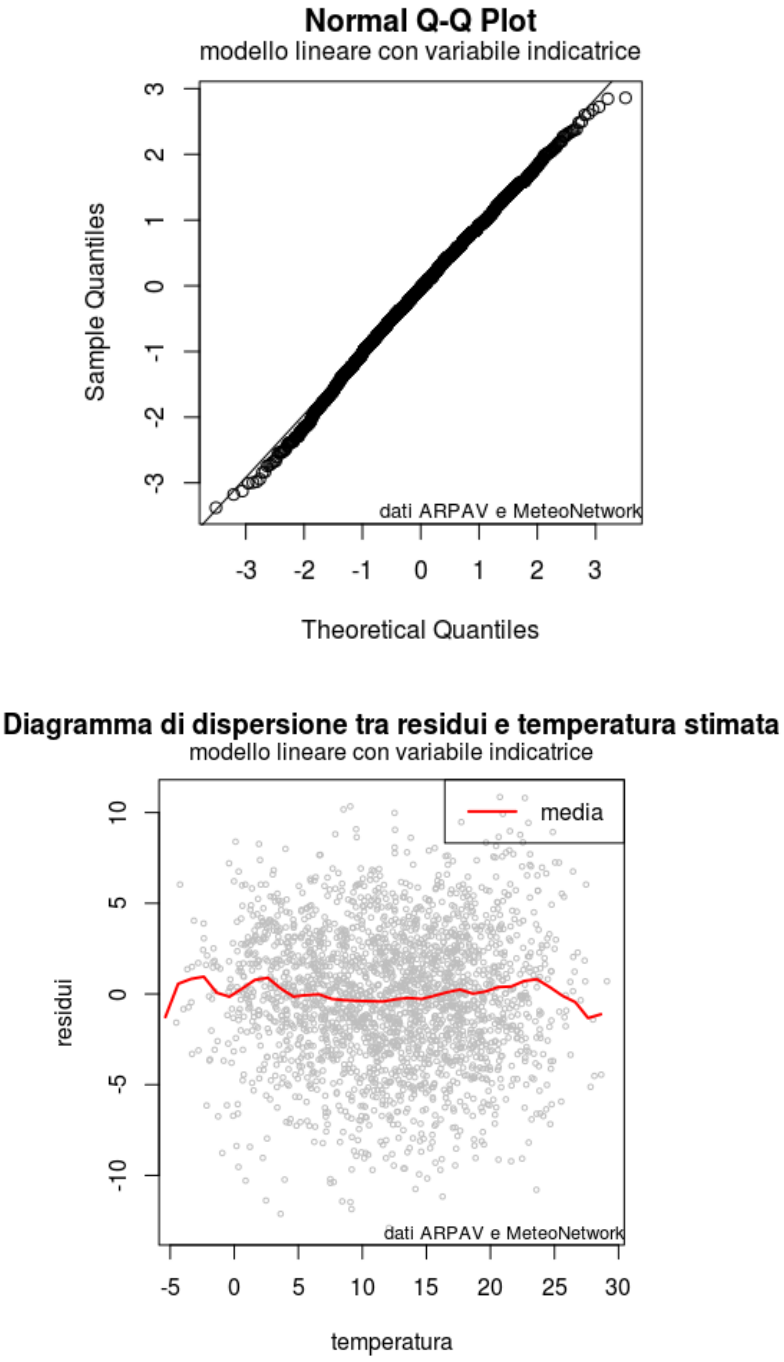


Figura 16: QQ-plot e diagramma di dispersione per i residui nel 2014.

Confronti tra singole stazioni

I due metodi visti prima si riferiscono ad un confronto su scala globale delle intere reti di stazioni meteorologiche. Ciò può risultare utile per verificare se la qualità dei due circuiti è simile e se nel complesso si comportano allo stesso modo.
In alcuni casi, però, sarebbe meglio sapere se ci sono delle stazioni statisticamente uguali. Questo perché, nel caso di una possibile integrazione, potrebbero portare a ridondanza delle informazioni ed essere dunque ridotte.
Si è provato quindi a scendere nel dettaglio analizzando coppie di centraline, una di ARPAV e una di MeteoNetwork.

Scelta delle stazioni

Per stabilire quali siano le coppie di stazioni più vicine ho calcolato la distanza tra ogni stazione ARPAV con ognuna di MeteoNetwork come $d_{km} = \Delta\sigma \cdot 6372$ dove $\Delta\sigma$ rappresenta una distanza calcolata in radianti che viene trasformata nel sistema metrico comune moltiplicandola per 6372km, un'approssimazione per il raggio medio della terra. [Wika]

$$\Delta\sigma = \arctan\left(\frac{\sqrt{(\cos(\phi_2)\cdot\sin(\Delta\lambda))^2 + (\cos(\phi_1)\cdot\sin(\phi_2) - \sin(\phi_1)\cdot\cos(\phi_2)\cdot\cos(\Delta\lambda))^2}}{\sin(\phi_1)\cdot\sin(\phi_2) + \cos(\phi_1)\cdot\cos(\phi_2)\cdot\cos(\Delta\lambda)}\right) \tag{7}$$

con ϕ_i=latitudine, $\Delta\lambda_i$=differenza di latitudine, dove $i = 1$ indica una stazione ARPAV e $i = 2$ una stazione MeteoNetwork.
Con la formula (7) calcolo la matrice di distanza e guardo quali sono i primi 50 valori più bassi tali però che le stazioni considerate non vengano ripetute in modo da poter effettuare confronti su più elementi. Di queste 50 coppie ne ho scelte 7 considerando non solo le più vicine ma anche la disposizione nel territorio in modo da avere coppie ben sparse su tutta la regione e la disponibilità di dati onde evitare di avere troppi dati man-

Tabella 7: Elenco delle coppie di stazioni ARPAV e MeteoNetwork selezionate.

Coppia	Stazione ARPAV	Stazione MeteoNetwork	Distanza km
1	s74	vnt240	0.08
2	s402	vnt124	0.33
3	s28	vnt233	0.62
4	s454	vnt177	0.77
5	s112	vnt33	0.94
6	s510	vnt189	1.26
7	s131	vnt102	2.75

canti. Infatti alcune stazioni di MeteoNetwork sono state attivate dopo il 2013 oppure sono rimaste ferme per alcuni mesi o sono state eliminate parzialmente dopo aver effettuato il controllo preliminare dei data set perché presentavano anomalie. Successivamente ho verificato anche che le altitudini delle stazioni selezionate fossero a due a due compatibili controllando un file contenente delle informazioni anagrafiche sulle centraline.

Le coppie di stazioni selezionate sono descritte nella Tabella 7 e rappresentate in Figura 17.

L'altitudine non è stata presa in considerazione subito in quanto per alcune stazioni non era presente questo dato nelle anagrafiche, infatti per la coppia s74-vnt240 non è stato possibile un confronto altimetrico.

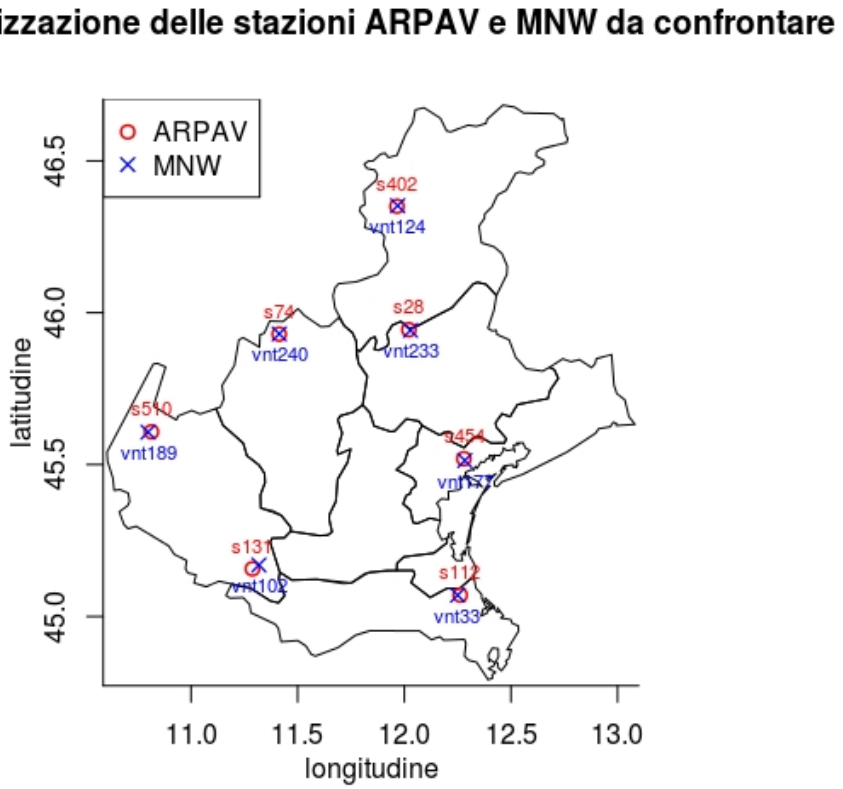


Figura 17: Mappa del Veneto con le coppie di stazioni scelte.

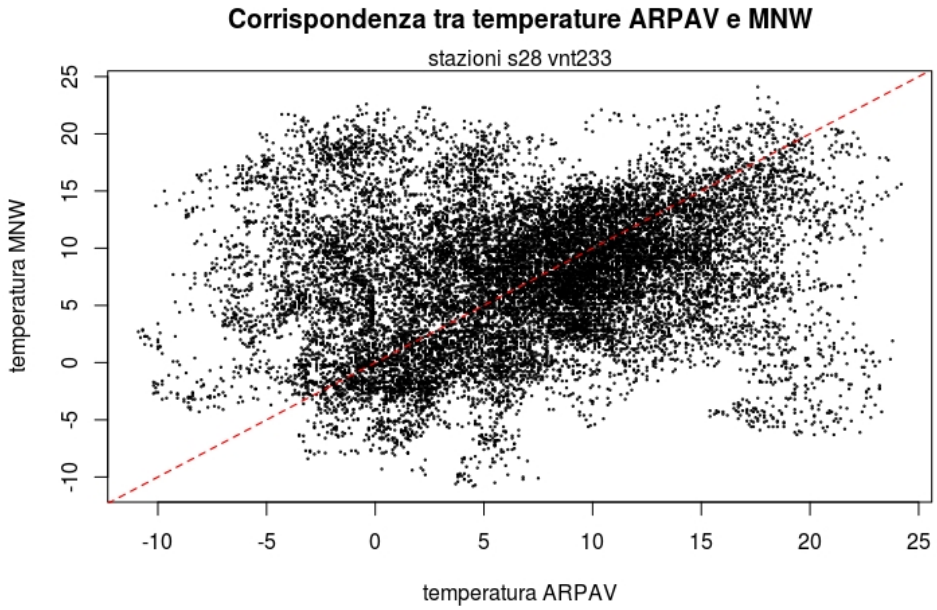
Descrizione grafica

Durante l'analisi si sono prodotti vari grafici per confrontare la compatibilità delle stazioni all'interno di ciascuna coppia.
I diagrammi di dispersione in Figura 18 rappresentano gli esempi agli estremi di quanto si corrispondano le temperature in un dato giorno alla stessa ora. Per una relazione perfetta i punti dovrebbero giacere tutti sulla bisettrice o al massimo su una parallela. Nelle coppie 1, 2 e 6 le temperature non sembrano corrispondere perfettamente tra i due sistemi di rilevazione però sembra ci sia una correlazione positiva e i valori stanno all'interno di una fascia i cui estremi sono paralleli alla bisettri-

Tabella 8: Indice di correlazione tra le temperature osservate nelle stazioni ARPAV e MeteoNetwork.

Coppia	1	2	3	4	5	6	7
Correlazione	0.431	0.531	0.218	0.784	0.771	0.603	0.873

ce. Nella coppia 3 invece, i punti sembrano disposti a caso, non si nota un andamento sistematico che evidenzi una relazione tra le temperature rilevata nelle due stazioni, tranne in una piccola zona centrale. Nella coppia 4 e 5 al contrario, la corrispondenza tra le due rilevazioni è quasi perfetta e i punti seguono la bisettrice. Anche nella coppia 7 si nota lo stesso andamento, con la differenza che la fascia di punti sembra allargarsi all'aumentare della temperatura. L'indice di correlazione per ogni coppia è riportato in Tabella 8.
Dall'analisi dei grafici delle serie storiche in Figura 19 si nota che il trend all'interno di ogni coppia è lo stesso, ossia la temperatura oscilla allo stesso modo e questo è un aspetto positivo. Le differenze non sono molte, in alcuni casi i valori sono più alti nella stazione ARPAV altre volte il contrario, ma ciò dipende anche dalla diversa posizione geografica. Inoltre anche i range variano in qualche coppia, in alcuni casi è più ampio quello della stazione MeteoNetwork, in altri accade il contrario.



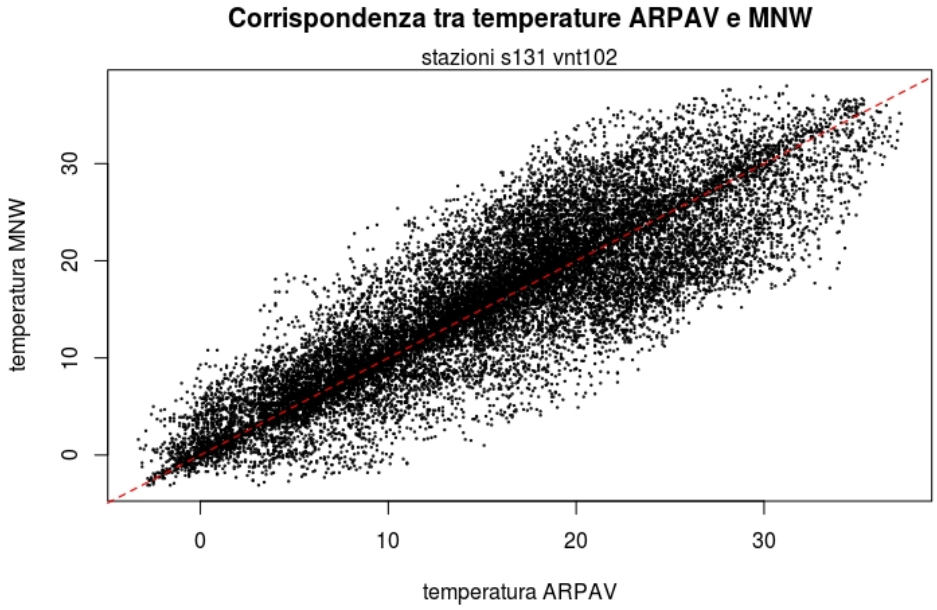


Figura 18: Confronto tra la corrispondenza della temperatura in due coppie. Nel primo caso la correlazione vale 0.22, nel secondo 0.87.

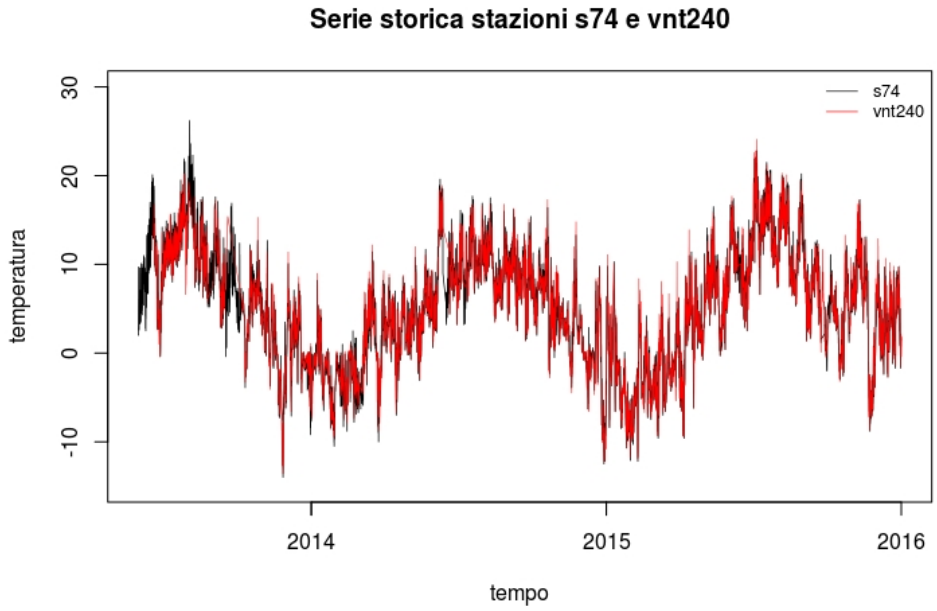


Figura 19: Serie storica della temperatura per una stazione di ARPAV e una di MeteoNetwork.

Considerazioni finali

In questo lavoro si è cercato di capire se le rilevazioni delle stazioni del circuito MeteoNetwork potessero portare delle informazioni aggiuntive per ARPAV. Per quanto riguarda la qualità dei dati della rete privata bisogna fare attenzione. Si è notato infatti che c'erano parecchi valori anomali registrati che costituiscono informazione fittizia e possono distorcere i risultati delle analisi. Una nota positiva è risultata però dal fatto che questi valori anomali sono diminuiti con il passare del tempo e ciò fa presumere che siano aumentati i controlli sui dati rilevati. Il primo passo resta comunque quello di valutare bene la qualità dei dati prima di utilizzarli per scopi specifici.
Il miglioramento delle informazioni di MeteoNetwork si è notato anche in altri ambiti. Infatti le analisi del 2015 sono risultate più precise delle precedenti. Ciò è emerso particolarmente quando si è effettuato un primo confronto con i valori predetti da un modello stimato sui dati provenienti dal circuito privato. Nel modello di confronto infatti, è risultato che i valori previsti hanno colto la maggior parte delle informazioni provenienti dalle dimensioni tempo e spazio, utili a stimare la temperatura.

L'analisi qui proposta prevedeva delle variabili di contorno facili da misurare e senza alcun costo, ma visti i buoni risultati ottenuti si può pensare di estenderla a parametri meteorologici più difficilmente rilevabili. In questo modo MeteoNetwork potrebbe portare informazioni aggiuntive ed integrare il sistema di ARPAV.

Riferimenti bibliografici

Bibliografia

[ARP16] ARPAV. *Controllo dei dati della rete di monitoraggio*. Agenzia Regionale per la Prevenzione e Protezione Ambientale del Veneto, 2016.

[AS12] Adelchi Azzalini e Bruno Scarpa. *Data Analysis and Data Mining: an introduction*. Oxford University Press, USA, 2012. Cap. 4.5.

[HTF09] Trevor Hastie, Robert Tibshirani e Jerome Friedman. *The Elements of Statistical Learning*. 2ª ed. Vol. Data Mining, Inference and Prediction. Springer Serie in Statistics. Springer-verlag New York Inc., 2009. Cap. 9.1.

[R C13] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2013. URL: `http://www.R-project.org/`.

[SAP16] Matteo Sartori, Luca Jacopo Avaldi e Paolo Patruno. *Studio statistico dell'impatto della rete MeteoNetwork sulla stima di parametri meteo superficiali in Emilia-Romagna*. Agenzia Re-

gionale per la Prevenzione e Protezione Ambientale dell'Emilia Romagna, 2016.

Sitografia

[MNW] MNW. *MeteoNetwork*. URL: http://www.meteonetwork.it/associazione.

[Rin] Rinsula. *R*. URL: http://www.insular.it.

[Rme] Rmetref. *R*. URL: http://www.agnesevardanega.eu/metref/.

[Sun] Sunearthtools. *Distanze terrestri*. URL: https://www.sunearthtools.com/it/tools/distance.php.

[Wika] Wikipedia. *Great-circle distance*. URL: https://en.wikipedia.org/wiki/Great-circle_distance.

[Wikb] Wikipedia. *Norme OMM*. URL: http://wiki.meteonetwork.it/index.php/Norme_OMM.