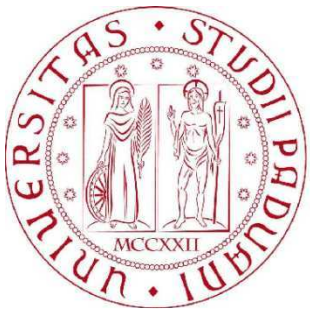
**UNIVERSITA' DEGLI STUDI DI PADOVA**




**FACOLTA' DI SCIENZE STATISTICHE**

**TESI DI LAUREA MAGISTRALE IN SCIENZE STATISTICHE**

# MODELLI STATISTICI PER LA PREVISIONE DEL CONSUMO GIORNALIERO DI GAS

**RELATORE**

**Prof. BRUNO SCARPA**

**LAUREANDO**

**DOLCETTO FEDERICO**

**ANNO ACCADEMICO 2010-2011**

*Un ringraziamento speciale*
*ai miei genitori,*
*ai miei amici,*
*all'Ing. Marco Soldà*
*e Main Consulting,*
*e a tutte le persone*
*che mi sono state vicine*
*e mi hanno supportato.*

# INDICE

**INTRODUZIONE** **11**

**CAPITOLO 1: IL SETTORE DEL GAS E DELL'ENERGIA** **15**

1.1 ENERGIA E GAS 15
1.1.1 Il ruolo del gas nel contesto energetico italiano 15
1.1.2 Il mercato del gas 16

1.2 IL CONTESTO NORMATIVO DI RIFERIMENTO 17
1.2.1 Il decreto legislativo 164/00 (Decreto Letta) 17

1.3 IL RUOLO DI SNAM RETE GAS 19
1.3.1 I clienti 19
1.3.2 La Capacità di trasporto 19
1.3.3 La struttura del settore 21

1.4 IL CONCETTO DI CAPACITA' 23
1.4.1 Capacità ai Punti di Entrata della Rete Nazionale di Gasdotti 23
1.4.2 Capacità ai Punti di Uscita della Rete Nazionale di Gasdotti 24
1.4.3 Capacità ai Punti di Riconsegna 25
1.4.4 Modalità di determinazione della Capacità di trasporto 25

1.5 OBIETTIVO DELLE ANALISI 27

**CAPITOLO 2: I DATI** **31**

2.1 IL DATASET 31

2.2 IL CONSUMO GIORNALIERO DI GAS 33

2.3 LE VARIABILI GEOGRAFICHE 35
2.7.1 La Provincia 35
2.7.2 La Regione 40

2.4 LE VARIABILI TEMPORALI 43
2.4.1 Il giorno della settimana 43
2.4.2 Il giorno del mese 46
2.4.3 Il mese 48
2.4.4 L'anno 50
2.4.5 Le festività 51

2.5 IL NUMERO DI CONTRATTI E LA CAPACITA' PRENOTATA 54

2.6 LA DESTINAZIONE D'USO DEL PUNTO DI RICONSEGNA 59

2.7 I DATI CLIMATICI 64
2.7.1 La zona climatica 64
2.7.2 Le informazioni atmosferiche 68

2.8 LA SUDDIVISIONE DEL DATASET: INSIEME DI STIMA E INSIEME DI VERIFICA 73

**CAPITOLO 3: IL *DATA MINING*** **75**

3.1 LA VARIABILE RISPOSTA 76

3.2 L'ANALISI DI CLASSIFICAZIONE 77
3.2.1 Il modello di regressione lineare 78
3.2.2 Il modello di regressione logistico 81
3.2.3 L'analisi discriminante lineare 83
3.2.4 Gli alberi di classificazione 84
3.2.5 Le Reti Neurali 88
3.2.6 Il *Bagging* 91
3.2.7 Il *Boosting* 93
3.2.8 Il confronto fra i modelli 95

3.3 L'ANALISI DI REGRESSIONE 97
3.3.1 Il modello di regressione lineare 97
3.3.2 I modelli non parametrici: *MARS* 102
3.3.3 I modelli non parametrici: *GAM* 107
3.3.4 I modelli non parametrici: *Projection Pursuit* 109
3.3.5 I modelli non parametrici: Reti Neurali 110
3.3.6 Il confronto fra i modelli 112

3.4 LA PREVISIONE DEL CONSUMO DI GAS 113

3.5 CONCLUSIONI 115

**CAPITOLO 4: L'ANALISI DELLE SERIE STORICHE 117**

4.1 L'ANALISI UNIVARIATA 117
4.1.1 Argelato (destinazione d'uso civile) 121
4.1.2 Tocco Da Casauria (destinazione d'uso civile e industriale) 128
4.1.3 Cartiera di Ferrara (destinazione d'uso industriale) 134

4.2 I MODELLI A FUNZIONE DI TRASFERIMENTO 139
4.2.1 Argelato (destinazione d'uso civile) 145
4.2.2 Tocco Da Casauria (destinazione d'uso civile e industriale) 154
4.2.3 Cartiera di Ferrara (destinazione d'uso industriale) 159

4.3 L'ANALISI MULTIVARIATA 165
4.3.1 La stima del modello 165
4.3.2 L'analisi strutturale del modello 182

**CAPITOLO 5: LA VALUTAZIONE DELLE PREVISIONI 189**

5.1 LA PREVISIONE COL *DATA MINING* 189

5.2 LA PREVISIONE CON LE SERIE STORICHE 191
5.2.1 I modelli di analisi univariata (*SARIMA*) 191
5.2.2 I modelli a funzione di trasferimento 197
5.2.3 I modelli di analisi multivariata (*VAR*) 200

5.3 CONCLUSIONI 206

**APPENDICE** **209**

**BIBLIOGRAFIA** **217**

**SITOGRAFIA** **218**

# INTRODUZIONE

Le risorse energetiche sono uno dei più importanti fattori che mantengono l'ordine economico e sociale in un Paese. In questo elaborato si tratterà l'argomento della previsione del consumo di risorse energetiche, in particolare del gas naturale; tali previsioni dovranno essere di buona qualità, in quanto la buona qualità delle previsioni del consumo di risorse energetiche permette lo sviluppo economico di un Paese e alti standard di vita per i suoi cittadini.

*La previsione del consumo di gas naturale è molto importante, soprattutto, per le aziende che operano nel settore del trasporto e della distribuzione; una buona previsione, infatti, consente l'ottimizzazione del portafoglio prodotti dal punto di vista della distribuzione, della vendita e dello stoccaggio, e permette, inoltre, una corretta programmazione delle azioni commerciali.*

In questo elaborato, l'orizzonte temporale di riferimento per la previsione del consumo di gas sarà il breve periodo (1-20 giorni), in quanto una corretta previsione nel breve periodo assicura il normale flusso energetico, e la puntuale ed efficiente distribuzione agli Utenti.

Il consumo di gas naturale può essere classificato in due categorie: il consumo civile e il consumo industriale: il consumo civile è caratterizzato dalle attività di riscaldamento delle abitazioni e degli edifici pubblici, e da tutte quelle attività domestiche che richiedono un apporto di gas, come l'utilizzo di acqua calda e la cucina, mentre il consumo industriale è caratterizzato, oltre che dalle attività di riscaldamento delle fabbriche e delle industrie, da tutte quelle attività che richiedono un apporto di gas all'interno dei processi industriali. Inoltre, è possibile osservare come il consumo civile risulti fortemente influenzato dalle condizioni atmosferiche (temperatura, umidità, … ), mentre il consumo industriale ne risulti molto meno condizionato.

Il presente elaborato sarà strutturato nel modo seguente:

- nel Capitolo 1, verrà presentato il settore del gas naturale in Italia e la sua regolamentazione, e verranno esposti i principali concetti legati al trasporto e all'approvvigionamento del gas naturale. Inoltre, verranno esposti in modo molto sintetico gli obiettivi delle analisi che si effettueranno nei capitoli successivi;

- nel Capitolo 2, verranno presentati i dati disponibili, attraverso opportune analisi descrittive delle variabili a disposizione per l'analisi;

- nel Capitolo 3, verranno stimati una serie di modelli che rientrano nella vasta categoria dei modelli di *Data Mining* (modelli di regressione lineare e metodi di stima non parametrica);

- nel Capitolo 4, i dati verranno trattati sotto forma di serie storiche, e verranno stimati modelli univariati, multivariati e modelli a funzione di trasferimento;

- nel Capitolo 5, infine, verranno riassunti i risultati ottenuti e ne verrà discussa la loro qualità.

# CAPITOLO 1: IL SETTORE DEL GAS E DELL'ENERGIA

## 1.1 ENERGIA E GAS

### *1.1.1 Il ruolo del gas nel contesto energetico italiano*

Il consumo interno lordo di energia in Italia ha registrato (*fonte: www.snamretegas.it* ) dal 2002 al 2009 una flessione dello 0,7% medio annuo passando da 188 Mtep (*Milioni di tonnellate equivalenti di petrolio*, ovvero la quantità di energia rilasciata dalla combustione di una tonnellata di petrolio grezzo) del 2002 a 179 Mtep del 2009 (TAVOLA 1). La contrazione dei consumi di Energia Primaria evidenzia gli effetti della Crisi Economica che, manifestatasi dalla seconda metà del 2008, ha investito interamente il 2009. Il tasso di crescita medio annuo, positivo fino al 2007 (+0,6%), ha subito un rallentamento nel 2008 (+0,25%), per poi invertire il segno nel 2009. Nonostante tale effetto, la quota di gas naturale sui consumi energetici del Paese è cresciuta, passando dal 31% del 2002 a circa il 36% del 2009, a scapito soprattutto dei consumi di prodotti petroliferi scesi dal 48,9% del 2002 a circa il 41% del 2009; il calo è evidente soprattutto nel settore della generazione elettrica, a causa dell'affermazione della tecnologia a ciclo combinato a gas, che coniuga alta efficienza ed emissioni più contenute. La produzione termoelettrica da gas naturale è passata, infatti, da 99 TWh nel 2002 a circa 146 TWh nel 2009, con un incremento totale del 47%. Il gas naturale ha rappresentato l'unica fonte combustibile fossile interessata dalla crescita, con un aumento medio annuo pari all'1,5%. Il tasso di crescita medio, pur rimanendo positivo ha subito tuttavia una contrazione rispetto al periodo precedente la crisi (+3,8%), a causa principalmente della contrazione dei consumi industriali.

| CONSUMI DI ENERGIA IN FONTI PRIMARIE IN ITALIA (in Milioni di tonnellate equivalenti di petrolio) | | | QUOTA DELLE FONTI DI ENERGIA | |
|---|---|---|---|---|
| | 2009 (stima) | 2002 (bilancio) | 2009 (stima) | 2002 (bilancio) |
| Consumo interno lordo totale | 179,5 | 188,1 | 100% | 100% |
| Combustibili solidi | 13,5 | 14,2 | 7,50% | 7,50% |
| **Gas Naturale** | **64** | **58,1** | **35,70%** | **31,00%** |
| Prodotti Petroliferi | 73,9 | 92 | 41,20% | 48,90% |
| Fonti Rinnovabili (1) | 18,3 | 12,6 | 10,20% | 6,70% |
| Importazioni Nette di Energia Elettrica | 9,8 | 11,1 | 5,40% | 5,90% |

(1) Comprende idroelettrico, eolico, fotovoltaico, geotermico, biogas

**TABELLA 1.1: Consumo di energia in fonti primarie e quota delle fonti di energia**

### *1.1.2 Il mercato del gas*

Il consumo di gas naturale dal 2002 al 2009 è passato da 70,5 miliardi di metri cubi a 78,1 miliardi (TAVOLA 2). La crescita maggiore si è registrata nel settore termoelettrico (+3,2% medio annuo), e nel settore residenziale e terziario (+3,3% medio annuo), poco sensibile alla crisi economica. Nello stesso periodo, il consumo relativo al settore industriale ha registrato una contrazione di circa 5,3 miliardi di metri cubi pari ad un decremento medio annuo del 3,9% circa, determinato principalmente dagli effetti della Crisi Economica in atto.

| | 2002 (bilancio) | 2009 (preconsuntivo) | 2009/2002 (tasso medio di crescita) |
|---|---|---|---|
| RESIDENZIALE E TERZIARIO | 25,4 | 31,8 | 3,3% |
| INDUSTRIA E ALTRI SETTORI (1) | 22,0 | 16,7 | -3,9% |
| TERMOELETTRICO | 22,6 | 28,2 | 3,2% |
| CONSUMI E PERDITE | 0,6 | 1,4 | 13,6% |
| **TOTALE CONSUMI ITALIA** | **70,5** | **78,1** | **1,5%** |

(1) Comprende anche Agricoltura, Autotrazione, Usi non energetici

**TABELLA 1.2: Consumi di gas naturale in Italia**

La crescita della domanda di gas naturale dal 2002 al 2009 è stata soddisfatta facendo ricorso in modo consistente alle importazioni, che sono cresciute da circa 59 a circa 69 miliardi di metri cubi. Il ruolo delle importazioni sulle disponibilità complessive nel periodo è così passato, al netto dello stoccaggio, dall' 80% al 90%.

## 1.2 IL CONTESTO NORMATIVO DI RIFERIMENTO

Il settore del gas naturale è stato oggetto di rilevante regolamentazione a livello nazionale e comunitario. In particolare, il processo di regolamentazione è stato avviato a livello europeo dalla Direttiva Gas (Direttiva 98/30 CE del Parlamento e del Consiglio Europeo del 22 giugno 1998), recante le norme comuni per il trasporto, la distribuzione, la fornitura e lo stoccaggio del gas naturale. La Direttiva Gas, è stata recepita in Italia nel maggio 2000 con il Decreto Legislativo 23 maggio 2000, n. 164 (Decreto Letta).

### *1.2.1 Il Decreto Legislativo 164/00 (Decreto Letta)*

Il Decreto Legislativo in oggetto ha introdotto norme che definiscono modalità e tempi del processo di liberalizzazione così come previsti dalla stessa Direttiva Gas, individuando e definendo i ruoli dei diversi segmenti della "catena" del gas naturale quali: importazione, coltivazione, esportazione, trasporto e dispacciamento, stoccaggio, distribuzione e vendita. Relativamente all'attività di trasporto, il Decreto Legislativo prevede tra l'altro:

1) la regolamentazione dell'attività di trasporto e dispacciamento, in modo da garantire che tali servizi siano offerti agli Utenti a tariffe regolamentate e a parità di condizioni;

2) la separazione societaria dell'attività di trasporto e dispacciamento da tutte le altre attività del settore del gas, ad eccezione dell'attività di stoccaggio, che è comunque oggetto di separazione contabile e gestionale dall'attività di trasporto e dispacciamento;
3) la definizione, con delibera da parte dell'Autorità per l'energia elettrica e il gas, dei "*criteri atti a garantire a tutti gli Utenti della rete la libertà di accesso a parità di condizioni, la massima imparzialità del trasporto e del dispacciamento, in condizioni di normale esercizio e gli obblighi dei soggetti che svolgono le attività di trasporto e dispacciamento del gas*";
4) l'adozione da parte delle imprese di gas naturale del proprio Codice di Rete – entro tre mesi dalla pubblicazione della delibera dell'Autorità per l'energia elettrica e il gas - che è trasmesso all'Autorità per la verifica di conformità ai suddetti criteri: trascorsi tre mesi senza comunicazioni da parte dell'Autorità per l'energia elettrica e il gas, il Codice di Rete si intende conforme.

Il Decreto attribuisce ruoli e responsabilità rilevanti al Ministero dello Sviluppo Economico ed all'Autorità per l'Energia Elettrica e il Gas. Il Ministero stabilisce le linee guida strategiche per il settore gas e garantisce la sicurezza e lo sviluppo economico del settore. L'Autorità è un organo governativo indipendente formato da un Presidente e da quattro membri che, nominati dal Consiglio dei Ministri, restano in carica per sette anni e non sono rieleggibili; tale organo è operativo dal 1997 ed è preposto alla regolamentazione dei mercati nazionali dell'energia elettrica e del gas naturale. Tra le sue funzioni vi sono la determinazione e l'aggiornamento delle tariffe, nonché la predisposizione delle regole per l'accesso alle infrastrutture e per l'erogazione dei servizi relativi alle attività di trasporto, di rigassificazione del GNL (con l'acronimo GNL si intende il Gas Naturale Liquefatto; si ottiene sottoponendo il gas naturale, dopo opportuni trattamenti di depurazione e disidratazione, a successive fasi di raffreddamento e condensazione. Il prodotto che ne deriva si presenta come un liquido inodore e trasparente costituito da una miscela composta prevalentemente da metano e

quantità minori di etano, propano, butano ed azoto, avente una temperatura di ebollizione di circa -160 °C a pressione atmosferica e di stoccaggio).

## 1.3 IL RUOLO DI SNAM RETE GAS

Snam Rete Gas è il principale operatore italiano di trasporto e dispacciamento di gas naturale sul territorio nazionale, disponendo della quasi totalità delle infrastrutture di trasporto in Italia, con oltre 31.000 km di gasdotti in alta e media pressione (circa il 96% dell'intero sistema di trasporto). La Società possiede l'unico impianto attualmente operativo in Italia per la rigassificazione del GNL attraverso il quale viene importato Gas Naturale Liquefatto trasportato da navi metaniere.

### *1.3.1 I clienti*

I clienti di Snam Rete Gas sono gli *Shipper* (termine per indicare gli operatori che utilizzano le reti per trasportare il proprio gas ai punti finali), che prenotano capacità nel sistema di trasporto di Snam Rete Gas per destinare quantitativi di gas immessi a proprio titolo in base alle loro esigenze. Snam Rete Gas, quindi, trasporta il gas per conto degli *Shipper*, consegnandolo sulla base delle loro istruzioni. Gli *Shipper* producono o importano gas, oppure lo acquistano da produttori nazionali o da altri *Shipper*, per rivenderlo ai clienti finali (industrie e centrali termoelettriche), grossisti, o altri *Shipper*.

### *1.3.2 La capacità di trasporto*

Il trasporto del gas naturale è un servizio integrato che consente la movimentazione del gas a partire dai punti di entrata nella Rete Nazionale fino ai

punti di riconsegna della Rete Regionale per conto del soggetto che lo ha immesso. Snam Rete Gas conferisce capacità di trasporto agli *Shipper* che ne fanno richiesta, i quali acquisiscono il diritto di immettere e ritirare, in qualsiasi giorno dell'Anno Termico (è il periodo che và dall'1 Ottobre al 30 Settembre dell'anno successivo, ed è il periodo che viene preso in considerazione nel settore del gas e dell'energia per tutte le analisi) rispettivamente ai punti di entrata e di uscita della Rete Nazionale, ai punti di riconsegna sulla Rete Regionale e al Punto di Scambio Virtuale (Punto Virtuale situato tra i punti di entrata e di uscita della Rete Nazionale Gasdotti presso il quale gli Utenti possono effettuare, su base giornaliera, scambi e cessioni di gas, immesso nella Rete Nazionale) un quantitativo di gas non superiore alla portata giornaliera conferita.

Il gas immesso nella Rete Nazionale Gasdotti proviene da importazioni e, in minor quantità, da produzione nazionale. Il gas naturale proveniente dall'estero viene immesso nella Rete Nazionale di Gasdotti attraverso 6 punti di entrata in corrispondenza delle interconnessioni con i metanodotti di importazione (Tarvisio, Gorizia, Passo Gries, Mazara del Vallo, Gela) e del terminale di rigassificazione GNL di Panigaglia. Il gas di produzione nazionale viene immesso in corrispondenza dei 67 punti di entrata dai campi di produzione o dai loro centri di raccolta e trattamento. Anche i campi di stoccaggio gas sono collegati alla rete di trasporto (2 punti virtuali di entrata). I punti di uscita dalla Rete Nazionale di Gasdotti sono costituiti da 17 aree di prelievo (ossia aggregazioni territoriali di punti di riconsegna), coincidenti generalmente con i confini amministrativo-regionali, da 5 punti di interconnessione con i gasdotti internazionali per le esportazioni (Tarvisio, Gorizia, Passo Gries, Bizzarrone, Repubblica di San Marino) e da due punti di uscita verso gli "hub" di stoccaggio. Il gas in uscita dalla Rete Nazionale di gasdotti viene trasportato sulla Rete Regionale fino ai punti di riconsegna, presso i quali avviene il ritiro del gas da parte degli Utenti e la sua misurazione.

### *1.3.3 La struttura del settore*

Il diagramma illustra la struttura dei principali segmenti del settore del gas naturale in Italia.

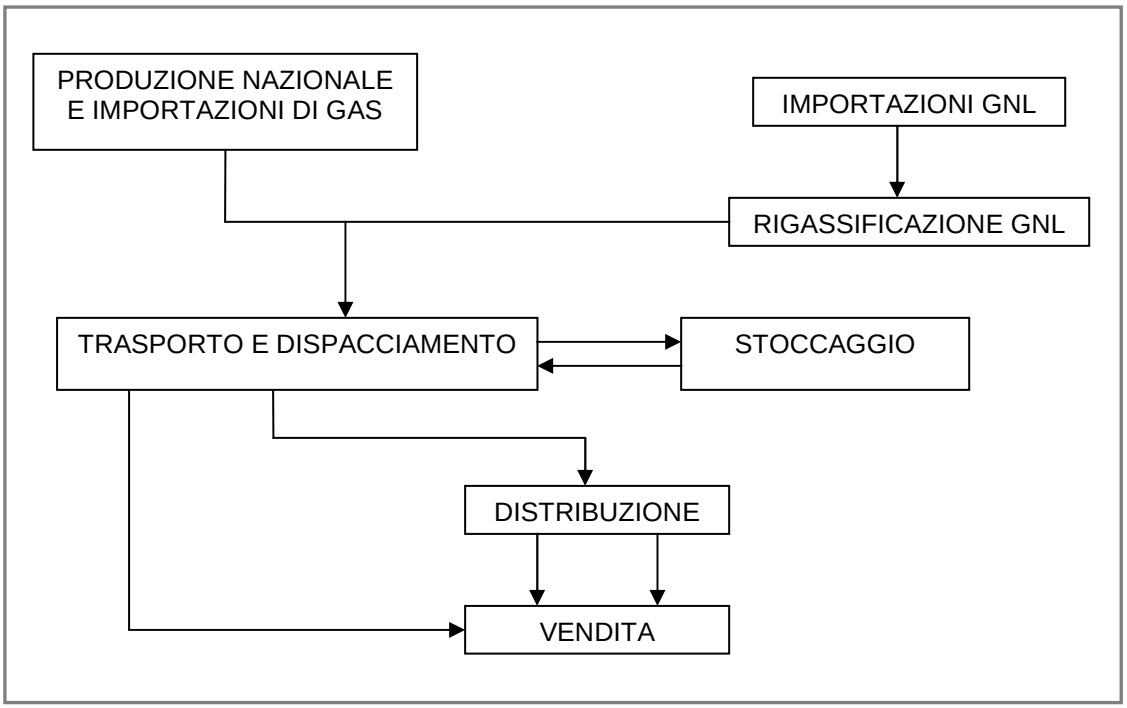



**GRAFICO 1.1: Il settore del gas**

- PRODUZIONE NAZIONALE E IMPORTAZIONI DI GAS: l'approvvigionamento di gas naturale in Italia avviene attraverso lo sfruttamento di riserve nazionali o l'importazione da paesi stranieri produttori di gas. L'attività di importazione include il trasporto di gas attraverso gasdotti internazionali o con navi metaniere che trasportano gas naturale liquefatto (GNL).
- IMPORTAZIONI GNL: nei paesi produttori il gas viene liquefatto in appositi impianti e caricato su navi metaniere che lo trasportano negli impianti di rigassificazione dove il gas torna al suo stato gassoso. In Italia

attualmente esiste un solo impianto, quello di Panigaglia, che è interconnesso alla rete nazionale di gasdotti.

- RIGASSIFICAZIONE DEL GNL: l'attività di rigassificazione del GNL comprende la discarica del GNL, lo stoccaggio, la sua rigassificazione e l'immissione nel sistema di trasporto presso il punto di entrata di Panigaglia. In base al Decreto Letta, l'attività di rigassificazione (come pure il trasporto e il dispacciamento) sono dichiarate di pubblico interesse e sono regolamentate.
- TRASPORTO E DISPACCIAMENTO: una volta prodotto o importato, il gas viene introdotto nel sistema di condotte ad alta e media pressione distribuite sul territorio italiano, che trasportano il gas sino a grandi clienti finali quali centrali termoelettriche e impianti industriali, direttamente collegati a tali condotte, o sino all'imbocco delle reti delle società di distribuzione locale.
- STOCCAGGIO: parte del gas naturale importato o prodotto viene destinato a riserve subito disponibili per fare fronte al consumo invernale (circa tre volte quello estivo) e alla domanda di gas in condizioni eccezionali. Lo stoccaggio deve poter soddisfare la maggiore domanda causata da un freddo anche eccezionale e deve essere in grado di integrare eventuali temporanee riduzioni delle importazioni del gas causate da motivi tecnico-commerciali o crisi nel settore del gas naturale. Anche le attività di stoccaggio sono regolamentate e soggette a concessione da parte del Ministero.
- DISTRIBUZIONE: le società di distribuzione, pubbliche e private, trasportano il gas utilizzando le condotte locali in media e bassa pressione, che coprono principalmente le aree urbane. Ai sensi del Decreto Letta, la distribuzione del gas è un servizio pubblico, assegnato alle società mediante gare d'appalto per periodi della durata massima di dodici anni.
- VENDITA: gli *Shipper*, una volta acquistato il gas da produttori, da importatori o da altri *Shipper*, vendono il gas direttamente a clienti finali, oppure ad altri venditori che provvedono a fornire il gas ai clienti finali,

direttamente o tramite l'utilizzo delle reti di distribuzione. In Italia la vendita di gas naturale è libera. Ai sensi del Decreto Letta, l'attività di vendita al dettaglio, ossia ai clienti finali, può essere effettuata dietro autorizzazione rilasciata dal Ministero.

## 1.4 IL CONCETTO DI CAPACITA'

Ora risulta opportuno descrivere le prestazioni della rete in situazioni di esercizio normale e speciale, e vengono descritte la modalità con cui tali prestazioni sono determinate, tenuto conto dei vincoli tecnici e gestionali esistenti.

*La capacità di trasporto è la massima quantità di gas che può essere immessa nel sistema (o prelevata da esso), nel corso del Giorno-gas, in uno specifico punto, nel rispetto dei vincoli tecnici e gestionali stabiliti in ciascuna sezione delle condotte e delle prestazioni massime degli impianti collocati lungo le stesse. La valutazione di tali capacità è effettuata mediante simulazioni idrauliche della rete, eseguite in scenari di trasporto appropriati e secondo standard tecnici riconosciuti.*

### *1.4.1 Capacità ai Punti di Entrata della Rete Nazionale di Gasdotti*

La capacità di trasporto presso i Punti di Entrata interconnessi con l'estero è la massima capacità che può essere messa a disposizione degli Utenti per il servizio di trasporto, di tipo continuo o interrompibile. La capacità di trasporto presso i Punti di Entrata da produzione nazionale e gli stoccaggi è la portata giornaliera di gas che il sistema di trasporto è in grado di ricevere e trasportare fino ai Punti di Riconsegna, in base alle verifiche tecniche effettuate dal Trasportatore. Dal momento che la capacità di trasporto in un punto è strettamente dipendente dalle capacità dei punti di immissione e prelievo attigui, non è possibile definire un

valore univoco di capacità massima caratteristico di ciascun Punto di Entrata da produzione nazionale o da stoccaggi: ciò a maggior ragione nel caso di reti magliate, qual è la rete di Snam Rete Gas. I valori pubblicati sono pertanto da intendersi come "valori di riferimento". Per tale motivo il Trasportatore è disponibile a rivedere al rialzo, previa ulteriore verifica tecnica, le capacità di trasporto presso i Punti di Entrata da produzione nazionale o da stoccaggi in funzione delle richieste effettuate dagli Utenti della rete in specifici punti della rete.

### *1.4.2 Capacità ai Punti di Uscita della Rete Nazionale di Gasdotti*

La capacità di trasporto presso i Punti di Uscita da esportazione è la massima capacità che può essere messa a disposizione degli Utenti nel Giorno-gas per il servizio di trasporto di tipo continuo. La capacità di trasporto presso i Punti di Uscita verso stoccaggio è la portata giornaliera di gas che il sistema di trasporto è in grado di ricevere e trasportare fino a tali Punti di Uscita, in base alle verifiche tecniche effettuate dal Trasportatore. Dal momento che la capacità di trasporto in un punto è strettamente dipendente dalle capacità dei punti di immissione e prelievo attigui, non è possibile definire un valore univoco di capacità massima caratteristico di ciascun Punto di Uscita verso stoccaggio: ciò a maggior ragione nel caso di reti magliate, qual è la rete di Snam Rete Gas. I valori pubblicati sono pertanto da intendersi, anche in questo caso, come "valori di riferimento". Per tale motivo il Trasportatore è disponibile a rivedere al rialzo, previa ulteriore verifica tecnica, le capacità di trasporto presso i Punti di Uscita verso stoccaggio in funzione delle richieste effettuate dagli Utenti della rete in specifici punti della rete. La capacità di trasporto in uscita per le Aree di Prelievo, in coerenza con la definizione delle stesse, è definita come sommatoria delle capacità dei Punti di Riconsegna afferenti a ciascuna di dette Aree.

### *1.4.3 Capacità ai Punti di Riconsegna*

La capacità di trasporto ai Punti di Riconsegna rappresenta la portata giornaliera di gas di cui può essere assicurata la riconsegna, in base alle verifiche tecniche effettuate. Anche in questo caso la capacità di trasporto in un punto è strettamente dipendente dalle capacità dei punti attigui: non è perciò possibile definire un valore univoco di capacità massima caratteristico di un Punto di Riconsegna (a maggior ragione per la rete magliata del Trasportatore). I valori pubblicati sul sito Internet sono da intendersi quindi come "valori di riferimento". Snam Rete Gas assicura la propria disponibilità a rivedere al rialzo, previa ulteriore verifica tecnica, le capacità di trasporto presso i Punti di Riconsegna in funzione delle richieste effettuate dagli Utenti della rete in specifici punti. Per Punti di Riconsegna costituiti dall'aggregato di punti fisici interconnessi a valle dalla rete di distribuzione, la capacità di trasporto pubblicata deriva dalla sommatoria delle capacità di trasporto dei singoli punti fisici. I valori di capacità di trasporto sono definiti considerando le prestazioni della rete, a prescindere dalla potenzialità degli impianti REMI (Impianti di Regolazione e Misurazione della capacità) che non fanno parte del sistema Snam Rete Gas. Pertanto in alcuni casi potrebbe verificarsi che gli impianti a valle non siano adeguati alle prestazioni indicate per il relativo Punto di Riconsegna.

### *1.4.4 Modalità di determinazione delle capacità di trasporto*

Le capacità di trasporto ai Punti di Entrata interconnessi con l'estero e presso i Punti in Uscita da esportazione vengono determinate mediante simulazioni idrauliche di trasporto, utilizzando criteri differenti per le capacità di tipo continuo e per quelle di tipo interrompibile. La valutazione delle capacità di trasporto di tipo continuo, la cui disponibilità deve essere garantita in ogni situazione ed in ogni periodo dell'Anno Termico, oltre che ai vincoli gestionali fa riferimento anche ai vincoli tecnici più gravosi: in particolare, per quanto

riguarda gli scenari di trasporto, si considerano le condizioni di prelievo più severe, prevedibili nel corso dell'Anno Termico, per gli Utenti collocati sulla rete (condizioni di esercizio speciali). La valutazione circa le capacità di tipo interrompibile, a parità di vincoli gestionali, sfrutta invece i margini di trasporto esistenti con vincoli tecnici meno severi (condizioni di esercizio normali). Le capacità di trasporto presso i Punti di Entrata da produzione nazionale sono determinate sulla base di uno scenario di immissioni previste, che deriva dalle portate utilizzate negli anni termici precedenti e dalle previsioni di immissione fornite dagli operatori dei campi di produzione. La verifica di tali portate viene effettuata a mezzo di simulazioni idrauliche che considerano lo scenario più gravoso relativamente ai prelievi del mercato. Le capacità di trasporto ai Punti di Entrata da stoccaggio ed ai Punti di Uscita verso stoccaggio sono determinate sulla base di uno scenario di immissioni e prelievi previsti sulla RN. Le portate immesse da/erogate verso ciascun stoccaggio sono valutate sulla base delle prestazioni massime conosciute e ad una distribuzione gravosa delle portate tra gli stoccaggi appartenenti a ciascun "pool". La verifica di tali portate viene effettuata mediante simulazioni idrauliche che tengono in considerazione differenti scenari possibili di prelievi del mercato. La capacità di trasporto ai Punti di Riconsegna è individuata sulla base di verifiche idrauliche che si basano su scenari di fabbisogno di capacità dell'area geografica interessata e che derivano dai dati storici disponibili e da eventuali contatti con i Clienti Finali (utenze industriali ed Imprese di Distribuzione). Tali capacità possono essere aggiornate, previa verifica tecnica di trasportabilità, sulla base degli incrementi richiesti dagli Utenti, in corrispondenza dell'inizio di un nuovo Anno Termico o mensilmente, nel caso di Anno Termico avviato.

Relativamente ai Punti di Riconsegna occorre sottolineare che:

1) i valori di capacità di trasporto pubblicati sul sito Internet di Snam Rete Gas sono espressi in metri cubi/giorno; le verifiche di rete tengono conto invece di portate "di picco" espresse in metri cubi/ora. La conversione delle portate

giornaliere (di riferimento e conferite) in metri cubi/ora viene fatta avendo analizzato, per ogni Punto di Riconsegna telemisurato, i dati storici relativi alla sua profilatura oraria di prelievo, così da determinare il legame statistico tra portata giornaliera e punta oraria massima associata. Ai Punti di Riconsegna per i quali tali dati non sono disponibili (ad es. nel caso di impianti non teleletti), è stato utilizzato un procedimento analogo a quello di altri Punti appartenenti al medesimo settore merceologico (siti industriali) o appartenenti alla medesima area climatica (siti civili);

2) per i Punti di Riconsegna costituiti dall'aggregato di punti fisici interconnessi a valle dalla rete di distribuzione, la capacità di trasporto pubblicata deriva dalla sommatoria delle capacità di trasporto dei singoli punti fisici. Ciò significa anche che, a fronte di incrementi di capacità richiesti dagli Utenti in aggregato su un Punto di Riconsegna, il Trasportatore deve, ai fini delle verifiche, oltre a convertire la richiesta in punta oraria, suddividere la stessa sui singoli punti fisici.

## 1.5 OBIETTIVO DELLE ANALISI

L'obiettivo delle analisi che si effettueranno in questo elaborato è fornire a *Main Consulting*, azienda che si occupa dello sviluppo di algoritmi e software per la gestione dei consumi di gas e del *Gas Shipping*, la previsione del consumo giornaliero nei vari punti di riconsegna nel breve periodo (1-20 giorni), in modo che da poter valutare la Capacità da prenotare da Snam Rete Gas nella rete di distribuzione. In questo modo, si aumenta l'efficienza, in quanto migliora l'ottimizzazione del portafoglio prodotti dal punto di vista della distribuzione, della vendita e dello stoccaggio, e, di conseguenza, migliora la programmazione delle azioni commerciali.

Inoltre, l'azienda cerca, grazie all'attendibilità delle previsioni, di ridurre al minimo le possibilità di incorrere nel pagamento delle penali a Snam Rete Gas,

in quanto, se la quantità di gas distribuita da Snam Rete Gas a uno *Shipper* eccede la capacità prenotata dallo *Shipper* stesso, si incorre in una sanzione pecuniaria, calcolata in relazione all'eccesso di Capacità consegnata, rispetto a quella prenotata.

# CAPITOLO 2: I DATI

Prima di passare alla costruzione di modelli che permettano di prevedere il consumo di gas giornaliero in ciascun punto di riconsegna, è necessario analizzare i dati a disposizione, su cui si effettueranno le analisi. Si analizzeranno, quindi, soprattutto per mezzo di supporti grafici, le variabili a disposizione e tutte quelle create per rendere più complete possibili le analisi: il consumo giornaliero di gas, le variabili geografiche (provincia e regione), le variabili temporali (giorno della settimana, giorno del mese, mese, anno, festività), la capacità prenotata e il numero dei contratti, la destinazione d'uso, e le variabili climatiche (zona climatica e dati atmosferici).

## 2.1 IL DATASET

Il dataset è composto da oltre 340.000 osservazioni relative al consumo giornaliero di gas, rilevato su 897 punti di riconsegna collocati in tutta Italia. Il periodo di osservazione non è lo stesso per tutti i punti, ma va da un minimo di 30 giorni a un massimo di 577 giorni, con una media di giorni osservati nel complesso dei punti di circa 383 giorni. Anche le date di rilevazione di tali informazioni non sono omogenee per tutti i punti, in quanto le osservazioni di un punto iniziano nel momento in cui l'azienda stipula un contratto con un Utente e si concludono nel momento in cui tale contratto si esaurisce. Il lasso di tempo all'interno del quale sono state rilevate le informazioni presenti nel nostro dataset va dall'1 ottobre 2009 (giorno di inizio dell'anno termico 2009/2010) fino al 30/04/2011, in un periodo dunque inferiore ai due anni.

Il dataset presenta la seguente struttura:

| PUNTO | DESCRIZIONE | DATA | gJ | sm3 | m3 |
|---|---|---|---|---|---|
| 30009501 | ALESSANDRIA (2) | 01/05/2011 | -9,6 | -251,969 | -251,969 |
| 30009501 | ALESSANDRIA (2) | 02/05/2011 | -44,1 | -1157,480 | -1157,480 |
| 30009501 | ALESSANDRIA (2) | 03/05/2011 | -42,7 | -1120,735 | -1120,735 |
| 30009501 | ALESSANDRIA (2) | 04/05/2011 | -43,5 | -1141,732 | -1141,732 |
| 30009501 | ALESSANDRIA (2) | 05/05/2011 | -43,5 | -1141,732 | -1141,732 |
| 30009501 | ALESSANDRIA (2) | 06/05/2011 | -36 | -944,882 | -944,882 |
| 30009501 | ALESSANDRIA (2) | 07/05/2011 | 0 | 0 | 0 |
| 30009501 | ALESSANDRIA (2) | 08/05/2011 | -9,1 | -238,845 | -238,845 |
| 30009501 | ALESSANDRIA (2) | 09/05/2011 | -44,5 | -1167,979 | -1167,979 |
| 30009501 | ALESSANDRIA (2) | 10/05/2011 | -42,8 | -1123,360 | -1123,360 |
| 30009501 | ALESSANDRIA (2) | 11/05/2011 | -25,9 | -679,790 | -679,790 |
| 30009501 | ALESSANDRIA (2) | 12/05/2011 | -34 | -892,388 | -892,388 |
| 30009501 | ALESSANDRIA (2) | 13/05/2011 | 0 | 0 | 0 |
| ……. | ……... | …… | …… | ……. | ……... |

**TABELLA 2.1: La struttura del dataset**

Ogni punto di riconsegna è caratterizzato da un codice identificativo (PUNTO) e da una descrizione, in cui viene rilevato il comune in cui è situato il punto o l'industria a cui il punto fornisce il gas (DESCRIZIONE); vengono poi fornite 3 misure equivalenti della quantità di gas consumata in un determinato giorno (DATA): i giga Joule (gJ), unità di misura dell'energia, i metri cubi (m3), ovvero il volume di gas racchiuso da un cubo avente gli spigoli di 1 metro, e i metri cubi standard (sm3), ovvero la quantità di gas necessaria ad occupare 1 metro cubo di volume a 15 °C di temperatura e 1,01325 bar assoluti di pressione.

Per le analisi si utilizzerà, in accordo con l'azienda, il consumo di gas in metri cubi.

Nel dataset a disposizione non sono state rilevate altre variabili; risulta quindi necessario ricercare autonomamente ulteriori informazioni al fine di costruire modelli statistici utili alla previsione del consumo di gas.

## 2.2 IL CONSUMO GIORNALIERO DI GAS

Il consumo giornaliero di gas rilevato sugli 897 punti di riconsegna non è costante nell'arco del periodo di osservazione; indipendentemente dalla diversa distribuzione temporale dei consumi di ciascun punto, è possibile osservare la presenza di un certo andamento stagionale dei consumi: è lecito supporre, infatti, che il consumo di gas, in quanto strettamente correlato alle condizioni climatiche, risulta maggiore nel periodo invernale rispetto a quello estivo. E' possibile osservare, inoltre, una certa stagionalità settimanale: questo è principalmente dovuto al fatto che i punti di riconsegna forniscono gas anche a utenze industriali che nel fine settimana restano inattive e, di conseguenza, non consumano alcun quantitativo di gas.

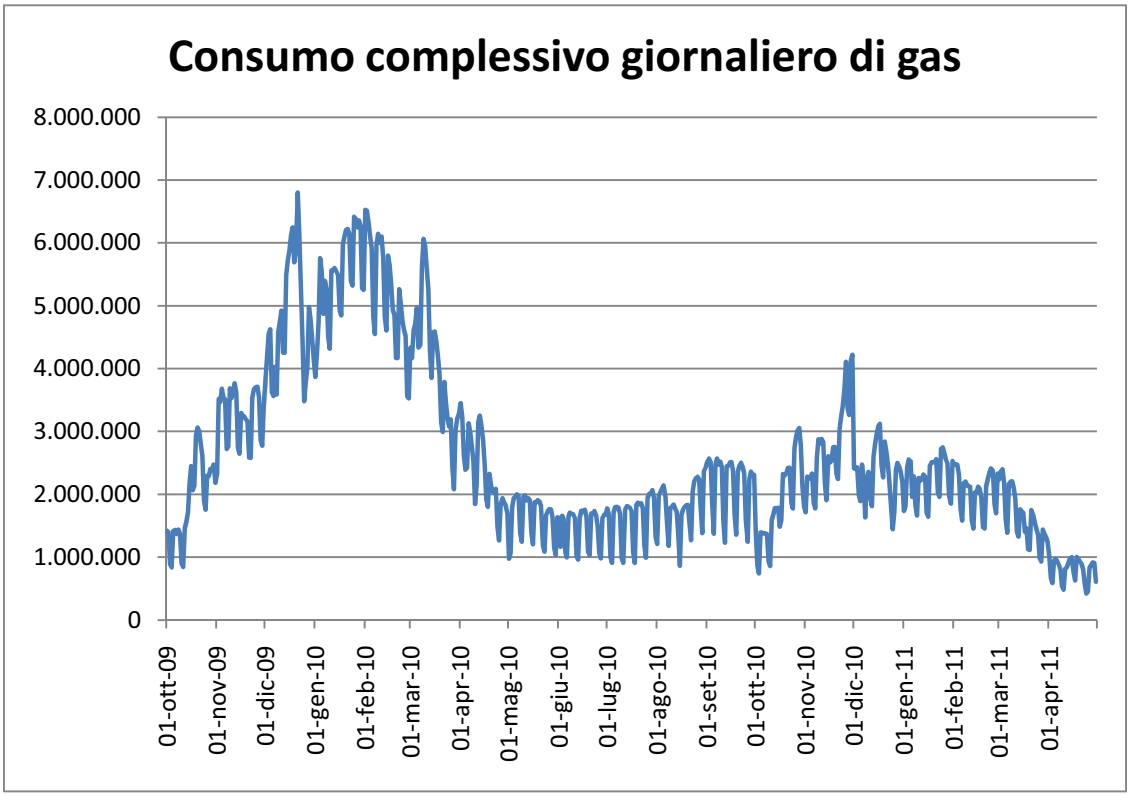



**GRAFICO 2.1: Consumo complessivo giornaliero di gas**

Il Grafico 2.1 evidenzia, oltre alle deduzioni precedenti, che la quantità di gas consumata sembra diminuire tra l'inverno 2009-2010 e l'inverno 2010-2011; una delle possibili cause di questa diminuzione può essere ricondotta alla crisi economica che negli ultimi anni ha colpito l'Italia, e che ha portato alla chiusura di numerose industrie e aziende.

Appare interessante, inoltre, analizzare anche la distribuzione della quantità di gas consumata giornalmente nei punti di riconsegna. La Tabella 2.2 ci mostra le principali misure della distribuzione:

| | |
|---|---|
| MEDIA | 4.324 |
| VALORE MINIMO | 0 |
| 1° QUARTILE | 17 |
| MEDIANA | 272 |
| 3° QUARTILE | 1.767 |
| VALORE MASSIMO | 1.467.752 |
| DEVIAZIONE STANDARD | 24.070 |
| INDICE DI ASIMMETRIA DI FISHER-PEARSON | 25 |
| INDICE DI CURTOSI | 939 |

**TABELLA 2.2: Statistiche descrittive "Quantità di gas complessiva giornaliera"**

La prima considerazione che suggerisce la Tabella 2.2 riguarda la notevole differenza tra la media e la mediana: questo fatto evidenzia che la distribuzione è fortemente asimmetrica, e questo è probabilmente dovuto alla presenza di alcuni consumi giornalieri molto elevati (il valore massimo della distribuzione è addirittura superiore al milione di metri cubi giornaliero, probabilmente registrato in una grande città o in una vasta area industriale).

Il valore minimo della distribuzione è pari a 0, ed è un valore plausibile, in quanto è lecito supporre, ad esempio, che in un punto di riconsegna esclusivamente civile non sia registrato alcun consumo di gas in un giorno estivo. La deviazione standard risulta molto elevata e gli indici di asimmetria e curtosi confermano che la distribuzione del consumo giornaliero di gas è molto distante da una distribuzione normale; queste considerazioni suggeriscono di costruire, al momento della stima di un modello, un'adeguata trasformazione della variabile.

## 2.3 LE VARIABILI GEOGRAFICHE

Per ciascun punto di riconsegna è possibile risalire, attraverso la descrizione, alla provincia e alla regione in cui è localizzato; questa rilevazione può essere utile al fine di ricercare eventuali differenze di consumo dovute alla localizzazione geografica dei punti di riconsegna.

### *2.3.1 La Provincia*

I punti di riconsegna che compongono il dataset provengono da 90 province italiane, collocate in modo abbastanza omogeneo sull'intero territorio nazionale. La Tabella 2.3 riporta il numero di punti di riconsegna presenti in ciascuna provincia.

| N. | PROV | PUNTI |
|---|---|---|
| 1 | AV | 48 |
| 2 | MI | 48 |
| 3 | VI | 47 |
| 4 | VR | 46 |
| 5 | PD | 34 |
| 6 | CR | 29 |

| N. | PROV | PUNTI |
|---|---|---|
| 31 | TO | 10 |
| 32 | AL | 9 |
| 33 | AQ | 9 |
| 34 | TN | 9 |
| 35 | GO | 8 |
| 36 | PG | 8 |

| N. | PROV | PUNTI |
|---|---|---|
| 61 | TE | 4 |
| 62 | AG | 3 |
| 63 | BA | 3 |
| 64 | BR | 3 |
| 65 | FC | 3 |
| 66 | MC | 3 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 7 | SA | 27 | 37 | LE | 7 | 67 | MT | 3 |
| 8 | BG | 26 | 38 | LC | 7 | 68 | PT | 3 |
| 9 | PC | 26 | 39 | LO | 7 | 69 | RG | 3 |
| 10 | PR | 23 | 40 | NA | 7 | 70 | RA | 3 |
| 11 | CS | 22 | 41 | PE | 7 | 71 | VT | 3 |
| 12 | FE | 22 | 42 | RE | 7 | 72 | FG | 2 |
| 13 | PN | 21 | 43 | BI | 6 | 73 | IM | 2 |
| 14 | UD | 21 | 44 | CE | 6 | 74 | SP | 2 |
| 15 | VA | 20 | 45 | FI | 6 | 75 | LT | 2 |
| 16 | MN | 19 | 46 | VC | 6 | 76 | RI | 2 |
| 17 | PV | 16 | 47 | MC | 5 | 77 | RN | 2 |
| 18 | RO | 15 | 48 | SI | 5 | 78 | TR | 2 |
| 19 | TV | 15 | 49 | AR | 4 | 79 | TS | 2 |
| 20 | BO | 14 | 50 | AT | 4 | 80 | AO | 1 |
| 21 | VE | 14 | 51 | BL | 4 | 81 | AP | 1 |
| 22 | CO | 13 | 52 | BZ | 4 | 82 | BAT | 1 |
| 23 | MB | 13 | 53 | CT | 4 | 83 | BN | 1 |
| 24 | VB | 13 | 54 | CH | 4 | 84 | CL | 1 |
| 25 | MO | 12 | 55 | FM | 4 | 85 | GE | 1 |
| 26 | NO | 12 | 56 | LU | 4 | 86 | GR | 1 |
| 27 | BS | 11 | 57 | ME | 4 | 87 | IS | 1 |
| 28 | CN | 11 | 58 | PU | 4 | 88 | LI | 1 |
| 29 | RM | 11 | 59 | PI | 4 | 89 | PA | 1 |
| 30 | PZ | 10 | 60 | PO | 4 | 90 | TP | 1 |

**TABELLA 2.3: Punti di riconsegna presenti in ciascuna provincia**

Ci sono alcune province particolarmente rappresentate, mentre altre lo sono meno. Le province più presenti sono Avellino (48), Milano (48), Vicenza (47) e Verona (46); queste province contengono numerosi punti di riconsegna in quanto particolarmente vaste e ricche di aziende e industrie.

Considerazioni diverse, invece, suggerisce la Tabella 2.4, che riporta il consumo medio di gas per ogni provincia (dobbiamo comunque ricordare che tali dati sono

condizionati dalla distribuzione non omogenea dei punti di riconsegna sul territorio e dalla differente dimensione temporale delle rilevazioni): le province maggiormente industrializzate, infatti, passano in secondo piano e sono superate dalle province che presentano un minor numero di punti di riconsegna.

| N. | PROV | CONS |
|---|---|---|
| 1 | GO | 26.216 |
| 2 | FM | 18.926 |
| 3 | CN | 17.744 |
| 4 | AR | 16.070 |
| 5 | PI | 10.574 |
| 6 | TS | 10.385 |
| 7 | VI | 9.912 |
| 8 | FE | 8.668 |
| 9 | MI | 7.976 |
| 10 | SA | 6.716 |
| 11 | UD | 6.484 |
| 12 | PC | 6.465 |
| 13 | BS | 6.411 |
| 14 | PT | 5.992 |
| 15 | PN | 5.952 |
| 16 | TO | 5.942 |
| 17 | VR | 5.601 |
| 18 | CT | 5.566 |
| 19 | LC | 4.802 |
| 20 | SI | 4.759 |
| 21 | BR | 4.346 |
| 22 | VC | 4.334 |
| 23 | PO | 4.187 |
| 24 | TV | 4.096 |
| 25 | RM | 3.771 |
| 26 | AV | 3.613 |
| 27 | PZ | 3.391 |
| 28 | BZ | 3.345 |

| N. | PROV | CONS |
|---|---|---|
| 31 | BL | 3.127 |
| 32 | TN | 2.918 |
| 33 | FG | 2.766 |
| 34 | MN | 2.633 |
| 35 | AT | 2.596 |
| 36 | BO | 2.425 |
| 37 | CL | 2.189 |
| 38 | PE | 2.029 |
| 39 | CR | 1.804 |
| 40 | VE | 1.731 |
| 41 | MO | 1.689 |
| 42 | PD | 1.545 |
| 43 | MB | 1.448 |
| 44 | RI | 1.435 |
| 45 | BG | 1.402 |
| 46 | PR | 1.393 |
| 47 | LO | 1.355 |
| 48 | RE | 1.344 |
| 49 | MC | 1.292 |
| 50 | TE | 1.236 |
| 51 | FI | 1.230 |
| 52 | CH | 1.168 |
| 53 | ME | 1.167 |
| 54 | VA | 1.104 |
| 55 | AQ | 959 |
| 56 | AP | 888 |
| 57 | AG | 765 |
| 58 | BA | 765 |

| N. | PROV | CONS |
|---|---|---|
| 61 | AL | 582 |
| 62 | VB | 573 |
| 63 | NO | 566 |
| 64 | PU | 555 |
| 65 | RO | 536 |
| 66 | CE | 535 |
| 67 | LU | 455 |
| 68 | PG | 438 |
| 69 | AO | 436 |
| 70 | CS | 419 |
| 71 | RA | 283 |
| 72 | GE | 203 |
| 73 | PV | 142 |
| 74 | RG | 133 |
| 75 | VT | 129 |
| 76 | BN | 128 |
| 77 | FC | 72 |
| 78 | SP | 71 |
| 79 | IS | 59 |
| 80 | RN | 47 |
| 81 | MC | 28 |
| 82 | GR | 25 |
| 83 | IM | 14 |
| 84 | LT | 10 |
| 85 | TR | 7 |
| 86 | BAT | 3 |
| 87 | LI | 3 |
| 88 | MT | 3 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 29 | BI | 3.267 | 59 | LE | 652 | 89 | TP | 3 |
| 30 | NA | 3.261 | 60 | CO | 599 | 90 | PA | 0 |

**TABELLA 2.4: Consumo medio di gas per ciascuna provincia**

E' possibile osservare come la provincia sia una variabile di tipo fattoriale contenente un numero abbastanza elevato (80) di livelli; l'inserimento di una variabile di questo tipo in un modello di regressione potrebbe quindi portare a una certa inefficienza delle stime e, quindi, delle previsioni. Si può quindi pensare di raggruppare tra loro le province simili, ovvero quelle che non sembrano presentare un consumo di gas differente tra loro, basandosi sulle stime dei coefficienti in un modello di regressione lineare semplice avente come variabile risposta il consumo giornaliero di gas e come variabile risposta la provincia.

La classificazione che ne risulta è la seguente:

| | |
|---|---|
| 1 | Alessandria, Bergamo, Biella, Brindisi, Chieti, Cosenza, Genova, Lecce, Macerata, Mantova, Reggio Emilia, Teramo |
| 2 | Agrigento, Aosta, Napoli, Como, Ragusa, Rovigo, Verbano-Cusio-Ossola, Viterbo |
| 3 | Asti, Avellino, Catania, Lecco, Milano, Messina, Pescara, Pistoia, Trento, Treviso, Trieste, Udine, Vercelli, Verona, Prato, Siena |
| 4 | Bari, Caserta, Latina, Pavia, Ravenna, Rieti, Terni, Massa Carrara |
| 5 | Barletta-Andria-Trani, Foggia, Imperia, La Spezia, Matera, Rimini |
| 6 | Belluno, Cuneo, Gorizia, Piacenza, Pisa, Pordenone, Salerno |
| 7 | Bologna, Bolzano, Roma, Venezia |
| 8 | Brescia, Ferrara, Firenze, Lucca, Monza e Brianza, Potenza |

| | |
|---|---|
| 9 | L'Aquila, Lodi, Novara, Padova, Parma, Pesaro e Urbino, Torino, Vicenza |
| 10 | Cremona, Varese |
| 11 | Fermo |
| 12 | Latina |
| 13 | Modena |
| 14 | Perugia |

**TABELLA 2.5: Classificazione delle Province**

Dopo aver stimato un modello di regressione lineare semplice, avente come variabile risposta il consumo giornaliero di gas e come variabile esplicativa la provincia raggruppata secondo la Tabella 2.5, si valuta l'importanza della variabile provincia attraverso il test ANOVA, che confronta il modello stimato con il modello avente come unica variabile esplicativa l'intercetta:

```
Analysis of Variance Table

              Df Sum Sq Mean Sq F value    Pr(>F)
Provincia     13 223062 17158.6  2077.0 < 2.2e-16 ***
Residuals  116150 959559     8.3
```

**TABELLA 2.6: Anova del modello di regressione lineare semplice relativo alla Provincia**

Analizzando i risultati del test ANOVA, è possibile affermare che la variabile relativa alla provincia di appartenenza del punto di riconsegna risulta essere significativa e che, quindi, risulta opportuno inserirla all'interno dei modelli statistici che successivamente si andranno a stimare e valutare.

### *2.3.2 La Regione*

Considerazioni analoghe a quelle per le province si possono fare anche in base alle regioni; i punti di riconsegna provengono da 19 regioni italiane (non vi sono osservazioni riguardanti la Sardegna).

| N. | REGIONE | PUNTI |
|---|---|---|
| 1 | Lombardia | 209 |
| 2 | Veneto | 175 |
| 3 | Emilia Romagna | 112 |
| 4 | Campania | 89 |
| 5 | Piemonte | 71 |
| 6 | Friuli Venezia-Giulia | 52 |
| 7 | Toscana | 35 |
| 8 | Abruzzo | 24 |
| 9 | Calabria | 22 |
| 10 | Lazio | 18 |

| N. | REGIONE | PUNTI |
|---|---|---|
| 11 | Sicilia | 17 |
| 12 | Puglia | 16 |
| 13 | Marche | 14 |
| 14 | Basilicata | 13 |
| 15 | Trentino Alto-Adige | 13 |
| 16 | Umbria | 10 |
| 17 | Liguria | 5 |
| 18 | Molise | 1 |
| 19 | Valle d'Aosta | 1 |

**TABELLA 2.7: Punti di riconsegna presenti in ciascuna regione**

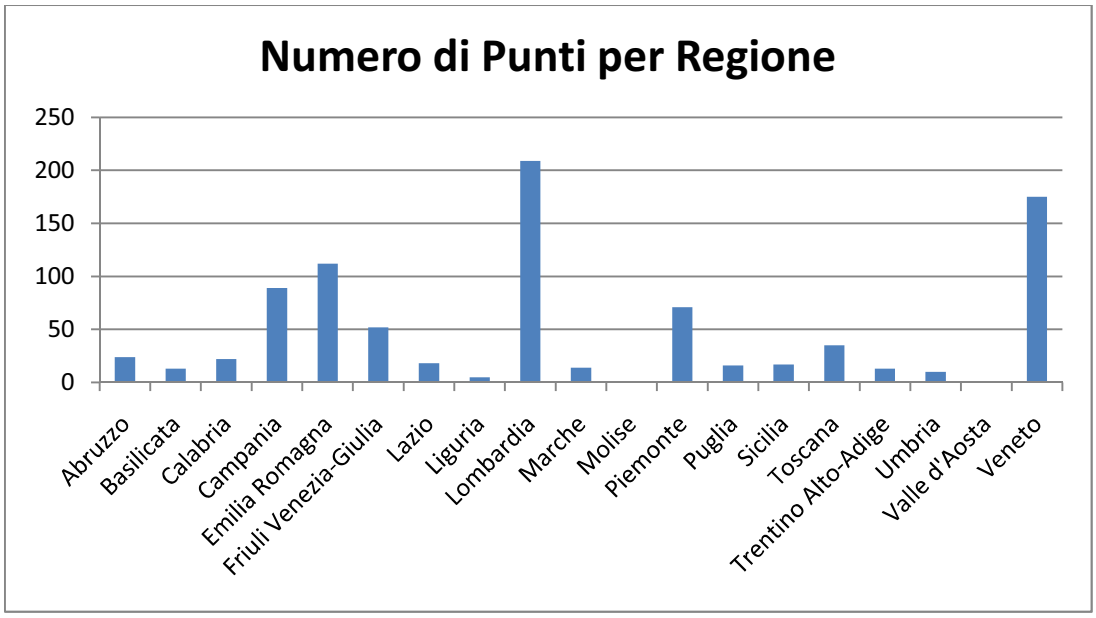



**GRAFICO 2.2: Punti di riconsegna presenti in ciascuna regione**

Per quanto riguarda la numerosità dei punti di riconsegna nelle 19 regioni italiane sembra confermata l'osservazione precedente: le regioni più rappresentate sono quelle corrispondenti alle zone più industrializzate del Paese (Lombardia, Veneto, Emilia Romagna, Campania, Piemonte).

Si analizza ora la quantità media di gas consumata in ciascuna regione.

| N. | REGIONE | CONS |
|---|---|---|
| 1 | Abruzzo | 1.328 |
| 2 | Basilicata | 2.824 |
| 3 | Calabria | 419 |
| 4 | Campania | 4.084 |
| 5 | Emilia Romagna | 4.147 |
| 6 | Friuli Venezia-Giulia | 10.181 |
| 7 | Lazio | 2.750 |
| 8 | Liguria | 57 |
| 9 | Lombardia | 3.191 |
| 10 | Marche | 8.424 |

| N. | REGIONE | CONS |
|---|---|---|
| 11 | Molise | 59 |
| 12 | Piemonte | 4.385 |
| 13 | Puglia | 1.330 |
| 14 | Sicilia | 1.962 |
| 15 | Toscana | 5.105 |
| 16 | Trentino Alto-Adige | 3.067 |
| 17 | Umbria | 328 |
| 18 | Valle d'Aosta | 436 |
| 19 | Veneto | 5.237 |

**TABELLA 2.8: Consumo medio di gas per ciascuna regione**

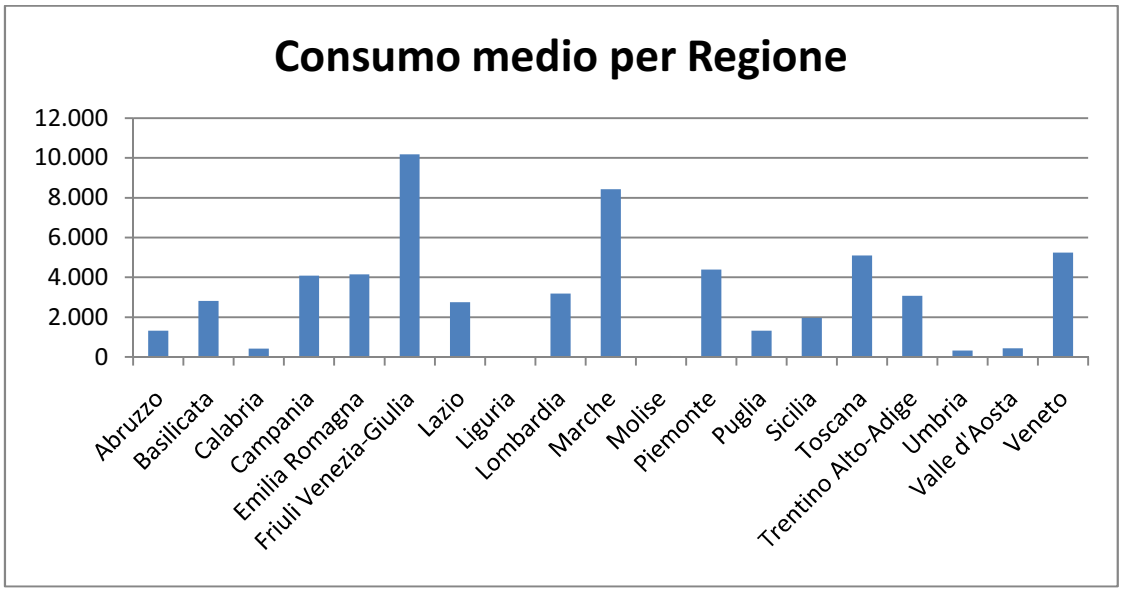



**GRAFICO 2.3: Consumo medio di gas per ciascuna regione**

Anche in questo caso le regioni maggiormente industrializzate passano in secondo piano e sono superate da alcune regioni che presentano un minor numero di punti di riconsegna, come Friuli Venezia-Giulia e Marche.

Come visto in precedenza per la variabile provincia, anche per la variabile regione si potrebbe valutare un raggruppamento delle regioni simili, ovvero quelle che non sembrano presentare un consumo di gas differente tra loro, basandosi sulle stime dei coefficienti in un modello di regressione lineare semplice avente come variabile risposta il consumo giornaliero di gas e come variabile risposta la regione.
La classificazione che ne risulta è la seguente:

| | |
|---|---|
| 1 | Abruzzo |
| 2 | Basilicata, Calabria, Emilia Romagna, Lombardia |
| 3 | Campania, Toscana, Trentino Alto-Adige |
| 4 | Friuli Venezia-Giulia, Marche |
| 5 | Lazio, Umbria |
| 6 | Liguria |
| 7 | Piemonte |
| 8 | Puglia, Valle d'Aosta |
| 9 | Sicilia |
| 10 | Veneto |

**TABELLA 2.9: Classificazione delle Regioni**

Dopo aver stimato un modello di regressione lineare semplice, avente come variabile risposta il consumo giornaliero di gas e come variabile esplicativa la

regione raggruppata secondo la Tabella 2.9, si valuta l'importanza della variabile regione attraverso il test ANOVA, che confronta il modello stimato con il modello avente come unica variabile esplicativa l'intercetta:

```
Analysis of Variance Table

              Df  Sum Sq Mean Sq F value    Pr(>F)
Regione        9   82457  9161.9   967.3 < 2.2e-16 ***
Residuals 116154 1100164     9.5
```

**TABELLA 2.10: Anova del modello di regressione lineare semplice relativo alla Regione**

Analizzando i risultati del test ANOVA, è possibile affermare che la variabile relativa alla regione di appartenenza del punto di riconsegna risulta essere significativa e che, quindi, risulta opportuno inserirla all'interno dei modelli statistici che successivamente si andranno a stimare e valutare.

## 2.4 LE VARIABILI TEMPORALI

Per ciascun punto di riconsegna è possibile risalire, attraverso la data, a una serie di variabili temporali che possono rivelarsi utili al fine di ricercare eventuali differenze di consumo dovute al differente momento di rilevazione.

### *2.4.1 Il giorno della settimana*

Una delle variabili temporali probabilmente più importante è il giorno della settimana in cui è stato rilevato il consumo di gas; è abbastanza intuitivo, infatti,

che il consumo giornaliero di gas sia fortemente condizionato da questo tipo di dato.

| GIORNO SETTIMANA | CONSUMO |
|---|---|
| Lunedì | 4.687 |
| Martedì | 4.743 |
| Mercoledì | 4.680 |
| Giovedì | 4.639 |
| Venerdì | 4.507 |
| Sabato | 3.582 |
| Domenica | 3.434 |

**TABELLA 2.11: Consumo medio di gas per giorno della settimana**

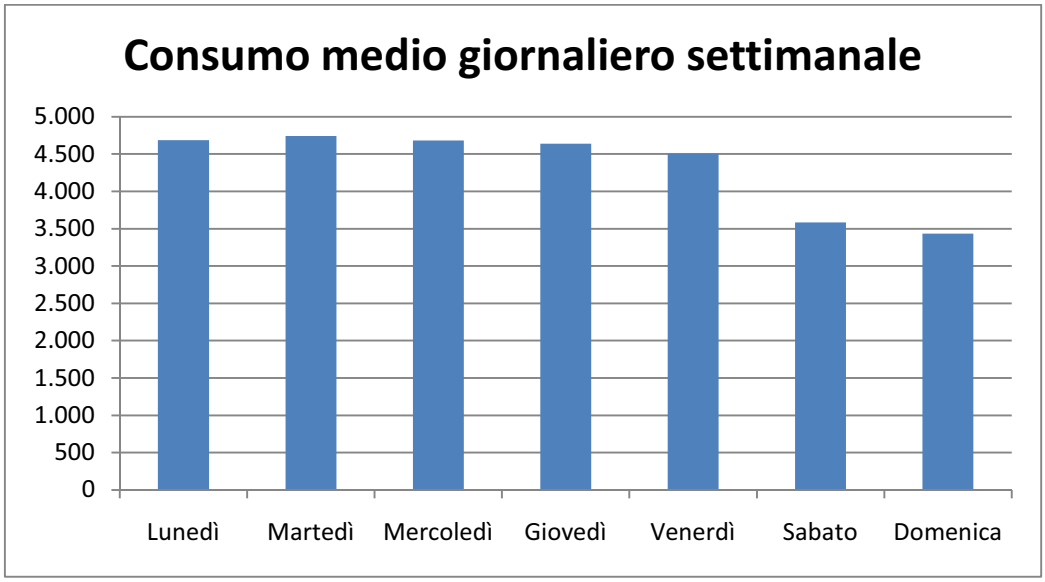



**GRAFICO 2.4: Consumo medio di gas per giorno della settimana**

L'istogramma riportato nel Grafico 2.4 conferma l'idea di partenza: il consumo di gas registra un calo, senza tuttavia azzerarsi, nel fine settimana e questo è probabilmente dovuto al fatto che il sabato molte aziende lavorano con orario ridotto e la domenica sono chiuse.

E’ evidente, inoltre, che all’interno della settimana vi sono alcuni giorni che presentano comportamenti simili nel consumo del gas; è possibile ad esempio osservare come nei giorni di lunedì, martedì, mercoledì e giovedì non sembrano esserci particolari differenze, e potrebbe quindi risultare utile raggrupparli insieme in un’unica variabile, mentre si continuerà a trattare come variabili diverse i restanti giorni di venerdì, sabato e domenica.

| | |
|---|---|
| 1 | Domenica |
| 2 | Lunedì, Martedì, Mercoledì, Giovedì |
| 3 | Venerdì |
| 4 | Sabato |

**TABELLA 2.12: Classificazione dei giorni della settimana**

Dopo aver stimato un modello di regressione lineare semplice, avente come variabile risposta il consumo giornaliero di gas e come variabile esplicativa il giorno della settimana raggruppato secondo la Tabella 2.12, si valuta l’importanza della variabile regione attraverso il test ANOVA, che confronta il modello stimato con il modello avente come unica variabile esplicativa l’intercetta:

```
Analysis of Variance Table

              Df  Sum Sq Mean Sq F value    Pr(>F)
gior_sett      3   22475  7491.5  750.09 < 2.2e-16 ***
Residuals  116160 1160146    10.0
```

**TABELLA 2.13: Anova del modello di regressione lineare semplice relativo al giorno della settimana**

Analizzando i risultati del test ANOVA, è possibile affermare che la variabile relativa alla giorno della settimana risulta essere significativa e che, quindi, risulta opportuno inserirla all'interno dei modelli statistici che successivamente si andranno a stimare e valutare.

### *2.4.2 Il giorno del mese*

Le rilevazioni dei consumi di gas vanno dal primo giorno del mese, contrassegnato dall'1, all'ultimo, contrassegnato dal 31.

| GIORNO | CONS |
|---|---|
| 1 | 4.108 |
| 2 | 4.260 |
| 3 | 4.339 |
| 4 | 4.398 |
| 5 | 4.226 |
| 6 | 4.079 |
| 7 | 4.197 |
| 8 | 4.255 |
| 9 | 4.416 |
| 10 | 4.310 |
| 11 | 4.405 |

| GIORNO | CONS |
|---|---|
| 12 | 4.324 |
| 13 | 4.282 |
| 14 | 4.361 |
| 15 | 4.504 |
| 16 | 4.511 |
| 17 | 4.441 |
| 18 | 4.463 |
| 19 | 4.423 |
| 20 | 4.408 |
| 21 | 4.548 |
| 22 | 4.555 |

| GIORNO | CONS |
|---|---|
| 23 | 4.404 |
| 24 | 4.193 |
| 25 | 4.136 |
| 26 | 4.303 |
| 27 | 4.219 |
| 28 | 4.342 |
| 29 | 4.204 |
| 30 | 4.114 |
| 31 | 4.281 |

**TABELLA 2.14: Consumo medio di gas per giorno del mese**

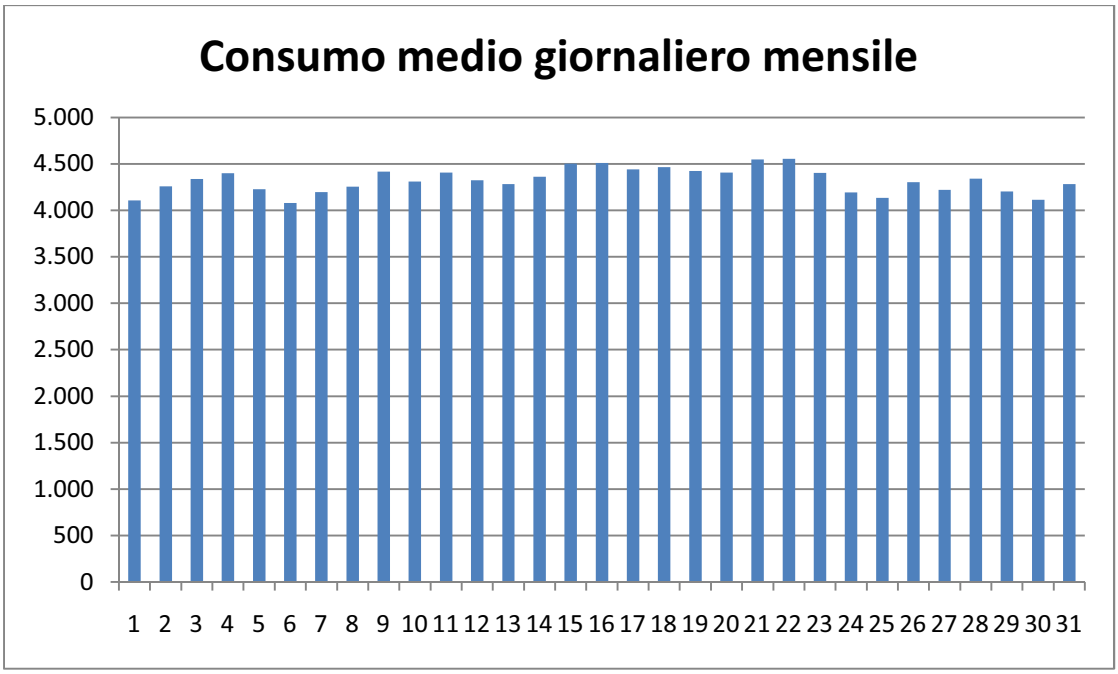



**GRAFICO 2.5: Consumo medio di gas per giorno del mese**

L'andamento del consumo medio giornaliero mensile è simile a un'onda. Questo andamento anomalo è determinato dal fatto che i giorni che riportano un consumo medio di gas più elevato (come il 21 e il 22) sono tali in quanto presentano un numero minore di osservazioni relative ai giorni di sabato e domenica che, come visto in precedenza, sono caratterizzati da un consumo medio di gas minore rispetto agli altri giorni della settimana. All'opposto, invece, i giorni che riportano un consumo medio di gas più basso (come il 6 e il 25), sono tali in quanto le osservazioni nei giorni di sabato e domenica sono maggiormente rappresentate.

Questo porta a concludere che l'andamento dell'istogramma è dovuto principalmente alla composizione del dataset e non a una reale differenza nel consumo medio giornaliero mensile di gas; per questo motivo tale variabile non sarà inserita nei modelli previsionali.

### *2.4.3 Il mese*

Di significatività ben diversa, invece, risultano essere le differenze che si notano nel consumo medio mensile annuale.

| MESE | CONS |
|---|---|
| Gennaio | 7.009 |
| Febbraio | 6.514 |
| Marzo | 4.936 |
| Aprile | 2.675 |
| Maggio | 2.464 |
| Giugno | 2.178 |

| MESE | CONS |
|---|---|
| Luglio | 2.255 |
| Agosto | 2.546 |
| Settembre | 3.039 |
| Ottobre | 3.292 |
| Novembre | 5.174 |
| Dicembre | 6.617 |

**TABELLA 2.15: Consumo medio di gas per mese dell'anno**

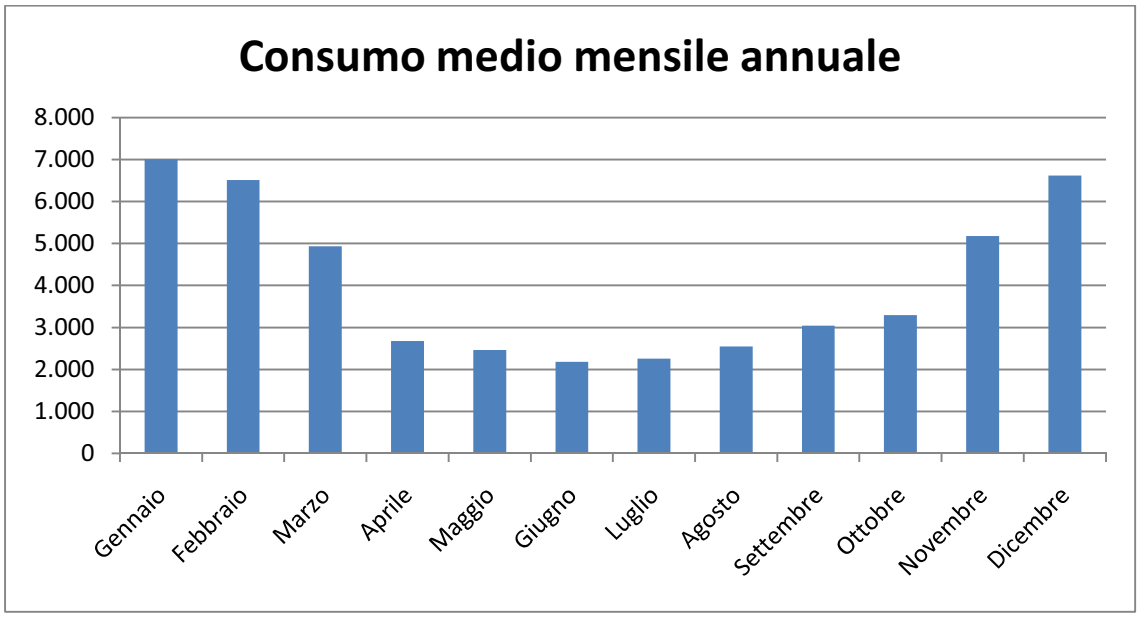



**GRAFICO 2.6: Consumo medio di gas per mese dell'anno**

Appare subito evidente che il consumo medio mensile annuale varia molto nel corso dell'anno; il consumo di gas risulta minore nei mesi estivi e maggiore nei

mesi invernali, evidenziando un trend decrescente passando dall'inverno alla primavera e crescente passando dall'autunno all'inverno. Questo andamento è dovuto principalmente alle condizioni climatiche: nei mesi estivi, nei quali si registrano temperature più elevate, il consumo di gas è minore rispetto ai mesi invernali, nei quali si registrano temperature di gran lunga inferiori. Il consumo di gas, tuttavia, non si azzera mai e questo è dovuto alla presenza di contratti stipulati con aziende e industrie che operano durante tutto l'arco dell'anno e che, di conseguenza, hanno bisogno continuo di gas per le proprie lavorazioni.
Il mese dell'anno, dunque, ci fornisce informazioni molto utili al fine di prevedere il consumo giornaliero di gas nei vari punti di riconsegna.

Per confermare le considerazioni precedenti, si valuta l'importanza della variabile regione, stimando un modello di regressione lineare semplice, avente come variabile risposta il consumo giornaliero di gas e come variabile esplicativa il mese, ed effettuando un test ANOVA, che confronta il modello stimato con il modello avente come unica variabile esplicativa l'intercetta:

```
Analysis of Variance Table

              Df  Sum Sq Mean Sq F value    Pr(>F)
mese          11   62509  5682.7  589.27 < 2.2e-16 ***
Residuals 116152 1120111     9.6
```

**TABELLA 2.16: Anova del modello di regressione lineare semplice relativo al mese**

Analizzando i risultati del test ANOVA, è possibile affermare che la variabile relativa alla mese risulta essere significativa e che, quindi, risulta opportuno inserirla all'interno dei modelli statistici che successivamente si andranno a stimare e valutare.

### *2.4.4 L'anno*

Un'altra variabile che permette di collocare temporalmente le osservazioni del dataset è l'anno di rilevazione del consumo di gas.

| ANNO | CONS |
|---|---|
| 2009 | 6.026 |
| 2010 | 4.237 |
| 2011 | 3.266 |

**TABELLA 2.17: Consumo medio di gas per anno**

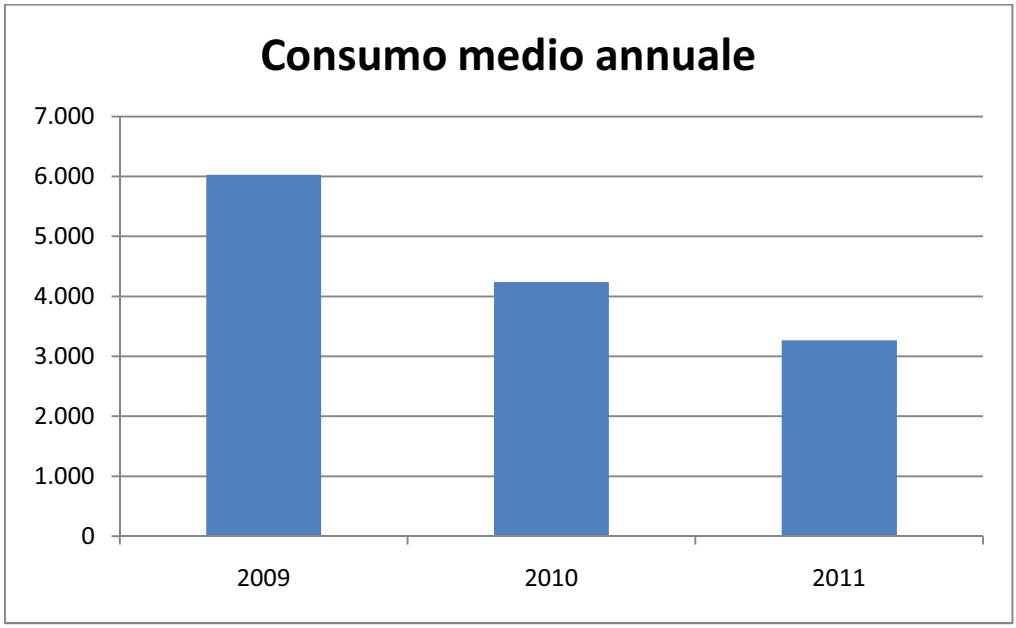



**GRAFICO 2.7: Consumo medio di gas per anno**

Il consumo medio annuale sembra diminuire, e questo coincide con quanto rilevato in precedenza; tuttavia occorre tenere in considerazione che nel dataset sono presenti osservazioni per gli ultimi 3 mesi del 2009 (Ottobre, Novembre e Dicembre), per tutti i mesi del 2010 e per i primi 4 mesi del 2011 (Gennaio,

Febbraio, Marzo, Aprile); i valori medi annuali, quindi, potrebbero essere condizionati dalla composizione del dataset, in quanto il consumo medio di gas è strettamente connesso al periodo dell'anno in cui viene rilevato; per questo motivo tale variabile non sarà inserita nei modelli previsionali.

### *2.4.5 Le festività*

L'ultima variabile che fornisce una dimensione temporale della rilevazione del consumo del gas è quella che riguarda le festività dell'anno. Questo tipo di informazione risulterà sicuramente molto importante nella costruzione dei modelli statistici, in quanto in alcuni giorni dell'anno la maggior parte delle industrie con cui sono stati stipulati contratti di fornitura del gas risulta chiusa a causa di festività religiose o civili. Risulta quindi opportuno creare una variabile dummy che assuma valore 0 se si tratta di un normale giorno feriale e valore 1 se si tratta di un giorno festivo. Oltre alle domeniche, le festività considerate sono le seguenti: l'1 Gennaio, Capodanno, il 6 gennaio, Epifania, il 5 Aprile 2010, Lunedì dell'Angelo, il 25 Aprile, festa della Liberazione nonché Lunedì dell'Angelo nel 2011, l'1 Maggio, festa del Lavoro, il 2 Giugno, festa della Repubblica, il 15 Agosto, Ferragosto, l'1 Novembre, festa di Tutti i Santi, l'8 Dicembre, Immacolata Concezione, il 25 e 26 Dicembre, Natale e Santo Stefano.

| GIORNO | CONS |
|---|---|
| Feriale | 4.482 |
| Festivo | 3.532 |

**TABELLA 2.18: Consumo medio di gas per tipo di giorno (feriale o festivo)**

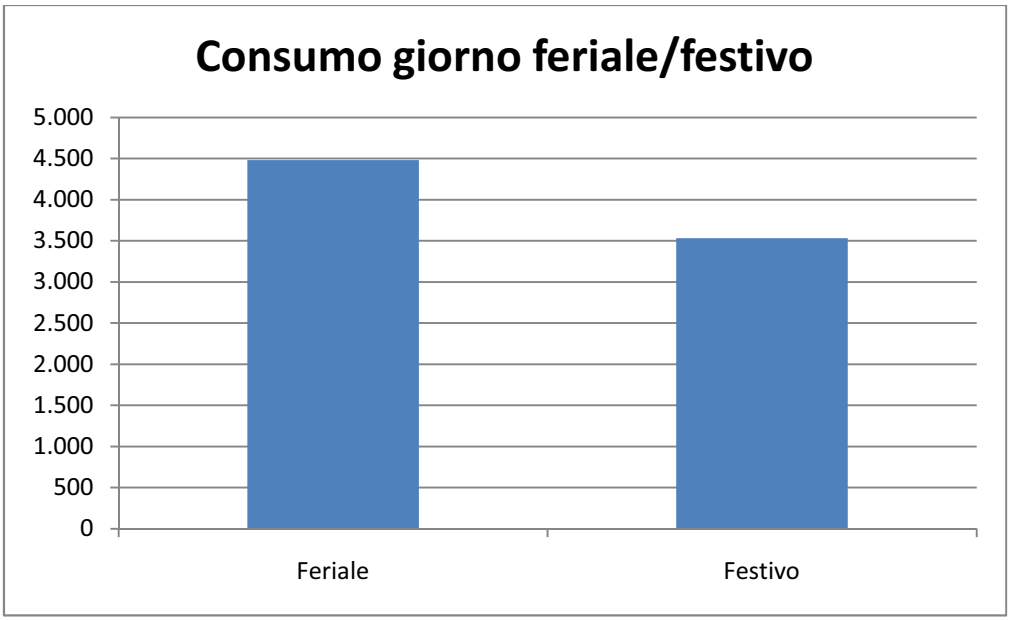



**GRAFICO 2.8: Consumo medio di gas per tipo di giorno (feriale o festivo)**

La Tabella 2.11 e il Grafico 2.8 suggeriscono che il consumo medio di gas nei giorni feriali è maggiore rispetto a quello dei giorni festivi; per verificare questa intuizione e averne una prima conferma è possibile condurre un test t che verifichi l'ipotesi nulla di uguaglianza delle medie nei due sottocampioni (consumo nei giorni feriali e consumo nei giorni festivi) contro un'ipotesi alternativa unilaterale destra.

$$\begin{cases} \mu_{FERIALE} = \mu_{FESTIVO} \\ \mu_{FERIALE} > \mu_{FESTIVO} \end{cases}$$

La verifica di questo sistema di ipotesi fornisce i seguenti risultati:

```
Two-sample t test with equal variances
------------------------------------------------------------------------------
Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
------------------------------------------------------------------------------
    0 |  286568    4481.543    45.35491    24279.41    4392.649    4570.437
    1 |   56771    3531.609    96.39718    22968.23     3342.67    3720.548

---------+--------------------------------------------------------------------
combined |  343339    4324.472    41.07865    24070.09    4243.959    4404.985
---------+--------------------------------------------------------------------
    diff |            949.9338    110.5646                733.2304    1166.637
------------------------------------------------------------------------------
diff = mean(0) - mean(1)                                      t =   8.5917
Ho: diff = 0                                     degrees of freedom =   343337

Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) =
0.0000
```

**TABELLA 2.19: Verifica di ipotesi consumo medio di gas nei giorni feriali e festivi**

Il test t, condotto per la verifica del sistema di ipotesi precedente, rifiuta, a un livello di significatività del 95%, l'ipotesi nulla di uguaglianza delle medie nei due sottocampioni (*p-value*=0); come ipotizzato in precedenza, è possibile quindi affermare che il consumo di gas nei giorni feriali è maggiore rispetto ai giorni festivi.

Per valutare l'importanza della variabile, si stima un modello di regressione lineare semplice, avente come variabile risposta il consumo giornaliero di gas e come variabile esplicativa il tipo di giorno (feriale o festivo), e si effettua un test ANOVA, che confronta il modello stimato con il modello avente come unica variabile esplicativa l'intercetta:

```
Analysis of Variance Table

               Df  Sum Sq Mean Sq F value    Pr(>F)
tipo_giorno     1   12816 12816.0  1272.6 < 2.2e-16 ***
Residuals   116162 1169805    10.1
```

**TABELLA 2.20: Anova del modello di regressione lineare semplice relativo al tipo di giorno**

Analizzando i risultati del test ANOVA, è possibile affermare che la variabile relativa al tipo di giorno risulta essere significativa e che, quindi, risulta opportuno inserirla all'interno dei modelli statistici che successivamente si andranno a stimare e valutare.

## 2.5 IL NUMERO DI CONTRATTI E LA CAPACITA' PRENOTATA DAGLI UTENTI

Il numero di contratti rappresenta quanti contratti sono in vigore in un determinato punto di riconsegna nel giorno in cui viene rilevato il consumo di gas; ogni punto di riconsegna, infatti, può portare il gas a più Utenti della rete e il numero di contratti in vigore potrebbe rappresentare una buona approssimazione della quantità di gas consumata in un determinato punto. Si può supporre, infatti, che all'aumentare del numero di contratti in vigore nel punto, aumenta la capacità di gas richiesta e, di conseguenza, la quantità di gas consumata.

| CONTRATTI | CONSUMO |
|---|---|
| 1 | 4.489 |
| 2 | 9.786 |
| 3 | 30.532 |
| 4 | 28.348 |
| 5 | 42.949 |
| 6 | 71412 |
| 7 | 18.177 |

**TABELLA 2.21: Consumo medio di gas per numero di contratti**

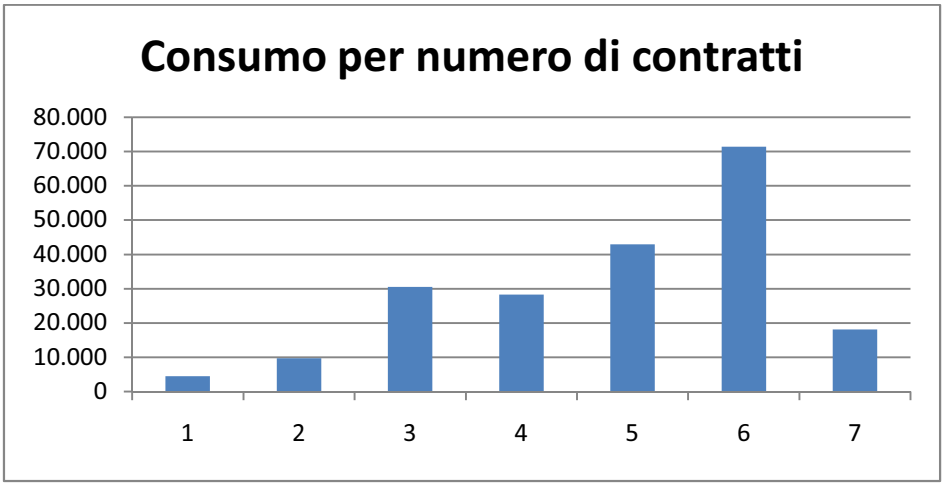



**GRAFICO 2.9: Consumo medio di gas per numero di contratti**

L'intuizione di partenza sembra essere confermata solo in parte dai dati: il consumo medio di gas sembra non variare significativamente per 3 e 4 contratti, e, addirittura, diminuisce quando i contratti passano da 6 a 7. Il numero di contratti non sembra quindi una buona approssimazione dell'ordine di grandezza del consumo di gas: potrebbe accadere, infatti, che la capacità prenotata in 7 contratti sia minore di quella prenotata in 4, 5 o 6 contratti.

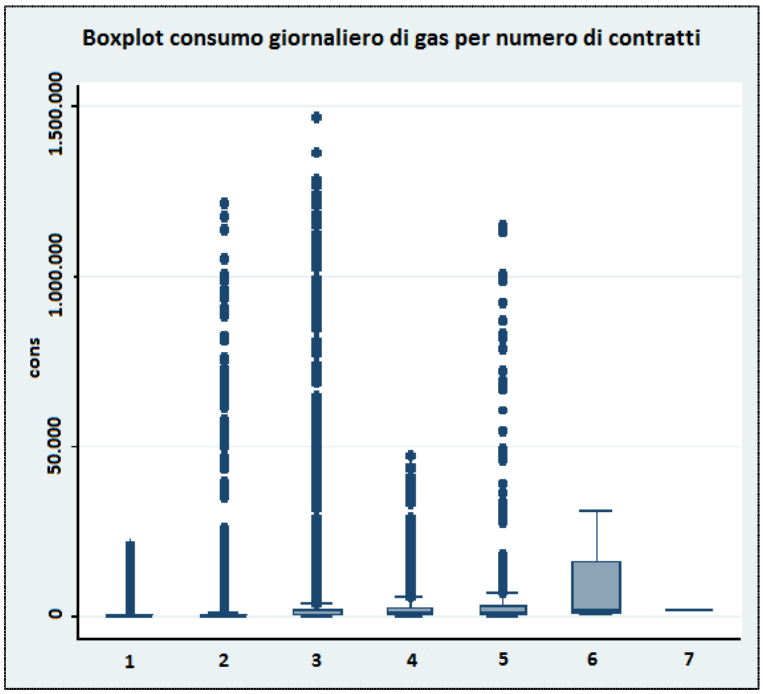



**GRAFICO 2.10: Boxplot consumo giornaliero di gas per numero di contratti**

Anche il Grafico 2.10 evidenzia il fatto che sono presenti osservazioni di consumi giornalieri di gas elevati anche in caso di un numero basso di contratti in vigore nel punto di riconsegna; anche questo grafico, quindi, testimonia che il numero di contratti non sembra una buona approssimazione dell'ordine di grandezza del consumo di gas

Si intende, comunque, valutare l'importanza della variabile, stimando un modello di regressione lineare semplice, avente come variabile risposta il consumo giornaliero di gas e come variabile esplicativa il numero di contratti, ed effettuando un test ANOVA, che confronta il modello stimato con il modello avente come unica variabile esplicativa l'intercetta:

```
Analysis of Variance Table

              Df  Sum Sq Mean Sq F value    Pr(>F)
contratti      1   86480   86480  9164.6 < 2.2e-16 ***
Residuals 116162 1096140       9
```

**TABELLA 2.22: Anova del modello di regressione lineare semplice relativo al numero di contratti**

Analizzando i risultati del test ANOVA, è possibile affermare che la variabile relativa al numero di contratti risulta essere significativa e che, quindi, risulta opportuno inserirla all'interno dei modelli statistici che successivamente si andranno a stimare e valutare.

Poiché la variabile relativa al numero di contratti non sembra una buona approssimazione dell'ordine di grandezza del consumo di gas, si decide di inserire nel modello la capacità prenotata giornalmente dagli Utenti in un determinato punto di riconsegna; questa variabile può risultare un'ottima approssimazione dell'ordine di grandezza della quantità di gas che verrà poi consumata, in quanto tale quantità non può superare la capacità prenotata, pena il pagamento di una penale.

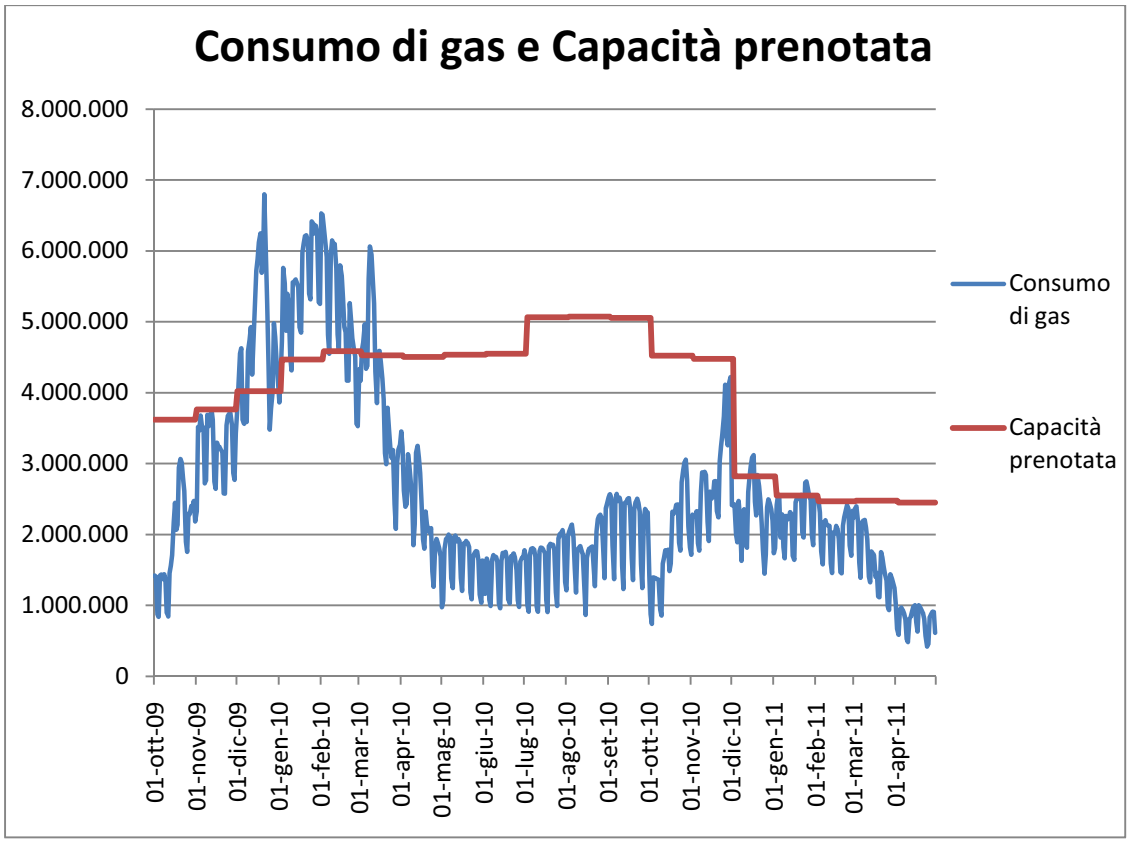



**GRAFICO 2.11: Consumo medio di gas e capacità prenotata dagli Utenti**

Dal Grafico 2.11 delle serie storiche del consumo di gas e della capacità prenotata è possibile evidenziare come, negli intervalli ottobre 2009-novembre 2009 e marzo 2010-aprile 2011, la serie storica della capacità prenotata sia praticamente sempre al di sopra della serie storica del consumo di gas: questo significa che la quantità di gas consumata è risultata praticamente sempre inferiore alla capacità prenotata. Nell'intervallo dicembre 2009-febbraio 2010, invece, la quantità di gas consumata risulta superiore alla capacità prenotata; questo è causato da una sottostima della previsione dei consumi di gas, che può essere dovuto a un'ondata di freddo non prevista o a un aumento non previsto della produzione industriale.

Per valutare l'importanza della variabile, si stima un modello di regressione lineare semplice, avente come variabile risposta il consumo giornaliero di gas e come variabile esplicativa la capacità prenotata dagli Utenti, e si effettua un test

ANOVA, che confronta il modello stimato con il modello avente come unica variabile esplicativa l'intercetta:

```
Analysis of Variance Table

               Df Sum Sq Mean Sq F value    Pr(>F)
capacita        1 726949  726949  185317 < 2.2e-16 ***
Residuals  116162 455672       4
```

**TABELLA 2.23: Anova del modello di regressione lineare semplice relativo alla Capacità**

Analizzando i risultati del test ANOVA, è possibile affermare che la variabile relativa alla capacità prenotata dagli Utenti risulta essere significativa e che, quindi, risulta opportuno inserirla all'interno dei modelli statistici che successivamente si andranno a stimare e valutare.

## 2.6 LA DESTINAZIONE D'USO DEL PUNTO DI RICONSEGNA

Un'altra variabile fondamentale da inserire in un modello statistico per la previsione dei consumi di gas è la destinazione d'uso del punto di riconsegna. Si possono individuare tre destinazioni d'uso:

1) **Uso civile**: in questa tipologia rientrano tutti quei punti di riconsegna all'interno dei quali sono stati stipulati esclusivamente contratti con Utenti che utilizzano il gas in ambito civile (abitazioni, case, uffici, scuole, ospedali, …); questi punti si distinguono dagli altri in quanto l'andamento del consumo è simile a quello rappresentato nella Figura.

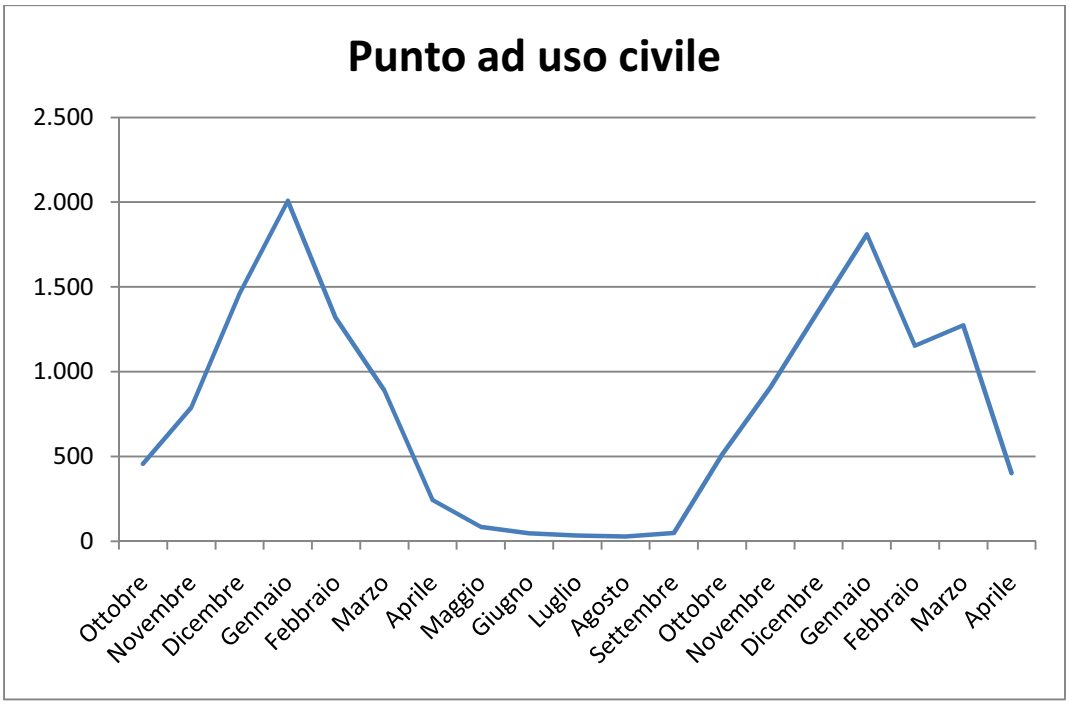



**GRAFICO 2.12: Punto di riconsegna con destinazione d'uso civile**

I punti ad uso civile sono caratterizzati da un consumo medio mensile di gas elevato nei mesi invernali e che cala fino a quasi 0 nei mesi estivi, in quanto le temperature più elevate portano gli Utenti a consumare quantitativi di gas molto bassi o praticamente nulli (impianti di riscaldamento praticamente sempre spenti nei mesi caldi dell'anno).

2) **Uso civile e uso industriale**: in questa tipologia rientrano tutti quei punti di riconsegna all'interno dei quali sono stati stipulati contratti con alcuni Utenti che utilizzano il gas in ambito civile (abitazioni, case, uffici, scuole, ospedali, …) e con altri che lo utilizzano in ambito industriale (fabbriche, industrie di lavorazione, industrie agricole); questi punti si distinguono dagli altri in quanto l'andamento del consumo è simile a quello rappresentato nella Figura.

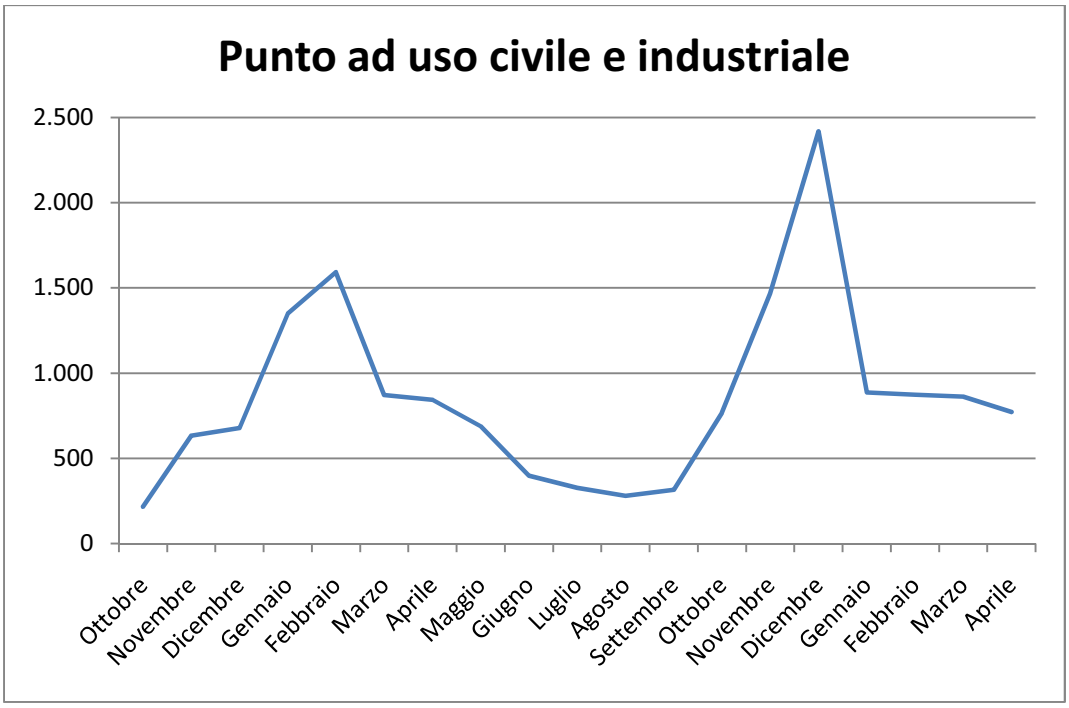



**GRAFICO 2.13: Punto di riconsegna con destinazione d'uso civile e industriale**

I punti ad uso civile e industriale sono caratterizzati da un consumo medio di gas elevato nei mesi invernali, dovuto ai contratti civili in essere nel punto (riscaldamento delle abitazioni e degli uffici); tale consumo, tuttavia, non si azzera nei mesi estivi ma resta significativamente maggiore di 0 anche durante l'estate, in quanto le industrie che hanno stipulato un contratto nel punto di riconsegna lavorano anche in quel periodo dell'anno e necessitano sempre di un determinato quantitativo di gas per restare in funzione, indipendentemente dalle condizioni climatiche.

3) **<u>Uso industriale</u>**: in questa tipologia rientrano tutti quei punti di riconsegna all'interno dei quali sono stati stipulati esclusivamente contratti con Utenti che utilizzano il gas in ambito industriale (fabbriche, industrie di lavorazione, industrie agricole); questi punti si distinguono dagli altri in quanto l'andamento del consumo è simile a quello rappresentato nella Figura.

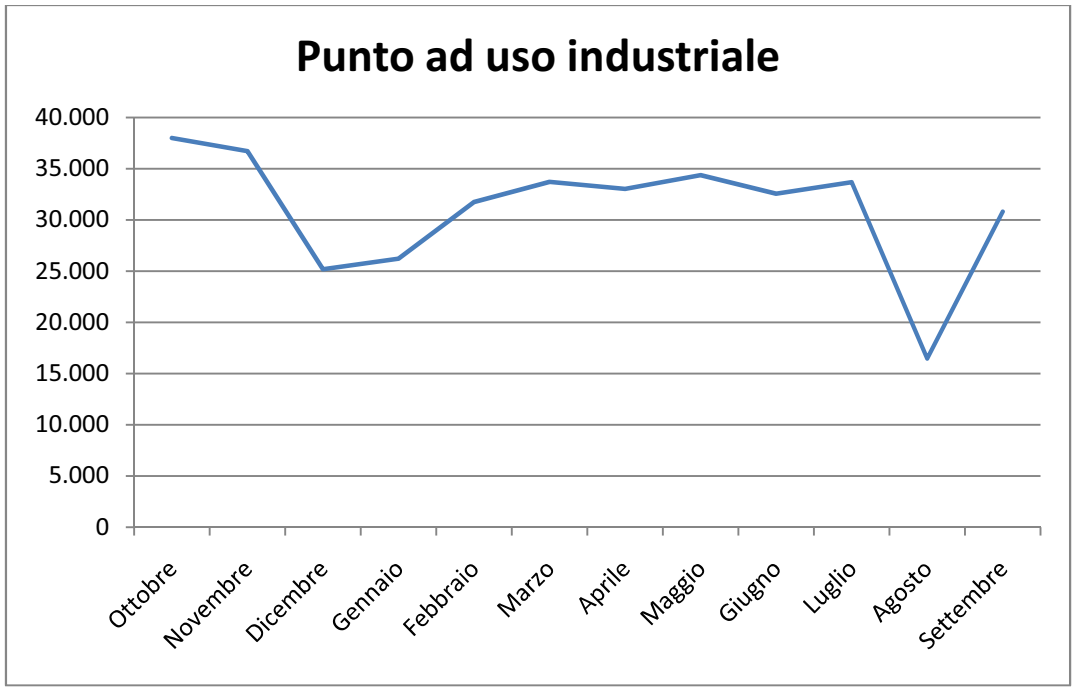



**GRAFICO 2.14: Punto di riconsegna con destinazione d'uso industriale**

I punti ad uso industriale sono caratterizzati da un consumo medio di gas pressoché costante, in quanto un'industria necessita sempre di un quantitativo di gas simile durante tutto l'arco dell'anno per poter lavorare e restare in funzione. Solitamente, gli unici valori significativamente più bassi rispetto agli altri si trovano in corrispondenza dei mesi di Dicembre e Agosto, in quanto le industrie solitamente osservano un periodo di chiusura rispettivamente per le vacanze di Natale e per le ferie estive.

Una volta suddivisi i punti di riconsegna del dataset in base alla destinazione d'uso, è possibile verificare se si possono riscontrare differenze significative nel consumo di gas.

| DESTINAZIONE | CONSUMO |
|---|---|
| Civile | 195 |
| Civile e Industriale | 6.173 |
| Industriale | 8.631 |

**TABELLA 2.24: Consumo medio di gas per destinazione d'uso**

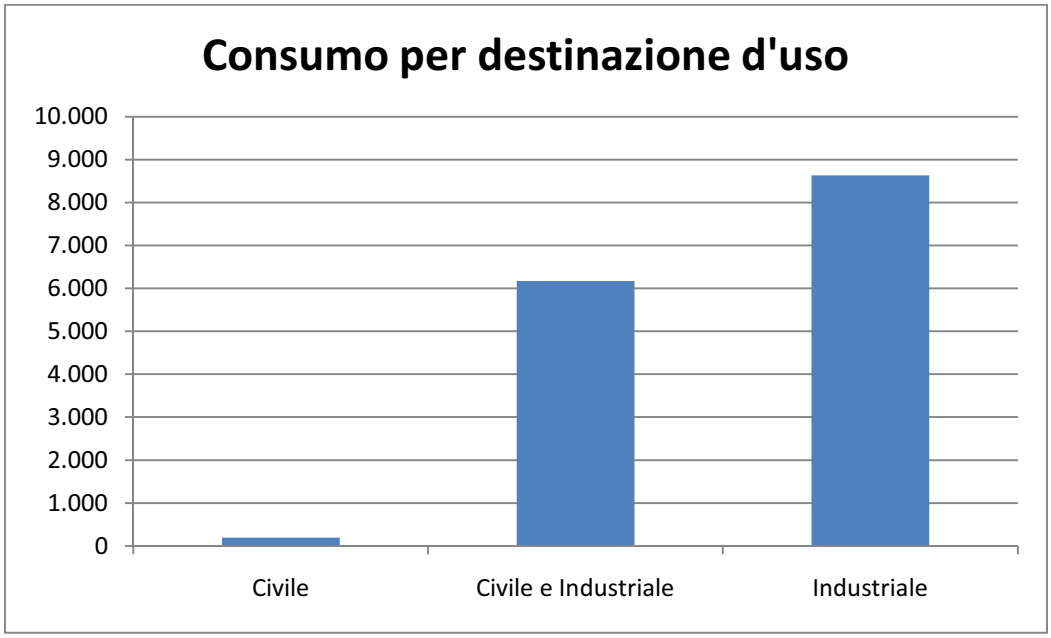



**GRAFICO 2.15: Consumo medio di gas per destinazione d'uso**

La classificazione dei punti per destinazione d'uso sembra essere rilevante per la previsione del consumo di gas, in quanto il consumo medio dei punti ad uso civile è decisamente inferiore rispetto a quello dei punti ad uso contemporaneamente civile e industriale e a quello dei punti esclusivamente ad uso industriale. Questo risultato era preventivabile, in quanto le industrie necessitano durante tutto l'arco dell'anno di un quantitativo continuo di gas decisamente superiore rispetto alle abitazioni.

Per valutare l'importanza della variabile, si stima un modello di regressione lineare semplice, avente come variabile risposta il consumo giornaliero di gas e come variabile esplicativa la destinazione d'uso del punto di riconsegna, e si effettua un test ANOVA, che confronta il modello stimato con il modello avente come unica variabile esplicativa l'intercetta:

```
Analysis of Variance Table

               Df Sum Sq Mean Sq F value    Pr(>F)
tipo_punto      2 446343  223171   35209 < 2.2e-16 ***
Residuals  116161 736278       6
```

**TABELLA 2.25: Anova del modello di regressione lineare semplice relativo alla destinazione d'uso**

Analizzando i risultati del test ANOVA, è possibile affermare che la variabile relativa alla destinazione d'uso del punto di riconsegna risulta essere significativa e che, quindi, risulta opportuno inserirla all'interno dei modelli statistici che successivamente si andranno a stimare e valutare.

## 2.7 I DATI CLIMATICI

### *2.7.1 La zona climatica*

La classificazione del territorio nazionale in zone climatiche risulta un aspetto molto rilevante ai fini del contenimento dei consumi di energia.
La Tabella 2.14 riporta, per ognuna delle zone climatiche, il periodo dell'anno e il numero massimo di ore giornaliere in cui è consentita l'accensione degli impianti di riscaldamento (nel caso di condizioni meteorologiche particolarmente avverse i singoli comuni possono consentire l'accensione degli impianti anche in periodi

diversi). L'unità di misura utilizzata per l'individuazione della zona climatica di appartenenza di ciascun comune è il grado-giorno, ovvero la somma, estesa a tutti i giorni di un periodo annuale convenzionale di riscaldamento, delle sole differenze positive giornaliere tra la temperatura dell'ambiente, convenzionalmente fissata a 20°C, e la temperatura media esterna giornaliera.

| ZONA | GRADI-GIORNO | ORE | ESEMPI |
|---|---|---|---|
| A | fino a 600 | 6 | Lampedusa, Linosa, Porto Empedocle |
| B | da oltre 600 a 900 | 8 | Agrigento, Catania, Crotone, Messina, Palermo, Reggio Calabria, Siracusa, Trapani |
| C | da oltre 900 a 1.400 | 10 | Bari, Benevento, Brindisi, Cagliari, Caserta, Catanzaro, Cosenza, Imperia, Latina, Lecce, Napoli, Oristano, Ragusa, Salerno, Sassari, Taranto |
| D | da oltre 1.400 a 2.100 | 12 | Ancona, Ascoli Piceno, Avellino, Caltanissetta, Chieti, Firenze, Foggia, Forli', Genova, Grosseto, Isernia, La Spezia, Livorno, Lucca, Macerata, Massa, Carrara, Matera, Nuoro, Pesaro, Pesaro, Pescara, Pisa, Pistoia, Prato, Roma, Savona, Siena, Teramo, Terni, Verona, Vibo Valentia, Viterbo |
| E | da oltre 2.100 a 3.000 | 14 | Alessandria, Aosta, Arezzo, Asti, Bergamo, Biella, Bologna, Bolzano, Brescia, Campobasso, Como, Cremona, Enna, Ferrara, Cesena, Frosinone, Gorizia, L'Aquila, Lecco, Lodi, Mantova, Milano, Modena, Novara, Padova, Parma, Pavia, Perugia, Piacenza, Pordenone, Potenza, Ravenna, Reggio Emilia, Rieti, Rimini, Rovigo, Sondrio, Torino, Trento, Treviso, Trieste, Udine, Varese, Venezia, Verbania, Vercelli, Vicenza |
| F | oltre 3.000 | 24 | Belluno, Cuneo |

**TABELLA 2.26: Zone climatiche italiane**

La Figura 2.1 riporta la carta geografica d'Italia con la suddivisione delle zone climatiche secondo la tabulazione del decreto legislativo.

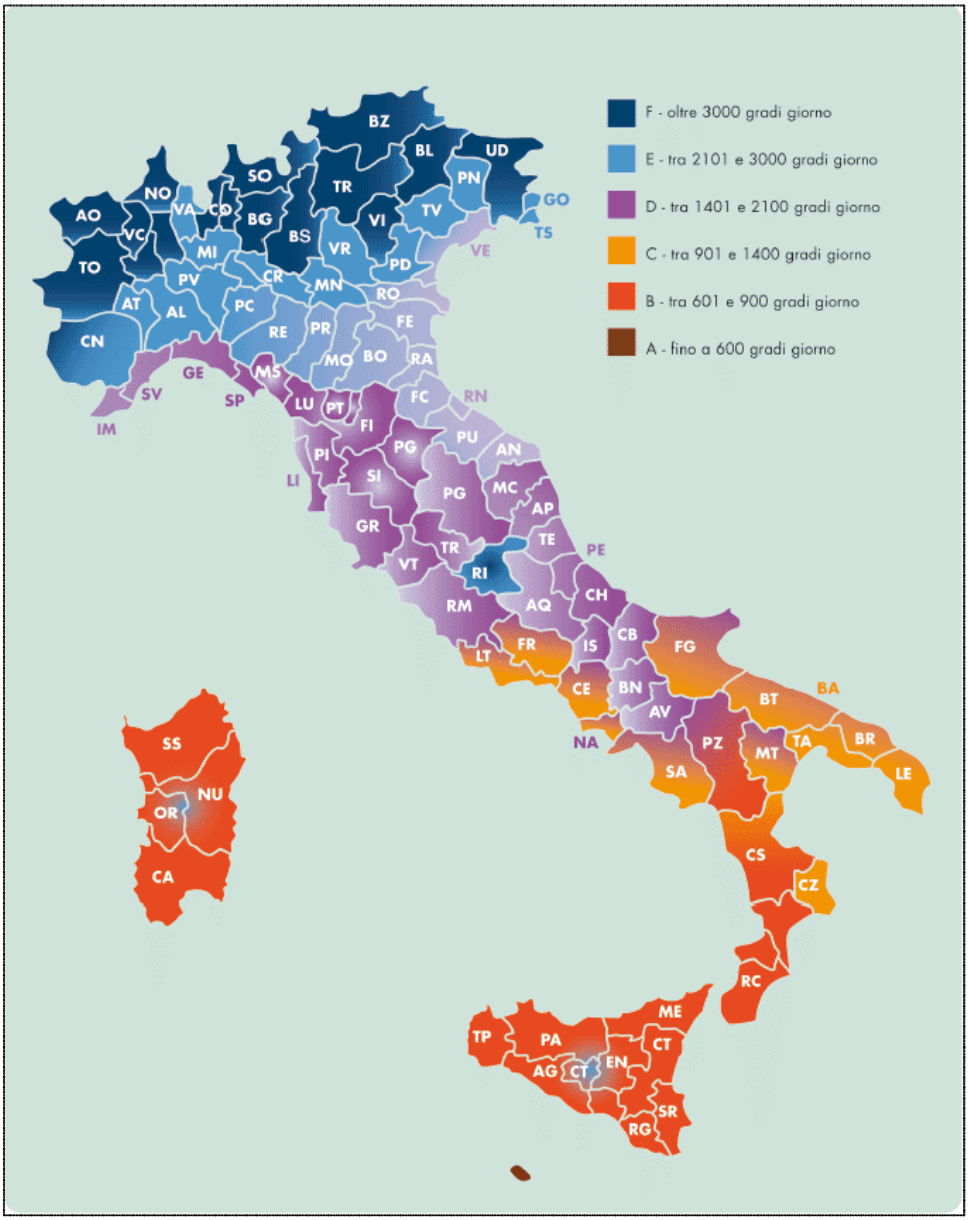



**FIGURA 2.1: Zone climatiche italiane**

I punti di riconsegna presenti nel dataset si collocano in tutte le zone climatiche, con l'eccezione della zona A; una volta individuate le zone climatiche di appartenenza, occorre verificare se effettivamente il consumo di gas varia a

seconda della zona climatica all'interno della quale si collocano il punto e gli Utenti a cui viene distribuito il gas.

| ZONA | CONSUMO |
|---|---|
| B | 928 |
| C | 2.960 |
| D | 2.896 |
| E | 4.624 |
| F | 6.366 |

**TABELLA 2.27: Consumo medio di gas per zona climatica**

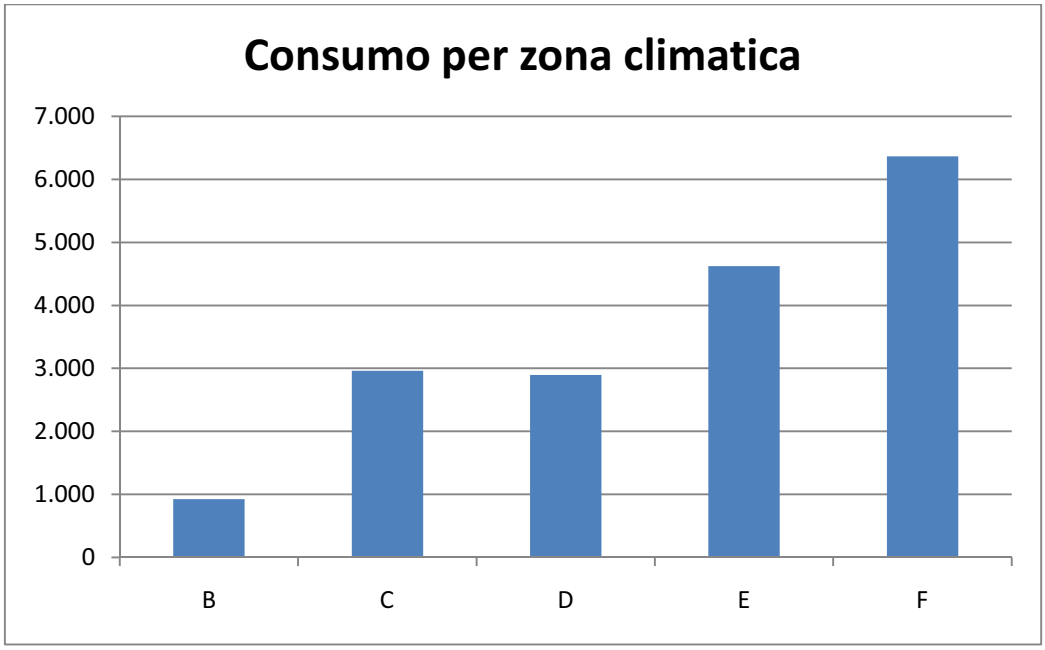



**GRAFICO 2.16: Consumo medio di gas per zona climatica**

Anche la zona climatica sembra risultare una variabile rilevante per la previsione del consumo di gas; il consumo medio di gas, infatti sembra variare da zona a zona, eccezion fatta per le zone C e D (centro e centro-sud dell'Italia) che sembrano registrare un consumo medio di gas pressoché analogo. Anche questo

risultato era preventivabile, in quanto la suddivisione in zone climatiche è strettamente connessa con la temperatura (suddivisione basata sui gradi-giorno).

Per valutare l'importanza della variabile, si stima un modello di regressione lineare semplice, avente come variabile risposta il consumo giornaliero di gas e come variabile esplicativa la zona climatica, e si effettua un test ANOVA, che confronta il modello stimato con il modello avente come unica variabile esplicativa l'intercetta:

```
Analysis of Variance Table

              Df  Sum Sq Mean Sq F value    Pr(>F)
zona           4   11569 2892.15  286.88 < 2.2e-16 ***
Residuals 116159 1171052   10.08
```

**TABELLA 2.28: Anova del modello di regressione lineare semplice relativo alla zona climatica**

Analizzando i risultati del test ANOVA, è possibile affermare che la variabile relativa alla zona climatica in cui si trova il punto di riconsegna risulta essere significativa e che, quindi, risulta opportuno inserirla all'interno dei modelli statistici che successivamente si andranno a stimare e valutare.

### *2.7.2 Le informazioni atmosferiche*

Le informazioni atmosferiche relative al comune in cui è localizzato il punto di riconsegna sono reperibili sul sito *www.ilmeteo.it*. Reperire informazioni atmosferiche per 897 punti di riconsegna, tuttavia, comporta un onere molto

elevato, sia in termini di tempo che in termini di quantità di informazioni da recuperare.

Per questi motivi, si è ritenuto opportuno selezionare un certo numero di punti di riconsegna in base alle variabili precedentemente analizzate, al fine di replicare la variabilità del dataset iniziale, e recuperare le informazioni atmosferiche solo per questi punti; sono stati selezionati 279 punti sugli 897 iniziali, e per questi sono state recuperate le seguenti informazioni atmosferiche:

- **Temperatura media**: temperatura media registrata nel corso del giorno.
- **Temperatura minima**: picco minimo della temperatura registrato nel corso del giorno.
- **Temperatura massima**: picco massimo della temperatura registrato nel corso del giorno.

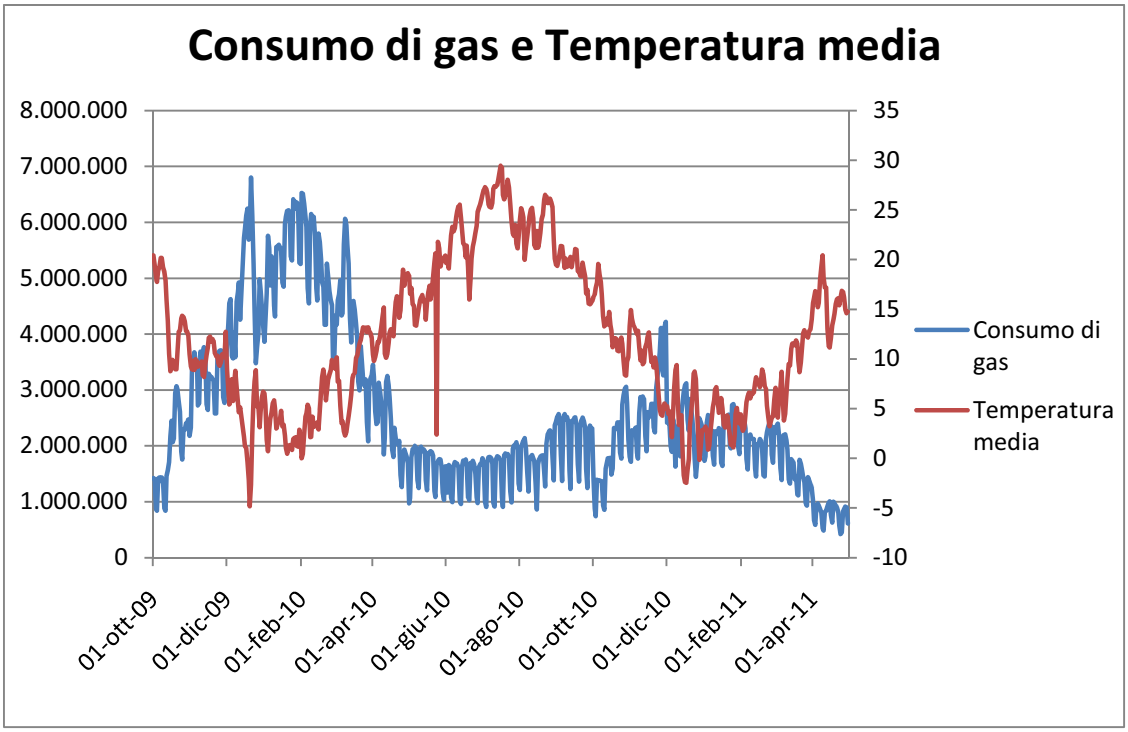



**GRAFICO 2.17: Consumo medio di gas e temperatura media**

Dal Grafico 2.17 si nota come la serie storica del consumo giornaliero complessivo di gas sembri essere correlata negativamente con la seria storica della temperatura media registrata: all'aumentare della temperatura media registrata il consumo giornaliero complessivo di gas diminuisce.

Altre analisi, relative ad ulteriori informazioni atmosferiche quali punto di rugiada, umidità, visibilità, velocità media e massima del vento, raffiche di vento e pressione atmosferiche, saranno inserite nella sezione Appendice (APPENDICE 1).

- **Fenomeni**: presenza di fenomeni atmosferici come pioggia, temporali, grandine, nebbia e neve (rilevati tramite la costruzione di opportune variabili dummy per ciascun fenomeno che assumono il valore 0 per indicare l'assenza del fenomeno e 1 per indicarne la presenza).

Risulta opportuno verificare se il consumo medio di gas è condizionato dalla presenza o meno dei fenomeni atmosferici precedentemente citati.

| | Pioggia | Temporale | Grandine | Nebbia | Neve |
|---|---|---|---|---|---|
| 0 | 7.292 | 7.717 | 7.509 | 6.799 | 7.239 |
| 1 | 7.982 | 4.883 | 5.610 | 10.803 | 15.388 |

**TABELLA 2.29: Consumo medio di gas per fattore atmosferico**

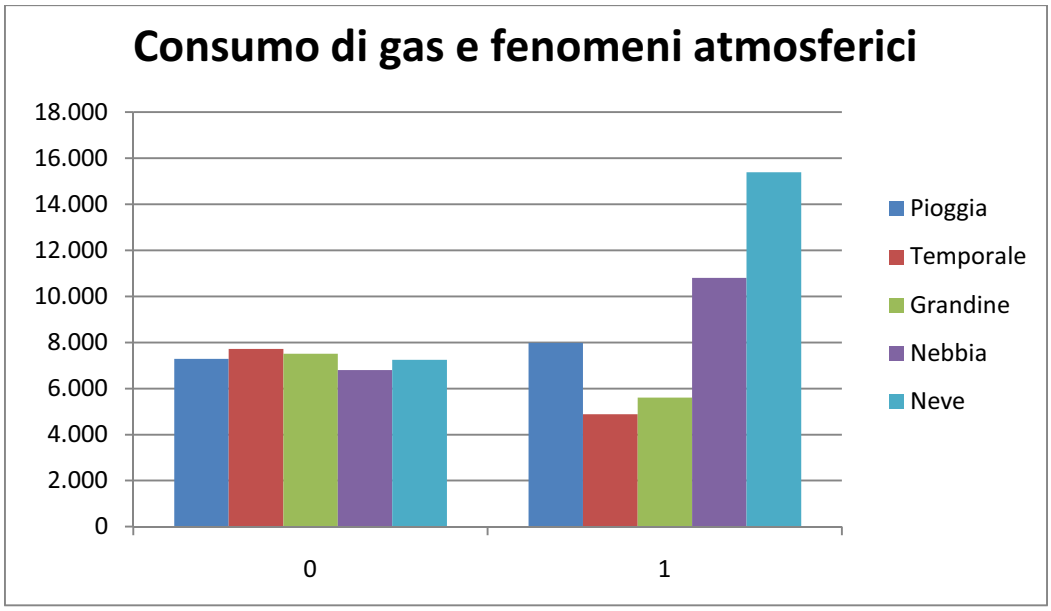



**GRAFICO 2.18: Consumo medio di gas per fattore atmosferico**

Il consumo medio di gas sembra maggiore quando piove, minore quando non c'è un temporale, indipendente dalla grandine, maggiore quando c'è nebbia e quando c'è neve.

Per verificare queste intuizioni è possibile effettuare dei test *t* per verificare l'ipotesi nulla di uguaglianza delle medie nei sottocampioni che vengono a crearsi in relazione alla presenza o assenza di un determinato fenomeno atmosferico (0 indica l'assenza del fenomeno atmosferico, 1 ne indica la presenza).

| | Pioggia | Temporale | Grandine | Nebbia | Neve |
|---|---|---|---|---|---|
| Media (0) < Media (1) | 0,0011 | 1,0000 | 0,6134 | 0,0000 | 0,0000 |
| Media (0) ≠ Media (1) | 0,0023 | 0,0000 | 0,7733 | 0,0000 | 0,0000 |
| Media (0) > Media (1) | 0,9989 | 0,0000 | 0,3866 | 1,0000 | 1,0000 |

**TABELLA 2.30: Verifica di ipotesi consumo medio di gas per fattore atmosferico**

Nella Tabella 2.30 vengono forniti i valori dei *p-value* a seconda dell'ipotesi alternativa considerata. E' quindi possibile affermare che, a un livello di significatività del 95%:

- Il consumo medio di gas quando non piove è minore di quando piove;
- Il consumo medio di gas quando non c'è temporale è maggiore di quando c'è temporale;
- Il consumo medio di gas è indipendente dalla presenza o assenza della grandine;
- Il consumo medio di gas quando c'è nebbia è maggiore di quando non ce n'è;
- Il consumo medio di gas quando nevica è maggiore di quando non nevica.

Quest'analisi, tuttavia, non è completa, in quanto vengono analizzate le situazioni di presenza o assenza di ogni singolo fenomeno atmosferico, mentre in uno stesso giorno potrebbero verificarsi contemporaneamente due o più fenomeni atmosferici (ad esempio "pioggia e temporale" o "nebbia e neve"); la valutazione dell'importanza dell'interazione tra i fenomeni atmosferici verrà rimandata alle analisi successive.

Poiché, in generale, il consumo giornaliero di gas è influenzato dalle condizioni atmosferiche, le variabili relative alle informazioni atmosferiche saranno tutte inserite nei modelli di regressione.

## 2.8 LA SUDDIVISIONE DEL DATASET: INSIEME DI STIMA E INISIEME DI VERIFICA

Dopo aver effettuato una prima veloce analisi descrittiva delle variabili che compongono il dataset, occorre chiarire in quale modo verranno trattati i dati. Il dataset, dopo aver selezionato un numero adeguato di punti per il reperimento delle informazioni atmosferiche, risulta composto da 121.764 osservazioni, sulle quali sono state rilevate le seguenti variabili:

- codice identificativo e descrizione del punto di riconsegna;
- provincia e regione in cui si colloca;
- il consumo giornaliero di gas;
- il giorno della settimana, il giorno del mese, il mese, l'anno e il tipo di giorno (feriale o festivo) in cui è stata effettuata la rilevazione del consumo;
- il numero di contratti giornalieri in vigore in un punto e la capacità prenotata dall'Utente;
- la destinazione d'uso del punto;
- la zona climatica;
- le informazioni atmosferiche.

Per evitare il problema di sovra-adattamento di un modello ai dati, e arrivare così a una scelta plausibile del modello previsivo, si è deciso di suddividere il dataset in un insieme di stima e un insieme di verifica, campionando temporalmente le osservazioni a nostra disposizione; in questo modo, dal dataset iniziale sono state escluse le ultime 20 osservazioni in ordine temporale di ciascun punto, componendo così l'insieme di stima, e tali osservazioni sono state collocate all'interno di un altro insieme detto di verifica.

L'insieme di stima risulta così composto da 116.164 osservazioni, mentre l'insieme di verifica da 5.600.

# CAPITOLO 3: IL *DATA MINING*

Il *Data Mining* è una disciplina recente, collocata al punto di intersezione tra varie aree scientifiche, specialmente la statistica, l'intelligenza artificiale e la gestione dei database; rappresenta l'attività di elaborazione in forma grafica o numerica di grandi raccolte di dati, con lo scopo di estrarre informazioni utili a chi detiene i dati stessi (*Azzalini-Scarpa*, *2009*).

Per prima cosa, si stimerà una serie di modelli di *Data Mining* sui dati disponibili; i modelli che si stimeranno, tuttavia, appartengono a un sottoinsieme di modelli nella vasta famiglia del *Data Mining*, e non tengono conto in alcun modo della correlazione seriale del consumo giornaliero di gas;è facile verificare, infatti, tramite un correlogramma, che i consumi giornalieri di gas, rilevati in un determinato punto di riconsegna, risultano caratterizzati da correlazione seriale, e che il consumo di giorni precedenti può, in qualche modo, influenzare il consumo di giorni successivi. Per questa serie di motivi, i modelli stimati potrebbero risultare poco efficienti e fornire previsioni non del tutto affidabili.

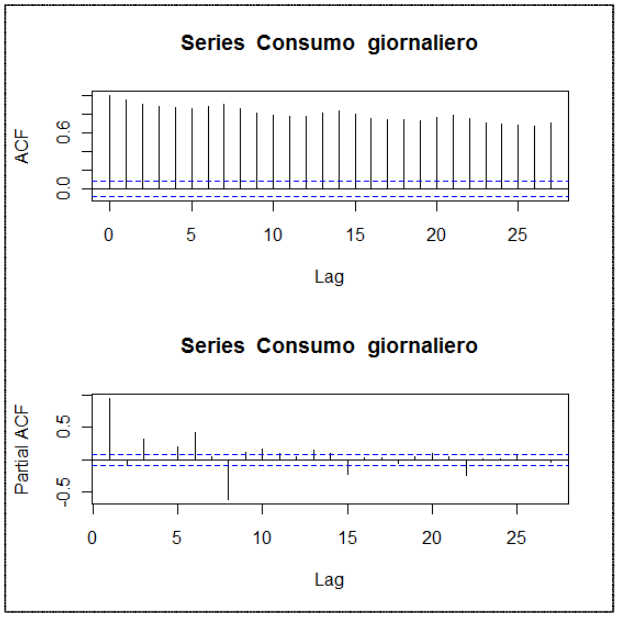



**GRAFICO 3.1: Correlogramma consumo giornaliero di gas**

## 3.1 LA VARIABILE RISPOSTA

La variabile risposta è senza dubbio il consumo giornaliero di gas, misurato in metri cubi. Le analisi descrittive svolte in precedenza, tuttavia, dimostrano che il consumo giornaliero di gas è una variabile caratterizzata da un'elevata variabilità interna, fattore che può influenzare la correttezza e l'efficienza delle stime e, di conseguenza, la bontà delle previsioni dei modelli stimati.

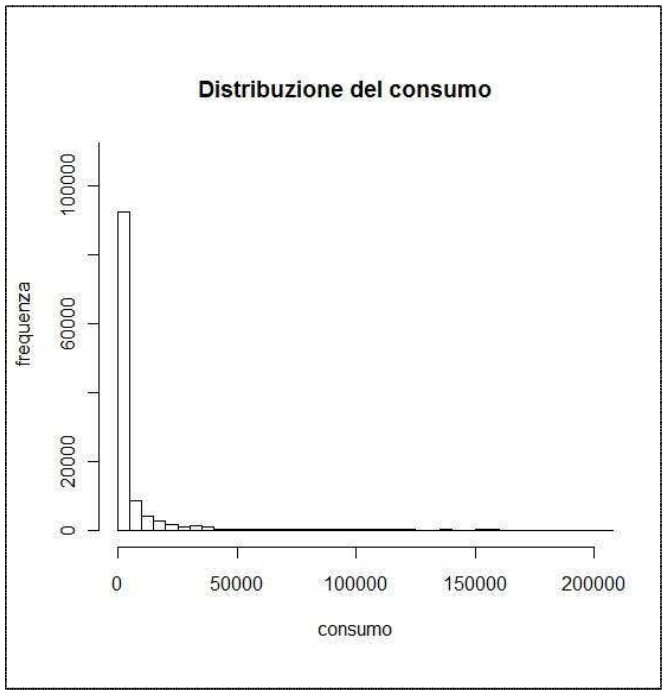



**GRAFICO 3.2: Consumo giornaliero di gas**

La distribuzione del consumo giornaliero di gas (Tabella 3.2), inoltre, risulta fortemente asimmetrica e molto concentrata intorno allo zero. In particolare, nell'insieme di stima sono presenti 8.230 osservazioni in cui il consumo giornaliero di gas è risultato pari a 0; questa caratteristica dei dati comporta qualche difficoltà nell'utilizzo dei tradizionali modelli di regressione, parametrici e non. È evidente, infatti, che la variabile risposta non può essere trattata come una variabile continua, ma presenta le caratteristiche di una variabile mista: è

infatti la combinazione di una componente continua per parte delle osservazioni e di una componente discreta e dicotomica per quell'altra parte di unità il cui consumo giornaliero di gas è risultato pari a 0. Per questi motivi, è ragionevole cercare di sfruttare al meglio anche questa informazione, al fine di costruire il modello di previsione più adeguato per il consumo giornaliero di gas; una possibilità è quella di organizzare un procedimento in due fasi: prima si adatta un modello per la probabilità che il consumo giornaliero di gas sia maggiore di 0 e poi, condizionatamente a questo evento, si costruisce un modello per il consumo quando assume valori positivi.

Occorre quindi creare una variabile risposta dicotomica, che assume il valore 0 quando il consumo di gas è pari a 0, e assume valore 1 quando il consumo di gas risulta maggiore di 0. Nell'insieme di stima vi sono 8.230 osservazioni in cui tale variabile assume il valore 0, e 107.934 osservazioni in cui assume valore 1; l'insieme di stima risulta quindi notevolmente sbilanciato, e tale sbilanciamento non consente di cogliere completamente le caratteristiche delle osservazioni presenti all'interno delle due classi. Per risolvere questo problema, è quindi necessario bilanciare l'insieme di stima, selezionando casualmente 8.230 osservazioni nella classe in cui la variabile dicotomica assume valore 1, facendo così in modo che la numerosità delle due classi risulti uguale, e riuscendo in questo modo a cogliere le caratteristiche delle due classi nella loro completezza.

## 3.2 L'ANALISI DI CLASSIFICAZIONE

L'obiettivo dell'analisi di classificazione è quello, a partire dal nuovo insieme di stima precedentemente descritto, di costruire una regola per comporre le informazioni disponibili sulle variabili esplicative, con l'obiettivo di allocare le osservazioni a una delle due classi. I modelli che si utilizzeranno a tal fine sono il modello di regressione lineare, il modello di regressione logistica, l'analisi

discriminante lineare, il modello additivo generalizzato (GAM), l'albero di classificazione, la rete neurale, il bagging e il boosting.

### *3.2.1 Il modello di regressione lineare*

Un modello di regressione lineare può essere descritto, considerando una variabile risposta $Y$ ed una o più variabili esplicative $X_1$, ... , $X_n$ ; la variabile $Y$ è rappresentabile tramite la relazione:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ ... \ + \beta_n X_n + \varepsilon$$

in cui $\beta_0$ ha il ruolo di intercetta, $\beta_1$, ... , $\beta_n$ sono i parametri che individuano la relazione tra $Y$ e le variabili esplicative, ed ε è la variabile dell'errore casuale, che rappresenta la parte di variabilità di $Y$ non riconducibile alla dipendenza da $X_1$, ... , $X_n$.

Utilizzando la notazione matriciale, si può scrivere:

$$y = X\beta + \varepsilon$$

dove $y = (y_1, \ ... \ , y_m)^T$ è il vettore contenente le $m$ osservazioni sulla variabile risposta, $X = [x_{ij}]$ è la matrice di regressione contenente i valori delle variabili esplicative, $\varepsilon = (\varepsilon_1, \ ... \ , \varepsilon_n)^T$ è il vettore contenente le $n$ componenti della variabile di errore, e $\beta = (\beta_0, \ ... \ , \beta_n)^T$ è il vettore dei parametri di regressione.

Si ipotizza dunque che i valori osservati della variabile risposta Y siano addizione di due componenti: la prima è la componente sistematica $X\beta$, detta anche parte deterministica del modello di regressione, che dipende dai valori assunti dalle variabili esplicative $X_i$, mentre la seconda è una componente di

disturbo aleatoria o erratica, l'errore $\varepsilon$, detta anche parte stocastica del modello di regressione, costituita da copie indipendenti di $N(0, \sigma^2)$ (*Pace-Salvan, 2001*).
Ipotesi fondamentali di questo modello sono, inoltre, che la relazione tra la variabile risposta e le variabili esplicative sia lineare nei parametri, e che le colonne della matrice $X$ di costanti note siano linearmente indipendenti (la matrice deve quindi essere di rango pieno e, di conseguenza, invertibile).

Per applicare il modello di regressione lineare nel caso di un'analisi di classificazione, si stima normalmente il modello di regressione lineare più adeguato, e successivamente si utilizza il valore $\hat{y} = \frac{1}{2}$ come soglia di discriminazione per la previsione delle due categorie, nel senso che si alloca un osservazione al gruppo 1 se il corrispondente $\hat{y}$ supera $\frac{1}{2}$ e al gruppo 0 nel caso contrario (*Azzalini-Scarpa, 2009*).

Si inizia quindi con la stima del modello lineare più adeguato, partendo da un modello avente come variabile risposta la variabile dicotomica relativa al consumo giornaliero di gas, resa numerica e non fattoriale, e come variabili esplicative tutte le informazioni a disposizione su ciascuna osservazione:

$$\begin{aligned} consumo = \ & \beta_0 + \beta_1 provincia + \beta_2 regione + \beta_3 gior_{sett} + \beta_4 mese \\ & + \beta_5 tipo_{giorno} + \beta_6 contratti + \beta_7 tipo_{punto} + \beta_8 zona \\ & + \beta_9 temp_{med} + \beta_{10} temp_{min} + \beta_{11} temp_{max} \\ & + \beta_{12} puntorugiada + \beta_{13} umidita + \beta_{14} visibilita \\ & + \beta_{15} vento_{media} + \beta_{16} vento_{max} + \beta_{17} raffica + \beta_{18} pressione \\ & + \beta_{19} pioggia + \beta_{20} temporale + \beta_{21} grandine + \beta_{22} nebbia \\ & + \beta_{23} neve + \beta_{24} capacita + \varepsilon \end{aligned}$$

In un modello di questo tipo, tuttavia, non tutte le stime risultano significative; vi sono infatti le stime di alcune variabili che risultano significativamente pari a 0. Di conseguenza, è opportuno effettuare un'operazione *passo a passo*, che

consenta di eliminare dal modello tutte quelle variabili che non contribuiscono a una diminuzione significativa della somma dei quadrati dei residui.

Si ottiene quindi il seguente modello:

$$\begin{aligned} consumo = {} & \beta_0 + \beta_1 provincia + \beta_2 regione + \beta_3 gior_{sett} + \beta_4 mese \\ & + \beta_5 tipo_{giorno} + \beta_6 contratti + \beta_7 tipo_{punto} + \beta_8 zona \\ & + \beta_9 temp_{med} + \beta_{12} puntorugiada + \beta_{14} visibilita \\ & + \beta_{21} grandine + \beta_{22} nebbia + \beta_{24} capacita + \varepsilon \end{aligned}$$

Dopo aver stimato il modello e aver classificato le osservazioni nell'insieme di verifica come spiegato in precedenza, occorre introdurre un indice sintetico della qualità del risultato ottenuto. I più immediati sono costituiti semplicemente dalla frazione di casi totali correttamente classificati, e dalla frazione di *falsi positivi* (consumi di gas pari a 0 classificati maggiori di 0) e *falsi negativi* (consumi di gas maggiori di 0 classificati pari a 0).

Per valutare questi indicatori, è possibile costruire una tabella a doppia entrata che conti il numero di casi previsti correttamente o meno, per ciascuna delle due modalità possibili; questa tabella è detta *matrice di confusione*.

| | | Osservati | |
|---|---|---|---|
| | | 0 | 1 |
| Previsti | 0 | 170 | 3.570 |
| | 1 | 353 | 1.507 |
| | Errore totale: 0,7005357 | | |
| | Falsi positivi: 0,1897849 | | |
| | Falsi negativi: 0,9545455 | | |

**TABELLA 3.1: Matrice di confusione modello di regressione lineare**

L’errore totale compiuto dal modello di regressione lineare risulta molto elevato e pari al 70% circa, mentre i falsi positivi e i falsi negativi risultano pari rispettivamente a circa il 19% e il 95%.

L’impiego di un modello di regressione lineare per la classificazione costituisce un po’ una forzatura: il campo di esistenza della variabile risposta è {0,1} e questo male si accorda con l’impostazione logica dei minimi quadrati, in quanto una funzione lineare di regressione non resta confinata entro questo insieme. Questo fatto si riflette sulla natura del termine di errore ε e provoca difficoltà nell’utilizzo di metodi inferenziali, in quanto gli errori non risultano omoschedastici e incorrelati. Per risolvere questi problemi occorre individuare una serie di modelli ad hoc per le analisi di classificazione.

### *3.2.2 Il modello di regressione logistico*

Il modello di regressione logistico prevede una variabile risposta di tipo categoriale, con due livelli indicati con “0” (consumo di gas pari a 0) e “1” (consumo di gas maggiore di 0), tale che una trasformata *logit* della probabilità di esito “0” ($P = \pi$) si può esprimere come combinazione lineare delle variabili esplicative:

$$
\begin{aligned}
logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right) \\
&= \beta_0 + \beta_1 provincia + \beta_2 regione + \beta_3 gior_{sett} + \beta_4 mese \\
&+ \beta_5 tipo_{giorno} + \beta_6 contratti + \beta_7 tipo_{punto} + \beta_8 zona \\
&+ \beta_9 temp_{med} + \beta_{10} temp_{min} + \beta_{11} temp_{max} \\
&+ \beta_{12} puntorugiada + \beta_{13} umidita + \beta_{14} visibilita \\
&+ \beta_{15} vento_{media} + \beta_{16} vento_{max} + \beta_{17} raffica + \beta_{18} pressione \\
&+ \beta_{19} pioggia + \beta_{20} temporale + \beta_{21} grandine + \beta_{22} nebbia \\
&+ \beta_{23} neve + \beta_{24} capacita + \varepsilon
\end{aligned}
$$

Come nel caso del modello lineare, tuttavia, non tutte le stime risultano significative; vi sono infatti le stime di alcune variabili che risultano significativamente pari a 0. Anche in questo caso, risulta quindi opportuno effettuare un'operazione *passo a passo*, che consenta di eliminare dal modello tutte quelle variabili che non contribuiscono a una diminuzione significativa della somma dei quadrati dei residui.

$$logit(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$
$$= \beta_0 + \beta_1 provincia + \beta_2 regione + \beta_3 gior_{sett} + \beta_4 mese$$
$$+ \beta_5 tipo_{giorno} + \beta_6 contratti + \beta_7 tipo_{punto} + \beta_8 zona$$
$$+ \beta_9 temp_{med} + \beta_{12} puntorugiada + \beta_{13} umidita$$
$$+ \beta_{14} visibilita + \beta_{21} grandine + \beta_{22} nebbia + \beta_{24} capacita + \varepsilon$$

Dopo aver stimato il modello e aver classificato le osservazioni nell'insieme di verifica, si inseriscono i risultati ottenuti nella *matrice di confusione*, al fine di valutare la bontà del modello.

| | | Osservati | |
|---|---|---|---|
| | | 0 | 1 |
| Previsti | 0 | 342 | 1.455 |
| | 1 | 181 | 3.622 |
| | Errore totale: 0,2921429 | | |
| | Falsi positivi: 0,0475940 | | |
| | Falsi negativi: 0,8096828 | | |

**TABELLA 3.2: Matrice di confusione modello di regressione logistico**

L’errore totale compiuto dal modello di regressione logistico diminuisce molto rispetto al modello di regressione lineare, risultando pari al 29% circa, mentre i falsi positivi e i falsi negativi risultano pari rispettivamente a circa il 5% e l’81%.

### *3.2.3 L’analisi discriminante lineare*

La regressione lineare e la regressione logistica non sono strumenti specificamente ideati per la classificazione, al contrario dell’analisi discriminante, che è un metodo ideato apposta per questa tipologia di analisi.
Ipotizzando che la popolazione complessiva sia composta da $k$ classi (in questo caso 2), questo metodo si basa sull’analisi di una *funzione discriminante*:

$$d_k(x_0) = log\ \pi_k + \log\ p_k(x_0)$$

dove $\pi_k$è il peso dato alla classe $k$, mentre $p_k$ rappresenta la funzione di densità della probabilità nella classe $k$. Il valore di $k$ che massimizza la funzione $d$ per un determinato $x_0$ è la classe a cui appartiene l’osservazione.

Per rendere operativo il procedimento, tuttavia, occorre stimare i pesi delle classi $\pi_k$, che, a meno di ulteriori informazioni, possono essere stimati con $\hat{\pi}_k = \frac{n_k}{n}$, e le funzioni di densità della probabilità $p_k(x)$, per le quali si aprono varie strade di approccio parametrico o non parametrico. Nell’analisi discriminante lineare, si adotta l’ipotesi parametrica più semplice: ciascuna densità $p_k(x)$ è Normale multipla, con parametri dipendenti da $k$, e tutte le matrici di varianza sono uguali ad una stessa $\Sigma$ (*Azzalini-Scarpa, 2009*).

Dopo aver applicato l’analisi discriminante lineare alle osservazione presenti nell’insieme di stima, e aver classificato le osservazioni nell’insieme di verifica, si inseriscono i risultati ottenuti nella *matrice di confusione*, al fine di valutare la bontà del modello.

| | | Osservati | |
|---|---|---|---|
| | | 0 | 1 |
| Previsti | 0 | 354 | 1.505 |
| | 1 | 169 | 3.572 |
| | Errore totale: 0,2989286 | | |
| | Falsi positivi: 0,04517509 | | |
| | Falsi negativi: 0,80957504 | | |

**TABELLA 3.3: Matrice di confusione analisi discriminante lineare**

L'errore totale compiuto dall'analisi discriminante lineare risulta molto simile a quello del modello di regressione logistico, essendo pari al 30% circa, mentre i falsi positivi e i falsi negativi risultano pari rispettivamente a circa il 5% e l'81%. Il metodo dell'analisi discriminante è appropriato per problemi di classificazione, permette di sfruttare eventuali informazioni a priori sulle osservazioni a disposizione ed è caratterizzato da una relativa semplicità di calcolo, fornendo risultati validi in un gran numero di casi e stabili rispetto all'ingresso di nuovi dati. Questo metodo, tuttavia, è costruito sotto ipotesi piuttosto dettagliate, e la stima, caratterizzata spesso da un elevato numero di parametri, non fornisce risultati robusti (*Azzalini-Scarpa, 2009*).

### *3.2.4 Gli alberi di classificazione*

Uno dei metodi più intuitivi per approssimare la probabilità che un'osservazione appartenga a una classe è quello di utilizzare una funzione a gradini, cioè una funzione costante a tratti su intervalli: questa è l'idea su cui si fonda la tecnica degli alberi di classificazione.

Questo metodo consente di classificare le osservazioni presenti nel dataset attraverso un apposito algoritmo, suddividendo ripetutamente, fino ad ottenere i risultati migliori, le osservazioni rispetto a determinate caratteristiche delle

variabili esplicative; la suddivisione produce una gerarchia ad albero, dove i sottoinsiemi di osservazioni vengono chiamati *nodi*, mentre il risultato finale viene rappresentato dalle *foglie*, che contengono la classe di appartenenza delle osservazioni.
Ogni elemento è quindi classificato seguendo un percorso lungo l'albero che porta dalla *radice*, il punto d'inizio, ad una *foglia*. I percorsi possibili sono rappresentati dai rami dell'albero, che forniscono una serie di regole di classificazione, espresse in funzione delle variabili dipendenti, allo scopo di costruire gruppi omogenei rispetto alla variabile risposta.

Nella stima degli alberi di classificazione, si ha il problema di trovare il giusto compromesso tra l'errore dovuto al campione su cui si lavora e quello del modello. Aumentando il numero di *foglie*, infatti, cresce l'errore dovuto alla specificità del campione, mentre diminuendo il numero di sottogruppi cresce l'errore del modello; occorre quindi individuare una procedura che permetta di minimizzare l'errore causato dal sovra-adattamento ai dati, mantenendo buona l'efficienza del modello. A tal fine l'insieme di stima verrà diviso casualmente in due sottoinsiemi: uno su cui verranno costruiti i modelli, composto da circa il 66% delle osservazioni, e un altro, composto dal restante 33%, su cui l'albero verrà potato (verrà ridotto il numero di foglie che aumentano l'instabilità del modello con lo scopo di minimizzare il sovra-adattamento ai dati).

Il metodo degli alberi di classificazione è utile soprattutto per la sua estrema semplicità: sarà infatti sufficiente seguire l'insieme di regole dettate dalla stima dell'albero per ottenere automaticamente la classificazione.
Nonostante questo indubbio pregio, però, esistono alcuni inconvenienti da tenere fortemente in considerazione: l'instabilità dei risultati (se si cambia, anche di poco, la composizione del dataset, ad esempio tenendolo costantemente aggiornato, risulta molto spesso diverso anche l'albero stimato) e l'impossibilità di compiere importanti operazioni statistiche, quali verifica di ipotesi e stima intervallare (*Azzalini-Scarpa, 2009*).

Per costruire l'albero più adeguato, si decide di utilizzare tutte le variabili esplicative a disposizione, sfruttando il fatto che gli alberi di classificazione selezionano automaticamente le variabili più importanti per la classificazione delle osservazioni.

L'albero verrà fatto crescere nella porzione dell'insieme stima riservato alla crescita, con l'inserimento di determinati parametri di controllo (il numero minimo di osservazioni presenti in ciascun *nodo* dovrà essere pari a 2 e il valore minimo di devianza interna a ciascun *nodo* dovrà essere pari a 0,001); una volta che l'albero avrà raggiunto la crescita massima, esso sarà potato nella porzione dell'insieme di stima riservata alla potatura.

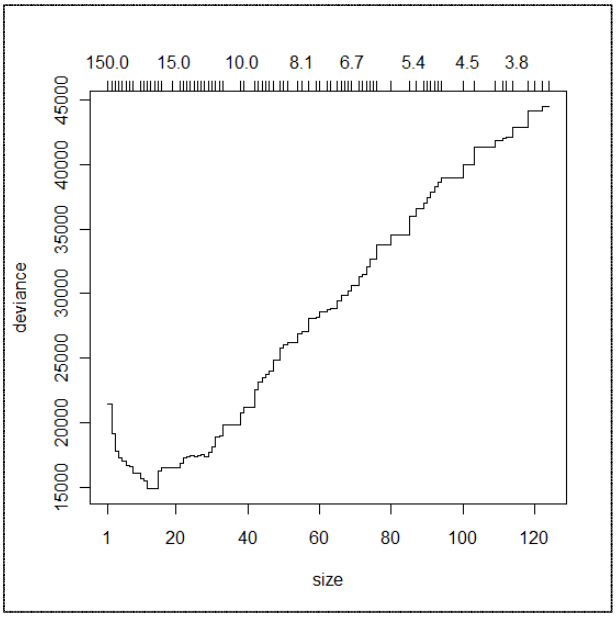



**GRAFICO 3.3: Scelta della size dell'albero**

Il Grafico 3.3 evidenzia che l'albero ottimale è composto da 13 foglie; la struttura è la seguente:

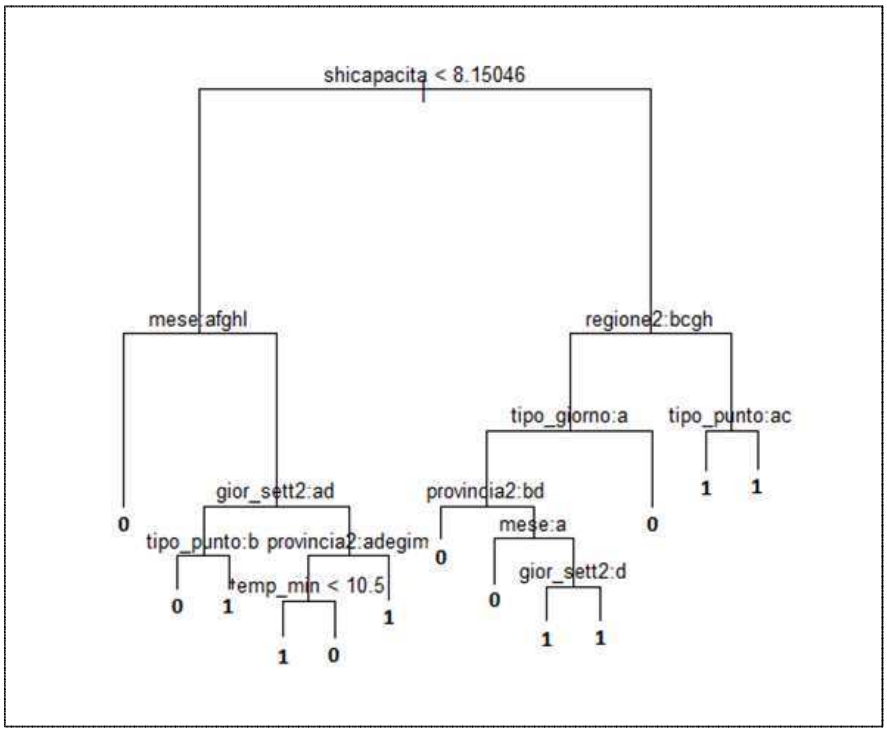



**GRAFICO 3.4: Albero di classificazione**

Dopo aver ottenuto l'albero di classificazione ottimale nell'insieme di stima, e dopo aver classificato le osservazioni nell'insieme di verifica, si inseriscono i risultati ottenuti nella *matrice di confusione*, al fine di valutare la bontà del modello.

| | | Osservati | |
|---|---|---|---|
| | | 0 | 1 |
| Previsti | 0 | 211 | 1.064 |
| | 1 | 312 | 4.013 |
| | Errore totale: 0,2457143 | | |
| | Falsi positivi: 0,07213873 | | |
| | Falsi negativi: 0,83450980 | | |

**TABELLA 3.4: Matrice di confusione albero di classificazione**

L’errore totale compiuto dall’albero di classificazione risulta inferiore rispetto ai modelli precedenti, essendo pari al 25% circa, mentre i falsi positivi e i falsi negativi risultano pari rispettivamente a circa il 7% e l’83%.

### *3.2.5 Le Reti Neurali*

Una rete neurale è sostanzialmente uno schema di regressione a due stadi, generalmente di tipo non lineare, in cui sono messe in relazione $p$ variabili esplicative (o di *input*) con $q$ variabili risposta (o di *output*), attraverso uno strato $r$ di *variabili latenti* (quindi non osservate), che si frappongono tra i due precedenti gruppi (nel senso che le variabili esplicative influenzano le variabili latenti e queste ultime influenzano le variabili risposta).
Indicate con $x_{\mathrm{h}}$ , $z_j$ e $y_k$ rispettivamente la generica variabile di input, latente e di output, e poste le “variabili costanti” $x_0 = z_0 = 1$, la relazione può essere espressa come:

$$z_j = f_0\left(\sum_{h\to j} \alpha_{hj} x_h\right) \qquad , \qquad y_k = f_1\left(\sum_{j\to k} \beta_{jk} z_j\right)$$

dove gli $\alpha_{hj}$ e $\beta_{jk}$ sono parametri da stimare, e le somme si estendono agli indici relativi a variabili per le quali è prevista una relazione di dipendenza.
Le funzioni di attivazione $f_0$ e $f_1$ sono solitamente poste rispettivamente entrambe pari alla funzione logistica, in quanto $f_1$ deve avere come codominio l’intervallo (0,1).

Ci sono due elementi da determinare: il numero $r$ di unità nello strato latente, e l’insieme di coefficienti $\alpha$ e $\beta$; non ci sono criteri facilmente utilizzabili per la scelta di $r$, se non quello di provarne diversi e confrontare i risultati ottenuti.

Una volta fissato $r$, occorre stimare i coefficienti $\alpha$ e $\beta$ sulla base di osservazioni campionarie. Essi si ottengono minimizzando una determinata funzione obiettivo, che ingloba un termine di penalizzazione per evitare problemi di sovra-adattamento; tale termine di penalizzazione prende il nome di *weight decay*, che viene normalizzato moltiplicandolo per un opportuno parametro di regolazione $\lambda$. La minimizzazione richiede un procedimento di ottimizzazione numerica e l'algoritmo più comunemente associato a questo problema è detto *algoritmo di back-propagation*, che consente di aggiornare successivamente le stime dei parametri quando nuove osservazioni entrano nel dataset di partenza.

Le reti neurali offrono una famiglia di modelli molto flessibili, in quanto le stime ottenute possono essere sequenzialmente aggiornate quando arrivano nuovi dati.
I modelli di questa classe, tuttavia, risultano fortemente arbitrari, in quando non vi sono criteri forti e oggettivi per scegliere il numero $r$ di nodi dello strato latente e il parametro di regolazione $\lambda$, difficili da stimare e di difficile interpretazione, soprattutto al crescere di $r$ (*Azzalini-Scarpa, 2009*).

Il numero di unità $r$ nello strato latente e il *weight decay*, come visto in precedenza, devono essere impostati dall'analista; per ottenere i valori ottimali sono state stimate più reti neurali con valori differenti di unità nello strato latente e di *weight decay*, e sono stati poi scelti quelli la cui rete corrispondente minimizza l'errore di classificazione nell'insieme di verifica.

| *r* | *weight decay* | *errore* | *falsi positivi* | *falsi negativi* |
|---|---|---|---|---|
| 3 | 0,001 | 33,9% | 3,9% | 81,9% |
| 3 | 0,002 | 21,8% | 7,0% | 81,3% |
| 3 | 0,005 | 20,4% | 5,5% | 76,3% |
| 3 | 0,010 | 35,7% | 3,6% | 82,3% |
| 3 | 0,022 | 22,8% | 4,8% | 77,0% |
| 3 | 0,046 | 9,4% | 9,2% | 66,7% |
| 3 | 0,100 | 25,2% | 4,6% | 78,4% |

| | | | | |
|---|---|---|---|---|
| 3 | 0,215 | 21,5% | 5,2% | 76,6% |
| 3 | 0,464 | 25,6% | 5,0% | 79,4% |
| 3 | 1,000 | 25,8% | 4,8% | 79,1% |
| 4 | 0,001 | 38,2% | 5,1% | 84,7% |
| 4 | 0,002 | 22,5% | 7,2% | 82,4% |
| 4 | 0,005 | 21,9% | 5,4% | 77,3% |
| 4 | 0,010 | 29,1% | 4,5% | 80,5% |
| 4 | 0,022 | 36,1% | 4,2% | 83,1% |
| 4 | 0,046 | 26,7% | 4,2% | 78,7% |
| 4 | 0,100 | 22,0% | 4,3% | 75,5% |
| **4** | **0,215** | **20,2%** | **3,8%** | **72,9%** |
| 4 | 0,464 | 21,0% | 4,6% | 75,0% |
| 4 | 1,000 | 21,2% | 4,7% | 75,5% |
| 5 | 0,001 | 32,0% | 3,1% | 80,3% |
| 5 | 0,002 | 26,6% | 4,9% | 79,8% |
| 5 | 0,005 | 18,6% | 4,6% | 72,4% |
| 5 | 0,010 | 18,2% | 6,0% | 75,0% |
| 5 | 0,022 | 20,9% | 4,3% | 74,4% |
| 5 | 0,046 | 20,5% | 4,2% | 73,8% |
| 5 | 0,100 | 25,2% | 4,8% | 78,8% |
| 5 | 0,215 | 19,3% | 5,0% | 74,1% |
| 5 | 0,464 | 19,2% | 4,3% | 72,6% |
| 5 | 1,000 | 20,6% | 4,8% | 75,1% |

**TABELLA 3.5: Reti Neurali – Scelta di *r* e del *weight decay***

Considerando l'errore di classificazione, i falsi positivi e i falsi negativi, il numero ottimale di unità nello strato latente è pari a *r=4*, con un corrispondente *weight decay=0,215*, e nel modello saranno inserite tutte le informazioni a disposizione sulle osservazioni. La classificazione delle osservazioni nell'insieme di verifica porta alla seguente *matrice di confusione*:

| | | Osservati | |
|---|---|---|---|
| | | 0 | 1 |
| Previsti | 0 | 360 | 970 |
| | 1 | 163 | 4.107 |
| | Errore totale: 0,2023214 | | |
| | Falsi positivi: 0,0381733 | | |
| | Falsi negativi: 0,7293233 | | |

**TABELLA 3.6: Matrice di confusione Rete Neurale**

L'errore totale compiuto dalla rete neurale risulta inferiore rispetto ai modelli precedenti, essendo pari al 20% circa, mentre i falsi positivi e i falsi negativi risultano pari rispettivamente a circa il 4% e l'73%.

### *3.2.6 Il Bagging*

Per tentare di migliorare la capacità previsiva dei modelli, una possibilità è quella di combinare le previsioni ottenute da metodi diversi. Sia $C(x)$ un classificatore ottenuto con uno dei metodi precedentemente utilizzati; seguendo la procedura che in statistica prende generalmente il nome di *bootstrap*, si consideri il campione ottenuto estraendo $n$ volte con ripetizione gli elementi dell'insieme di stima. Utilizzando questo nuovo campione è possibile ottenere un nuovo classificatore $C_1^*(x)$ che per una fissata $x$ generica sarà diverso dal precedente. Analogamente si possono estrarre B diversi campioni di dimensione $n$ attraverso il ricampionamento dell'insieme di stima, ottenendo da questi B nuovi classificatori $C_b^*(x)$. Si introduce in questo modo un nuovo classificatore che, per ogni $x$, sia un indicatore medio dei risultati ottenuti da ciascuno dei $C_b^*(x)$ su quella stessa $x$:

$$C_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B} C_b^*(x)$$

In questo modo, si classifica l'unità nella classe associata al valore 1 se $C_{bag}(x) > \frac{1}{2}$ e a quella associata al valore 0 altrimenti. Tale criterio è indicato come *voto di maggioranza* e il modello risultante dall'intera procedura viene chiamato *boostrap aggregating* (*bagging*). L'errore di classificazione del nuovo modello risulta mediamente inferiore rispetto a quello di ciascuno dei modelli originali (*Azzalini-Scarpa, 2009*).

Si stimeranno due modelli di *bootstrap aggregating*, nei quali verranno inserite tutte le variabili esplicative a disposizione, e in cui saranno estratti rispettivamente B=20 e B=100 diversi campioni di dimensione *n* attraverso il ricampionamento dell'insieme di stima.

La classificazione delle osservazioni dell'insieme di verifica, tramite il modello in cui sono stati estratti B=20 diversi campioni di dimensione *n* attraverso il ricampionamento dell'insieme di stima, determina la seguente *matrice di confusione*:

| | | Osservati | |
|---|---|---|---|
| | | 0 | 1 |
| Previsti | 0 | 384 | 537 |
| | 1 | 139 | 4.540 |
| | Errore totale: 0,1207143 | | |
| | Falsi positivi: 0,0297072 | | |
| | Falsi negativi: 0,5830619 | | |

**TABELLA 3.7: Matrice di confusione Bagging (B=20)**

La classificazione delle osservazioni dell'insieme di verifica, tramite il modello in cui sono stati estratti B=100 diversi campioni di dimensione $n$ attraverso il ricampionamento dell'insieme di stima, invece, determina la seguente *matrice di confusione*:

| | | Osservati | |
|---|---|---|---|
| | | 0 | 1 |
| Previsti | 0 | 382 | 504 |
| | 1 | 141 | 4.573 |
| | Errore totale: 0,1151786 | | |
| | Falsi positivi: 0,02991090 | | |
| | Falsi negativi: 0,56884876 | | |

**TABELLA 3.8: Matrice di confusione Bagging (B=100)**

Il modello di tipo *bagging* che classifica meglio le osservazioni dell'insieme di verifica risulta essere quello in cui sono stati estratti B=100 diversi campioni di dimensione $n$ attraverso il ricampionamento dell'insieme di stima, con un errore di classificazione pari a circa il 12%, e con le percentuali di falsi positivi e falsi negativi pari rispettivamente a circa il 3% e il 57%.

### *3.2.7 Il Boosting*

La strategia attuata dal *bagging* si basa sulla logica di combinare i risultati di un modello ottenuto usando diversi insiemi di dati, estratti attraverso un campionamento casuale che assegna a ciascuna unità la stessa probabilità di entrare nel campione.

Analogamente, il *boosting* consiste nel combinare i risultati di un modello adattato utilizzando diversi insiemi di dati, selezionando però le unità da inserire nel campione attraverso l'assegnazione a ciascuna unità di una diversa

probabilità di entrata; in particolare, si assegna maggior peso alle osservazioni che in precedenza sono state classificate peggio e, in questo modo, si intende migliorare la prestazione del nuovo modello, intervenendo su quelle osservazioni in cui il classificatore originale presentava maggiori difficoltà di classificazione. La procedura è iterativa, e l'insieme dei pesi viene aggiornato ad ogni iterazione, in funzione del tasso di errore globale ottenuto su un insieme di verifica; al termine della procedura si identifica un nuovo classificatore attraverso *voto di maggioranza* tra tutti i modelli costruiti nel corso delle iterazioni.
Procedure di tipo *boosting* hanno dimostrato una notevole capacità di produrre classificatori accurati in un'ampia varietà di situazioni, che dimostrano proprietà statistiche che giustificano le ottime prestazioni empiriche (*Azzalini-Scarpa, 2009*).

La classificazione delle osservazioni dell'insieme di stima, tramite un modello di tipo *boosting*, determina la seguente *matrice di confusione*:

| | | Osservati | |
|---|---|---|---|
| | | 0 | 1 |
| Previsti | 0 | 343 | 1.063 |
| | 1 | 180 | 4.014 |
| | Errore totale: 0,2219643 | | |
| | Falsi positivi: 0,04291845 | | |
| | Falsi negativi: 0,75604552 | | |

**TABELLA 3.9: Matrice di confusione Boosting**

Il modello di tipo *boosting* classifica le osservazioni dell'insieme di verifica con un errore di classificazione pari a circa il 22%, e con le percentuali di falsi positivi e falsi negativi pari rispettivamente a circa il 4% e il 76%.

### *3.2.8 Il confronto tra i modelli*

Una volta stimati i modelli di classificazione, occorre valutare i risultati ottenuti, al fine di determinare quale sia il modello che classifica meglio le unità nell'insieme di verifica, confrontando l'errore di classificazione e le percentuali di falsi positivi e falsi negativi, che derivano dalla classificazione delle osservazioni nell'insieme di verifica.

| | *errore* | *falsi positivi* | *falsi negativi* |
|---|---|---|---|
| Modello lineare | 70,1% | 19,0% | 95,4% |
| Modello logistico | 29,2% | 4,8% | 81,0% |
| Analisi discriminante lineare | 29,9% | 4,5% | 81,0% |
| Alberi di classificazione | 24,6% | 7,2% | 83,4% |
| Rete Neurale | 20,2% | 3,8% | 72,9% |
| ***Bagging (B=100)*** | **11,5%** | **3,0%** | **56,9%** |
| *Boosting* | 22,2% | 4,3% | 75,6% |

**TABELLA 3.10: Confronto tra i modelli di classificazione**

Il modello migliore risulta il modello di tipo *bagging*, in cui sono stati estratti B=100 diversi campioni di dimensione $n$ attraverso il ricampionamento dell'insieme di stima.

Anche se per alcuni aspetti è conveniente disporre di indicatori molto semplici e globali delle prestazioni di una procedura di classificazione, quali quelli evidenziati nella Tabella 3.10, in realtà è utile cercare di valutare in modo più analitico la capacità di classificazione dei vari metodi. Un metodo particolarmente diffuso è costituito dalla *funzione lift*, che fornisce una misura di miglioramento ottenuto dal modello in considerazione, rispetto alla classificazione casuale con probabilità uniforme pari alla frazione osservata nell'insieme di verifica (*Azzalini-Scarpa, 2009*).

Il confronto delle curve *lift* dei modelli di classificazione stimati in questo capitolo dà la seguente rappresentazione grafica:

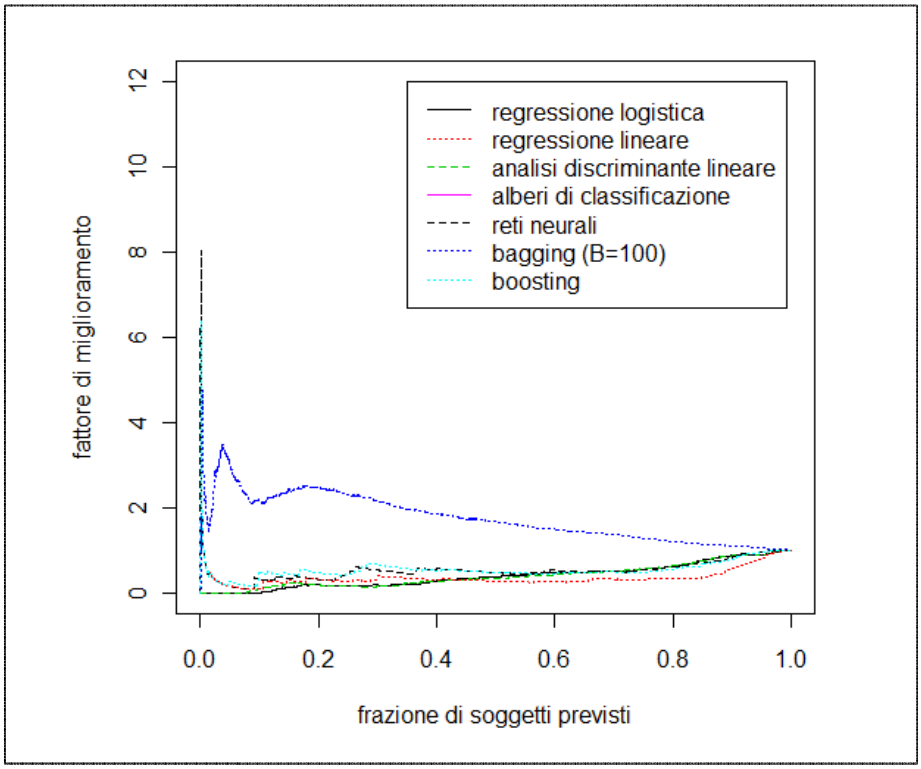



**GRAFICO 3.5: Confronto curve *lift***

Anche il confronto delle curve *lift* dei modelli stimati conferma che il modello migliore per la classificazione è il modello di tipo *bagging*, in cui sono stati estratti B=100 diversi campioni di dimensione $n$ attraverso il ricampionamento dell'insieme di stima.

Questo modello sarà quindi utilizzato per classificare le osservazione dell'insieme di verifica; le osservazioni che saranno assegnate alla classe 0 avranno un consumo giornaliero di gas previsto pari a 0, mentre per le osservazioni che saranno segnate alla classe 1 occorrerà prevedere il consumo giornaliero di gas attraverso un appropriato modello di regressione. La scelta del

modello di regressione adeguato sarà oggetto di discussione del prossimo paragrafo.

## 3.3 L'ANALISI DI REGRESSIONE

Dopo aver classificato le osservazioni nell'insieme di verifica, valutando se il consumo giornaliero di gas previsto fosse uguale a 0 (classe 0) o maggiore di 0 (classe 1), occorre prevedere un valore per le osservazioni classificate nella classe 1. Per fare questo è necessaria un'analisi di regressione sull'insieme di stima, privato delle 8.203 osservazioni in cui il consumo giornaliero di gas è stato rilevato pari a 0, poiché è di interesse analizzare l'andamento dei consumi quando essi risultano strettamente maggiori di 0.

L'insieme di stima risulta quindi composto da 107.934 osservazioni, e i modelli che si andranno a stimare e valutare sono il modello di regressione lineare, il modello *MARS* (*Multi-Adaptive Regression Splines*), il modello *GAM* (*Generalized Additive Model*), il modello *Projection Pursuit*, e la Rete Neurale.

### *3.3.1 Il modello di regressione lineare*

La stima del modello lineare più adeguato parte da un modello, avente come variabile risposta il consumo giornaliero di gas, e come variabili esplicative tutte le informazioni a disposizione su ciascuna osservazione:

$$
\begin{aligned}
cons = \ & \beta_0 + \beta_1 provincia + \beta_2 regione + \beta_3 gior_{sett} + \beta_4 mese \\
& + \beta_5 tipo_{giorno} + \beta_6 contratti + \beta_7 tipo_{punto} + \beta_8 zona \\
& + \beta_9 temp_{med} + \beta_{10} temp_{min} + \beta_{11} temp_{max} \\
& + \beta_{12} puntorugiada + \beta_{13} umidita + \beta_{14} visibilita \\
& + \beta_{15} vento_{media} + \beta_{16} vento_{max} + \beta_{17} raffica + \beta_{18} pressione \\
& + \beta_{19} pioggia + \beta_{20} temporale + \beta_{21} grandine + \beta_{22} nebbia \\
& + \beta_{23} neve + \beta_{24} capacita + \varepsilon
\end{aligned}
$$

In questo caso, però, la variabile risposta non ha distribuzione normale, neppure approssimata, e si pone il problema della trasformazione dei dati. Box e Cox, nel 1964, hanno proposto un metodo iterativo per individuare quale trasformazione dei dati può meglio normalizzare la loro distribuzione. Il metodo ricorre a una famiglia di trasformazioni di potenze mediante la formula:

$$
\begin{cases}
y_{tras} = \dfrac{y^{\lambda} - 1}{\lambda} & se\ \lambda \neq 0 \\
y_{tras} = \log(y) & se\ \lambda = 0
\end{cases}
$$

con λ che varia da -3 a +3. Il valore di λ che meglio normalizza la distribuzione è quello che rende massima la funzione L (nota come log-likelihood function):

$$
L = \frac{n-1}{2} \log(s^2_{tras}) + (\lambda - 1) \frac{n-1}{n} \sum_{i=1}^{n} y_i
$$

dove $s^2_{tras}$ è la varianza dei dati trasformati; inoltre è possibile calcolare l'intervallo di confidenza per λ, entro il quale è conveniente scegliere la trasformazione più adeguata. Benché possa teoricamente assumere qualsiasi

valore da –3 a +3 in una scala continua, in pratica λ ha significato pratico solo per alcuni valori (*Ricci, 2006*).

Il Grafico 3.6 mostra il risultato ottenuto con la trasformata di *Box-Cox*: in ascissa è riportato il valore che identifica la trasformazione, mentre in ordinata è riportato il valore della log-verosimiglianza.

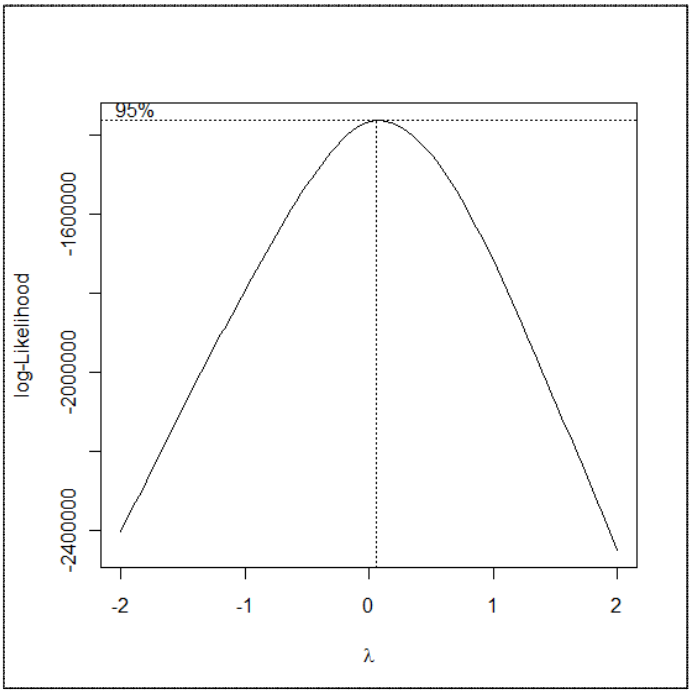



**GRAFICO 3.6: Trasformata di *Box-Cox***

Dal Grafico 3.6, inoltre, si deduce che il valore ottimale per λ è pari a circa 0,06. Si stima quindi il modello di regressione lineare con la variabile trasformata; in questo modello, tuttavia, non tutte le stime risultano significative; vi sono infatti le stime di alcune variabili che risultano significativamente pari a 0. Di conseguenza, è opportuno effettuare un'operazione *passo a passo*, che consenta di eliminare dal modello tutte quelle variabili che non contribuiscono a una diminuzione significativa della somma dei quadrati dei residui.

Si ottiene quindi il seguente modello con le relative stime:

$$cons = \beta_0 + \beta_1 provincia + \beta_2 regione + \beta_3 gior_{sett} + \beta_4 mese + \beta_5 tipo_{giorno} + \beta_6 contratti + \beta_7 tipo_{punto} + \beta_9 temp_{med} + \beta_{11} temp_{max} + \beta_{12} puntorugiada + \beta_{14} visibilita + \beta_{15} vento_{media} + \beta_{16} vento_{max} + \beta_{17} raffica + \beta_{20} temporale + \beta_{22} nebbia + \beta_{24} capacita + \varepsilon$$

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.887e+00  6.347e-02 -29.729  < 2e-16 ***
provincia2    -3.359e-01  2.498e-02 -13.447  < 2e-16 ***
provincia3     1.661e-01  1.858e-02   8.941  < 2e-16 ***
provincia4     3.523e-01  2.813e-02  12.521  < 2e-16 ***
provincia5     2.799e-01  5.558e-02   5.036 4.77e-07 ***
provincia6     2.973e-01  2.197e-02  13.533  < 2e-16 ***
provincia7    -1.091e-01  2.641e-02  -4.131 3.62e-05 ***
provincia8    -1.190e-01  2.201e-02  -5.406 6.46e-08 ***
provincia9    -1.639e-01  2.070e-02  -7.917 2.45e-15 ***
provincia10    3.243e-01  2.676e-02  12.119  < 2e-16 ***
provincia11    3.306e-01  5.480e-02   6.031 1.63e-09 ***
provincia12   -1.840e+00  9.271e-02 -19.848  < 2e-16 ***
provincia13   -1.071e+00  3.874e-02 -27.641  < 2e-16 ***
provincia14   -1.691e-01  6.258e-02  -2.702 0.006893 **
regione2       3.042e-01  2.947e-02  10.323  < 2e-16 ***
regione3       4.166e-01  3.071e-02  13.566  < 2e-16 ***
regione4       1.623e-01  3.264e-02   4.972 6.63e-07 ***
regione5       2.678e-01  4.403e-02   6.082 1.19e-09 ***
regione6      -5.084e-01  7.085e-02  -7.176 7.24e-13 ***
regione7       2.641e-01  3.260e-02   8.100 5.53e-16 ***
regione8       1.214e+00  4.144e-02  29.299  < 2e-16 ***
regione9       1.858e-01  4.529e-02   4.102 4.11e-05 ***
regione10      2.103e-01  3.020e-02   6.963 3.35e-12 ***
gior_sett2     1.676e-01  3.526e-02   4.752 2.02e-06 ***
gior_sett3     1.295e-01  3.683e-02   3.516 0.000438 ***
gior_sett4    -5.336e-01  3.628e-02 -14.705  < 2e-16 ***
```

```
meseAprile     6.930e-01  3.069e-02  22.578  < 2e-16 ***
meseDicembre   1.332e+00  3.772e-02  35.309  < 2e-16 ***
meseFebbraio   1.454e+00  3.613e-02  40.239  < 2e-16 ***
meseGennaio    1.424e+00  3.860e-02  36.887  < 2e-16 ***
meseGiugno     3.020e-01  2.901e-02  10.411  < 2e-16 ***
meseLuglio     3.519e-01  2.822e-02  12.472  < 2e-16 ***
meseMaggio     3.406e-01  3.011e-02  11.310  < 2e-16 ***
meseMarzo      1.299e+00  3.276e-02  39.640  < 2e-16 ***
meseNovembre   1.153e+00  3.217e-02  35.852  < 2e-16 ***
meseOttobre    5.301e-01  2.914e-02  18.190  < 2e-16 ***
meseSettembre  2.351e-01  3.050e-02   7.707 1.30e-14 ***
tipo_giorno1  -8.333e-01  3.295e-02 -25.289  < 2e-16 ***
contratti     -2.341e-02  6.929e-03  -3.379 0.000729 ***
tipo_punto2    1.142e+00  1.405e-02  81.337  < 2e-16 ***
tipo_punto3    1.889e+00  2.061e-02  91.610  < 2e-16 ***
temp_media    -3.817e-02  3.524e-03 -10.830  < 2e-16 ***
temp_max      -1.583e-02  2.582e-03  -6.131 8.74e-10 ***
puntorugiada  -8.793e-03  2.162e-03  -4.068 4.75e-05 ***
visibilita    -3.652e-05  1.424e-05  -2.564 0.010353 *
vento_media    8.181e-03  2.324e-03   3.520 0.000431 ***
vento_max      3.139e-03  8.838e-04   3.552 0.000383 ***
raffica       -1.983e-03  5.787e-04  -3.427 0.000611 ***
temporale1    -1.501e-01  2.048e-02  -7.329 2.34e-13 ***
nebbia1        4.039e-02  1.348e-02   2.996 0.002738 **
capacita3      8.651e-01  2.112e-03 409.604  < 2e-16 ***

Multiple R-squared: 0.846
```

**TABELLA 3.11: Stime del modello lineare**

La variabilità spiegata dal modello risulta pari a circa l’84,6% della variabilità complessiva; con i dati a disposizione, si può senza dubbio considerare un risultato abbastanza soddisfacente.

Per valutare l’effettiva qualità delle previsioni, evitando il problema di sovra-adattamento ai dati, occorre valutare l’errore di previsione nell’insieme di verifica, che risulta elevato e pari a circa 266.683.

### *3.3.2 I modelli non parametrici: Mars*

Il modello di regressione lineare, trattato in precedenza, può essere riscritto nella formula più generale:

$$y = f(x; \beta) + \varepsilon$$

Il metodo utilizzato per ottenere le soluzioni del modello di regressione lineare è di tipo parametrico, in quanto la funzione $f$, che esprime la relazione tra la variabile risposta e le variabili esplicative, è considerata uguale ad una retta, ponendo di fatto una restrizione non da poco.

I modelli non parametrici, invece, non considerano la funzione $f$ come appartenente a una predeterminata classe parametrica di funzioni, ma lascia che sia essa stessa ad adattarsi ai dati; questo non significa che seguirà esattamente le osservazioni, in quando ciò comporterebbe un evidente sovra-adattamento ai dati: verranno quindi posti alcuni vincoli, entro i quali la funziona potrà "liberamente esprimersi".

Uno dei metodi per la costruzione di modelli non parametrici è conosciuto con il nome di *spline di regressione*: vengono fissati sull'asse della ascisse un numero $k$ di punti, chiamati *nodi*, tali che: $A_1 < A_2 < A_3 < \cdots < A_k$.

La funzione che si costruirà dovrà passare esattamente all'interno dei punti fissati, mentre negli intervalli sarà una funzione di tipo polinomiale.

Un problema che si può riscontrare è causato dal fatto che, con l'aumentare del numero di dimensioni, la distanza tra le osservazioni cresce esponenzialmente, e diventa sempre più difficile per i modelli cogliere le relazioni al loro interno. Questa difficoltà nel cogliere in modo accurato la funzione $f(x)$ quando si è in presenza di un elevato numero di dimensioni (variabili esplicative) è conosciuta con il nome di *maledizione della dimensionalità*, e nell'analisi dei modelli non si potrà prescindere da tale problema (*Azzalini-Scarpa, 2009*).

I modelli di tipo *spline*, come quelli precedentemente descritti, sono molto difficili da utilizzare nel caso in cui vi sia un elevato numero di variabili esplicative; risulta quindi indispensabile utilizzare una procedura che fornisca una selezione automatica delle variabili da trasformare e, soprattutto, del numero e della posizione dei nodi da creare.
Permettono di compiere tale operazione le *MARS* (*Multivariate Adaptive Regression Spline*), che consistono in una particolare specificazione iterativa delle *spline di regressione*, finalizzata alla modellazione nel caso in cui siano presenti numerose variabili esplicative.

Il modello *MARS* ha la seguente forma:

$$f(x) = \beta_0 + \sum_{k=1}^{K} \beta_k \, h_k(x)$$

dove le $h_k$ rappresentano le diverse funzioni, dette basi, che sommate permettono di identificare il modello; una volta ottenute le $h_k$, i parametri $\beta_k$ sono stimati minimizzando la somma dei quadrati dei residui.
Risulta quindi necessario definire il parametro $K$, aggiungendo progressivamente una nuova base ad ogni passaggio successivo; procedendo in continuazione con tale operazione, tuttavia, si rischia di sovra-adattare il modello ai dati. Per individuare il valore ottimale di $K$ è possibile utilizzare la convalida incrociata generalizzata (*GCV*, da *Generalized Cross Validation*) (*Azzalini-Scarpa, 2009*).

Per costruire un modello di tipo *MARS* adeguato, si decide di utilizzare come variabili esplicative tutte le informazioni di cui si dispone; il numero di dimensioni risulterà elevato, ma la procedura iterativa del modello permetterà una selezione delle variabili da utilizzare e fornirà criteri adeguati per la scelta del numero di nodi necessari per ciascuna variabile.

| | 0/1 | size | RSS | GCV | | 0/1 | size | RSS | GCV |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 1 | 1 | 1697562.1 | 15.729 | 28 | 0 | 26 | 242515.5 | 2.251 |
| 2 | 1 | 2 | 412406.4 | 3.822 | 29 | 0 | 25 | 242746.6 | 2.253 |
| 3 | 1 | 3 | 328865.4 | 3.048 | 30 | 0 | 24 | 242983.3 | 2.255 |
| 4 | 1 | 4 | 301632.5 | 2.795 | 31 | 0 | 23 | 243236.4 | 2.257 |
| 5 | 1 | 5 | 286914.3 | 2.659 | 32 | 0 | 22 | 243585.2 | 2.261 |
| 6 | 1 | 6 | 275799.1 | 2.556 | 33 | 0 | 21 | 243962.8 | 2.264 |
| 7 | 1 | 7 | 267637.4 | 2.481 | 34 | 0 | 20 | 244354.7 | 2.267 |
| 8 | 1 | 8 | 262961.4 | 2.438 | 35 | 0 | 19 | 244813.3 | 2.271 |
| 9 | 1 | 9 | 258799.9 | 2.399 | 36 | 0 | 18 | 245412.0 | 2.277 |
| 10 | 1 | 10 | 255379.5 | 2.368 | 37 | 0 | 17 | 246014.2 | 2.282 |
| 11 | 1 | 11 | 253952.5 | 2.355 | 38 | 0 | 16 | 246666.8 | 2.288 |
| 12 | 1 | 12 | 251364.9 | 2.331 | 39 | 0 | 15 | 247598.6 | 2.297 |
| 13 | 1 | 13 | 250080.9 | 2.319 | 40 | 0 | 14 | 248638.4 | 2.306 |
| 14 | 1 | 14 | 248638.4 | 2.306 | 41 | 0 | 13 | 250080.9 | 2.319 |
| 15 | 1 | 15 | 247598.6 | 2.297 | 42 | 0 | 12 | 251364.9 | 2.331 |
| 16 | 1 | 16 | 246666.8 | 2.288 | 43 | 0 | 11 | 253952.5 | 2.355 |
| 17 | 1 | 17 | 246014.2 | 2.282 | 44 | 0 | 10 | 255379.5 | 2.368 |
| 18 | 1 | 18 | 245412.0 | 2.277 | 45 | 0 | 9 | 258799.9 | 2.399 |
| 19 | 1 | 19 | 244813.3 | 2.271 | 46 | 0 | 8 | 262961.4 | 2.438 |
| 20 | 1 | 20 | 244354.7 | 2.267 | 47 | 0 | 7 | 267637.4 | 2.481 |
| 21 | 1 | 21 | 243962.8 | 2.264 | 48 | 0 | 6 | 275799.1 | 2.556 |
| 22 | 1 | 22 | 243585.2 | 2.261 | 49 | 0 | 5 | 286914.3 | 2.659 |
| 23 | 1 | 23 | 243236.4 | 2.257 | 50 | 0 | 4 | 301632.5 | 2.795 |
| 24 | 1 | 24 | 242983.3 | 2.255 | 51 | 0 | 3 | 328865.4 | 3.048 |
| 25 | 1 | 25 | 242746.6 | 2.253 | 52 | 0 | 2 | 412406.4 | 3.821 |
| 26 | 1 | 26 | 242515.5 | 2.251 | 53 | 0 | 1 | 1697562.1 | 15.729 |
| 27 | 1 | 27 | 242177.1 | 2.248 | | | | | |

**TABELLA 3.12: MARS - Scelta del numero ottimale di nodi**

La Tabella 3.12 consente di verificare che il numero ottimale di nodi è pari a 27; il valore della somma dei quadrati dei residui, infatti, diminuisce fino a 27 e poi riprende ad aumentare.

I principali risultati del modello sono quindi i seguenti:

| | pred1 | knot1 | pred2 | knot2 | coefs | SE |
|---|---|---|---|---|---|---|
| 1 | 0 | NA | 0 | NA | 2.727 | 0.119 |
| 2 | 17 | NA | 0 | NA | 0.516 | 0.013 |
| 3 | 8 | NA | 0 | NA | -0.101 | 0.004 |
| 4 | 1 | NA | 0 | NA | 0.546 | 0.053 |
| 5 | 1 | NA | 8 | NA | 0.109 | 0.002 |
| 6 | 7 | NA | 0 | NA | -0.007 | 0.056 |
| 7 | 8 | NA | 17 | NA | -0.012 | 0.001 |
| 8 | 17 | 4.709 | 0 | NA | 0.543 | 0.013 |
| 9 | 5 | NA | 0 | NA | -0.236 | 0.028 |
| 10 | 1 | NA | 7 | NA | -1.032 | 0.023 |
| 11 | 8 | 6.000 | 0 | NA | -0.106 | 0.005 |
| 12 | 8 | 18.000 | 0 | NA | 0.179 | 0.007 |
| 13 | 2 | NA | 0 | NA | -0.103 | 0.006 |
| 14 | 1 | NA | 2 | NA | 0.081 | 0.003 |
| 15 | 1 | NA | 5 | NA | -0.210 | 0.010 |
| 16 | 2 | 10.000 | 0 | NA | 0.229 | 0.026 |
| 17 | 1 | NA | 2 | 10 | -0.231 | 0.014 |
| 18 | 5 | NA | 7 | NA | 0.299 | 0.018 |
| 19 | 2 | NA | 8 | NA | -0.002 | 0.0001 |
| 20 | 4 | NA | 0 | NA | -0.469 | 0.031 |
| 21 | 8 | 14.000 | 0 | NA | -0.094 | 0.009 |
| 22 | 1 | NA | 4 | NA | 0.197 | 0.015 |
| 23 | 12 | NA | 0 | NA | -0.047 | 0.006 |
| 24 | 7 | NA | 12 | NA | 0.039 | 0.004 |
| 25 | 2 | NA | 12 | NA | -0.005 | 0.0004 |
| 26 | 1 | NA | 12 | NA | 0.023 | 0.002 |
| 27 | 1 | NA | 8 | 14 | 0.008 | 0.003 |

Rsquared : 0.857

**TABELLA 3.13: MARS - Stima del modello**

La Tabella 3.13 riporta in che modo le basi entrano nel modello finale, e le variabili sono indicate con i numeri da 0 a 17 nel seguente modo:

0) l’intercetta;
1) la destinazione d’uso del punto di riconsegna;

2) la provincia in cui si trova il punto di riconsegna;
3) la regione in cui si trova il punto di riconsegna;
4) il numero di contratti in vigore;
5) il giorno della settimana in cui è stata effettuata la rilevazione;
6) il mese dell'anno in cui è stata effettuata la rilevazione;
7) il tipo di giorno (feriale o festivo) in cui è stata effettuata la rilevazione;
8) la temperatura media registrata il giorno della rilevazione;
9) la temperatura massima registrata il giorno della rilevazione;
10) il punto di rugiada registrato il giorno della rilevazione;
11) la visibilità registrata il giorno della rilevazione;
12) la velocità media del vento registrata il giorno della rilevazione;
13) la velocità massima del vento registrata il giorno della rilevazione;
14) la velocità massima delle raffiche di vento registrate il giorno della rilevazione;
15) la presenza di temporali il giorno della rilevazione;
16) la presenza di nebbia il giorno della rilevazione;
17) la capacità prenotata dagli Utenti nel giorno della rilevazione.

Il modello sembra evidenziare un leggero miglioramento delle stime e, di conseguenza, della capacità previsiva. L'$R^2$, infatti, sale all'85,7%% e, quindi, la quantità di varianza spiegata dal modello aumenta di circa 1 punto percentuale. E' possibile notare, inoltre, che nonostante le 17 variabili esplicative del modelle, solo alcune vi entrano come basi, oltre alla componente lineare: la destinazione d'uso del punto di riconsegna, la provincia in sui si trova il punto di riconsegna, il numero di contratti in vigore, il giorno della settimana, il tipo di giorno (feriale o festivo), la temperatura media del giorno in cui è stata effettuata la rilevazione, la velocità media del vento registrata il giorno della rilevazione e la capacità prenotata dagli Utenti.

Infine, occorre valutare la qualità delle previsioni, calcolando l'errore nell'insieme di verifica, che risulta pari a circa 266.280. L'errore di previsione

diminuisce leggermente, e quindi il modello *MARS* prevede un po' meglio del modello di regressione lineare.

### *3.3.3 I modelli non parametrici: Gam*

Oltre ai modelli di tipo *MARS*, un'altra famiglia di modelli non parametrici sono i modelli di tipo *GAM* (*Generalized Adaptive Model*), che consistono in un'estensione in ambito non parametrico dei *Modelli Lineari Generalizzati*. Essi permettono di creare una funzione ottimale per ciascuna variabile indipendente, secondo la seguente formula:

$$f(x_1, \dots, x_p) = \alpha + \sum_{j=1}^{p} f_j(x_J)$$

dove le $f_1, \dots, f_p$ sono funzioni in una variabile dall'andamento sufficientemente regolare, e $\alpha$ è una costante; per poter stimare le funzioni relative al modello additivo esiste una procedura iterativa, che prende il nome di *Algoritmo di backfitting*.
I *GAM* permettono di ottenere dei risultati più semplici da interpretare rispetto alle *MARS*: la funzione generata dai modelli additivi generalizzati per ogni variabile, infatti, non dipende da quelle ottenute per le altre (*Azzalini-Scarpa, 2009*).

Il modello *GAM* presenterà le stesse variabili indipendenti dei modelli precedenti, ad eccezione della visibilità registrata il giorno della rilevazione; per tutte le variabili continue sono state utilizzate le *spline di lisciamento* come stimatori non parametrici. Una volta eliminate le variabili non significative, il modello stimato risulta essere il seguente:

$$cons = a + f_1(gior_{sett}) + f_2(mese) + f_3(tipo_{giorno}) + f_4(contratti) + f_5(tipo_{punto}) + f_6(temp_{med}) + f_7(temp_{min}) + f_8(temp_{max}) + f_9(vento_{media}) + f_{10}(temporale) + f_{11}(nebbia) + f_{12}(capacita) + \varepsilon$$

Il modello permette di trovare la "migliore" funzione possibile, che riesca a cogliere l'andamento delle variabili esplicative rispetto alla variabile risposta. I risultati principali sono i seguenti:

| | Df | Npar | Df | Npar | F | Pr(F) | |
|---|---|---|---|---|---|---|---|
| (Intercept) | 1 | | | | | | |
| tipo_punto | 2 | | | | | | |
| s(contratti) | 1 | | 3 | | 128.01 | < 2.2e-16 | *** |
| gior_sett | 6 | | | | | | |
| mese | 11 | | | | | | |
| tipo_giorno | 1 | | | | | | |
| s(temp_media) | 1 | | 3 | | 37.18 | < 2.2e-16 | *** |
| s(temp_min) | 1 | | 3 | | 59.71 | < 2.2e-16 | *** |
| s(temp_max) | 1 | | 3 | | 52.85 | < 2.2e-16 | *** |
| s(vento_media) | 1 | | 3 | | 10.62 | 5.625e-07 | *** |
| temporale | 1 | | | | | | |
| nebbia | 1 | | | | | | |
| s(capacita3) | 1 | | 3 | | 533.88 | < 2.2e-16 | *** |

**TABELLA 3.14: Stima del modello GAM**

Occorre, infine, valutare la qualità delle previsioni, calcolando l'errore nell'insieme di verifica: esso risulta pari a circa 267.814, più elevato rispetto all'errore degli altri modelli precedentemente stimati.

### *3.3.4 I modelli non parametrici: Projection Pursuit*

I modelli precedenti hanno il limite che, all'aumentare del numero di variabili esplicative (e di conseguenza del numero di dimensioni del modello), faticano a cogliere le relazioni tra di esse e la variabile risposta.
Il metodo del *Project Pursuit Regression* costruisce un certo numero di curve di regressione, approssimate dalla generica funzione $f$, facendo ruotare l'asse delle ascisse in un numero $M$ di direzioni, dividendo così l'area del modello in diversi "settori" e cogliendo, in questo modo, le relazioni unidimensionali tra le variabili al loro interno, risparmiando di conseguenza un elevato numero di gradi di libertà.

La formula generica di tale metodo è la seguente:

$$Y = a_0 + \sum_{j=1}^{M} f_j(a_j^T X) + \varepsilon$$

dove le M rappresentano le diverse direzioni del modello, che devono essere indicate dall'analista che costruisce il modello.
Il maggior problema di questa tipologia di modelli consiste nella difficoltà di interpretare i risultati ottenuti se si utilizza un numero di direzioni $M > 1$; le relazioni unidimensionali che si potrebbero cogliere, inoltre, sono, in linea teorica, un numero molto elevato e si deve prestare attenzione che quelle identificate dal modello non generino un problema di sovra-adattamento ai dati.
Il pregio più evidente della *Project Pursuit Regression* consiste nel fatto che si riesce a contrastare il problema della *maledizione della dimensionalità*, dal momento che le relazioni colte sono sempre di tipo unidimensionale; questa caratteristica rende questa famiglia di modelli particolarmente interessanti per l'obiettivo dell'analisi.
Il numero di direzioni M, come visto in precedenza, deve essere impostato dall'analista. Per ottenere questo valore sono stati costruiti diversi modelli con

numeri differenti di direzioni, ed è stato scelto il numero che minimizza l'errore nell'insieme di verifica.

| *M* | *errore* |
|---|---|
| 1 | 268.153 |
| 2 | 266971 |
| 3 | 263.945 |
| 4 | 260.975 |
| 5 | 260.665 |
| 6 | 260.749 |
| 7 | 260.198 |
| 8 | 260.733 |
| 9 | 258.264 |
| **10** | **257.025** |

**TABELLA 3.15: Projection Pursuit – Scelta di M**

Il numero ottimale di direzioni risulta pari a M=10; le variabili inserite nel modello sono le stesse dei modelli precedenti, e l'errore nell'insieme di verifica risulta pari a circa 257.025, il valore più basso finora registrato. Il modello *Projection Pursuit* risulta quindi il miglior modello tra quelli stimati fino ad adesso.

### *3.3.5 I modelli non parametrici: Reti Neurali*

Una rete neurale è sostanzialmente uno schema di regressione a due stadi, generalmente di tipo non lineare, in cui sono messe in relazione *p* variabili esplicative (o di *input*) con *q* variabili risposta (o di *output*), attraverso uno strato *r* di *variabili latenti* (quindi non osservate), che si frappongono tra i due precedenti gruppi (nel senso che le variabili esplicative influenzano le variabili latenti e queste ultime influenzano le variabili risposta) (*Azzalini-Scarpa, 2009*).

Il numero di unità *r* nello strato latente e il *weight decay*, come visto in precedenza, devono essere impostati dall'analista; per ottenere i valori ottimali sono state stimate più reti neurali con valori differenti di unità nello strato latente e di *weight decay*, e sono stati poi scelti quelli la cui rete corrispondente minimizza l'errore nell'insieme di verifica.

| *r* | *weight decay* | *errore* |
|---|---|---|
| 3 | 0,1 | 264.927 |
| 3 | 0,2 | 262.747 |
| 3 | 0,3 | 264.850 |
| 3 | 0,4 | 263.281 |
| 3 | 0,5 | 263.530 |
| 3 | 0,6 | 266.634 |
| 3 | 0,7 | 266.348 |
| 3 | 0,8 | 266.186 |
| 3 | 0,9 | 265.168 |
| 3 | 1,0 | 262.312 |
| 4 | 0,1 | 264.661 |
| 4 | 0,2 | 263.614 |
| 4 | 0,3 | 265.518 |
| 4 | 0,4 | 263.646 |
| 4 | 0,5 | 262.608 |

| *r* | *weight decay* | *errore* |
|---|---|---|
| 4 | 0,6 | 262.826 |
| 4 | 0,7 | 262.367 |
| 4 | 0,8 | 261.887 |
| 4 | 0,9 | 266.832 |
| 4 | 1,0 | 261.076 |
| 5 | 0,1 | 263.491 |
| 5 | 0,2 | 263.204 |
| 5 | 0,3 | 261.296 |
| **5** | **0,4** | **259.703** |
| 5 | 0,5 | 263.260 |
| 5 | 0,6 | 260.435 |
| 5 | 0,7 | 261.238 |
| 5 | 0,8 | 259.939 |
| 5 | 0,9 | 260.420 |
| 5 | 1,0 | 262.736 |

**TABELLA 3.16: Reti Neurali – Scelta di *r* e del *weight decay***

Il numero ottimale di unità nello strato latente è pari a *r=5*, con un corrispondente *weight decay=0,4*, e le variabili inserite nel modello saranno le stesse dei modelli precedenti; la struttura della rete, tuttavia, non può essere

riportata in forma grafica, in quanto l'elevato numero di variabili di input ne rende difficile la visualizzazione.

Infine, si valuta la qualità delle previsioni, calcolando l'errore nell'insieme di verifica: esso risulta pari a circa 259.703, il secondo valore più basso finora ottenuto.

### *3.3.6 Il confronto fra i modelli*

Dopo aver stimato i modelli, occorre confrontare e valutare i risultati ottenuti; il termine di paragone sarà l'errore nell'insieme di verifica, che tiene conto, non solo dell'efficienza del modello, ma anche del problema del sovra-adattamento del modello ai dati.

| | *errore* |
|---|---|
| Modello di regressione lineare | 266.683 |
| *MARS* | 266.280 |
| *GAM* | 267.814 |
| ***Project Pursuit Regression*** | **257.025** |
| Rete Neurale | 259.703 |

**TABELLA 3.17: Confronto tra i modelli di regressione**

Analizzando la Tabella 3.17, occorre innanzitutto evidenziare come i modelli non parametrici, ad eccezione del *GAM*, migliorano i risultati ottenuti dal Modello di regressione lineare: soprattutto con il modello *Projection Pursuit* si passa da 266.683 a 257.025.

I modelli migliori, con i dati a disposizione, sono il *Project Pursuit Regression* e la Rete Neurale; quando tali modelli vengono utilizzati per effettuare delle

previsioni, tuttavia, occorre tenere conto dei pro e dei contro espressi in precedenza, come ad esempio le difficoltà di stima e di interpretazione dei risultati ottenuti.

Il modello *Project Pursuit* sarà quindi utilizzato per prevedere il consumo giornaliero di gas per le osservazioni dell'insieme di verifica che in precedenza l'analisi di classificazione aveva classificato nella classe 1 (consumo giornaliero di gas maggiore di 0).

## 3.4 LA PREVISIONE DEL CONSUMO DI GAS

Dopo aver identificato il modello migliore, occorre affrontare il problema della previsione del consumo di gas. Come spiegato in precedenza, questa operazione si compone di due fasi:

- nella prima fase, le osservazioni vengono classificate in due categorie tramite una procedura di tipo *bagging*: quelle che hanno un consumo di gas previsto uguale a 0 (classe 0), e quelle che hanno un consumo di gas previsto maggiore di 0 (classe 1). Per le osservazioni della classe 0, l'operazione di previsione si conclude, in quanto viene previsto un consumo di gas uguale a 0, mentre per le osservazioni della classe 1, si passa alla fase successiva;

- nella seconda fase, per le osservazioni della classe 1, viene previsto il valore del consumo giornaliero di gas tramite un modello *Projection Pursuit*. La probabilità che un punto di riconsegna consumi un certo quantitativo di gas, quindi, risulta essere condizionata al fatto che tale consumo sia maggiore di 0. Poiché per la costruzione del modello *Projection Pursuit* è stata utilizzata una trasformazione di *Box-Cox* ($\lambda = 0{,}06$) del consumo giornaliero di gas, per avere una previsione in metri cubi occorre applicare alle previsioni del modello la seguente trasformata inversa:

$$y = (0{,}06 y_{tras} + 1)^{1/0{,}06}$$

Una volta ultimate queste due fasi, si ottengono le previsioni del consumo giornaliero di gas; un esempio di previsioni per il Punto di Riconsegna *30009501*, situato nel comune di Alessandria, per il periodo che va dal 10 Aprile 2011 al 30 Aprile 2011, è contenuto nella Tabella 3.18:

| *Data* | *Classificazione* | *Previsione* |
|---|---|---|
| 10/04/2011 | 0 | 0 |
| 11/04/2011 | 1 | 2.282 |
| 12/04/2011 | 1 | 2.239 |
| 13/04/2011 | 1 | 1.974 |
| 14/04/2011 | 1 | 2.324 |
| 15/04/2011 | 1 | 2.442 |
| 16/04/2011 | 0 | 0 |
| 17/04/2011 | 0 | 0 |
| 18/04/2011 | 1 | 2.591 |
| 19/04/2011 | 1 | 2.506 |
| 20/04/2011 | 1 | 2.450 |
| 21/04/2011 | 1 | 2.418 |
| 22/04/2011 | 0 | 0 |
| 23/04/2011 | 0 | 0 |
| 24/04/2011 | 0 | 0 |
| 25/04/2011 | 0 | 0 |
| 26/04/2011 | 1 | 2.329 |
| 27/04/2011 | 1 | 2.191 |
| 28/04/2011 | 1 | 2.331 |
| 29/04/2011 | 1 | 2.256 |
| 30/04/2011 | 1 | 2.141 |

**TABELLA 3.18: Previsioni per il Punto di Riconsegna di Alessandria**

## 3.5 CONCLUSIONI

In questo capitolo sono stati individuati i modelli migliori, con le variabili a disposizione, per la classificazione e la previsione del consumo giornaliero di gas per quei punti per cui non si hanno a disposizione informazioni sullo storico dei consumi e per cui, quindi, non si possono utilizzare i modelli delle serie storiche.

L'errore totale dei modelli individuati, tuttavia, risulta sempre abbastanza elevato, e questo è dovuto:

- alla correlazione seriale dei consumi, che i modelli del *Data Mining* considerati non tengono in considerazione, e che è stato dimostrato essere presente e rilevante;

- alla mancanza di ulteriori informazioni sulle aziende e sulle industrie a cui viene distribuito il gas (ad esempio dimensioni e settore in cui operano); è possibile verificare, infatti, che le tipologie di punti su cui si commette un errore di previsione più elevato sono i punti ad uso civile e industriale e i punti esclusivamente ad uso industriale. E' facile intuire, infatti, che vi sono processi industriali che richiedono quantità di gas superiori rispetto ad altri, e che, allo stesso modo, industrie più grandi necessitano di un quantitativo di gas superiore rispetto ad industrie di dimensioni più ridotte.

Nel prossimo Capitolo si cercherà di risolvere il problema della correlazione seriale, e si prenderanno in considerazione i punti su cui si dispone di informazioni relative allo storico dei consumi, utilizzando modelli di tipo *ARIMA* o *VAR* per la stima e per le previsioni.

# CAPITOLO 4: L'ANALISI DELLE SERIE STORICHE

L'analisi del consumo giornaliero di gas attraverso l'utilizzo di modelli per le serie storiche è fondamentale, in quanto in questo modo si tiene conto della correlazione seriale dei consumi, cosa che non avveniva con i modelli precedenti. Per applicare questa classe di modelli, tuttavia, è necessario avere a disposizione informazioni sullo storico dei consumi e, di conseguenza, si possono ricavare modelli e previsioni solamente per i punti di riconsegna già in funzione, e presso i quali vi è già un flusso giornaliero di gas del quale si hanno informazioni.

## 4.1 L'ANALISI UNIVARIATA

Il primo tipo di analisi da effettuare è quella univariata su un singolo punto di riconsegna: questa tipologia di analisi consiste nell'adattare allo storico dei consumi del punto un modello stocastico, che descriva il processo generatore dei dati.

Un modello possibile è del tipo:

$$Y_t = f(t) + u_t$$

In un'espressione di questo tipo si assume che la serie osservata $Y_t$ sia il risultato della composizione di una sequenza deterministica $f(t)$, che costituisce la parte sistematica della serie  e di una sequenza di variabili casuali $u_t$ , che rappresenta la parte stocastica della serie ed obbedisce ad una determinata legge di probabilità. Le due sequenze non sono individualmente osservabili, ma vanno determinate sulla base del campione.

Il modello stocastico espresso può essere interpretato seguendo un approccio classico o un approccio moderno.

L'approccio classico (o tradizionale) suppone che esista una "legge di evoluzione temporale" del fenomeno, rappresentata da $f(t)$. La componente casuale $u_t$ viene invece assunta a rappresentare l'insieme delle circostanze, ciascuna di entità trascurabile, che non si vogliono o non si possono considerare esplicitamente in $Y_t$ . I residui di $Y_t$ , non spiegati da $f(t)$, vengono pertanto imputati al caso e trattati come errori accidentali. Da un punto di vista statistico ciò equivale ad ipotizzare che la componente stocastica del modello sia generata da un processo white noise, ossia da una successione di variabili casuali indipendenti, identicamente distribuite, di media nulla e varianza costante. Una successione di variabili così definita viene detta processo stocastico a componenti incorrelate (*Di Fonzo-Lisi, 2007*).
In sintesi, nell'approccio classico l'attenzione viene concentrata su $f(t)$, essendo $u_t$ considerato un processo a componenti incorrelate e dunque trascurabile.

Nell'approccio moderno si ipotizza invece che $f(t)$ manchi o sia già stata "eliminata" (mediante stima o altri metodi). L'attenzione viene posta quindi sulla componente stocastica $u_t$, che si ipotizza essere un processo a componenti correlate del tipo:

$$u_t = g(Y_{t-1}, Y_{t-2}, \dots\dots\dots, \varepsilon_{t-1}, \varepsilon_{t-2}, \dots\dots\dots) + \varepsilon_t$$

Si cerca quindi di estrarre qualche tipo di informazione da questo tipo di processo.
L'obiettivo non è più quindi di arrivare ad una stima delle componenti della serie, ma piuttosto quello di individuare un modello probabilistico che descriva l'evoluzione del fenomeno in esame, modello che può essere utilizzato sia a fini descrittivi che previsivi.

I principali processi stocastici sono i processi a media mobile *MA(q)*, i processi auto-regressivi *AR(p)*, i processi auto-regressivi a media mobile *ARMA(p,q)*, i processi auto-regressivi integrati a media mobile *ARIMA (p,d,q)* e i processi stagionali *SARIMA (p,d,q)*x$(P,D,Q)_s$ .

L'idea è di utilizzare, in questa analisi, un approccio moderno per l'adattamento dei modelli alle serie storiche sui punti di riconsegna. Poiché l'andamento del consumo giornaliero di gas è prevalentemente stagionale, il modello che si utilizzerà a fini descrittivi e previsivi è il processo stagionale di tipo *SARIMA (p,d,q)*x$(P,D,Q)_s$.
Nel caso di processi stagionali, infatti, la stagionalità risulta spesso essere una componente stocastica molto importante e correlata con le componenti non stagionali; l'idea che sta alla base dei processi stagionali *SARIMA* è che il processo deve poter descrivere due tipi di relazioni all'interno della serie osservata: la correlazione tra valori consecutivi, che è quella modellata dagli usuali modelli *ARIMA*, e la correlazione tra osservazioni che distano tra di loro un multiplo del periodo.

La procedura utilizzata è quella di Box e Jenkins (*Di Fonzo-Lisi, 2007*); consiste in un metodo statistico che consente di approssimare adeguatamente il processo generatore di una serie storica di dati. Tutte le operazioni necessarie per costruire il modello devono quindi fare riferimento solo alla serie storica osservata e si possono riassumere in 3 gruppi:

1) <u>IDENTIFICAZIONE</u>: il primo passo consiste nella specificazione dell'ordine del modello, ovvero nell'identificazione dei valori che lo contraddistinguono. I principali strumenti da usare sono la funzione di autocorrelazione e la funzione di autocorrelazione parziale, entrambe stimate sul campione. L'idea di base è quella di riconoscere nella struttura della funzione di autocorrelazione empirica la struttura di una funzione di autocorrelazione teorica.

2) STIMA DEI PARAMETRI: una volta che sono stati fissati tali valori, cioè dopo che il modello è stato identificato, si può passare alla fase di stima dei parametri che lo caratterizzano. Il metodo di base per compiere questa operazione è quello della massima verosimiglianza (esatta o condizionata); tale metodo richiede la conoscenza della distribuzione del termine di errore e fornisce stimatori con ottime proprietà statistiche.

3) CONTROLLO DIAGNOSTICO: una volta che un prefissato modello è stato stimato, il passo finale della procedura di costruzione è quello di controllarne l'adeguatezza facendo uso di opportune analisi e test diagnostici. Alla base di tali test c'è le considerazione che se il modello è stato correttamente identificato e stimato, allora sui residui devono potersi riscontrare determinate ipotesi fatte a priori. Tra questa la più importante è l'incorrelazione seriale. Alcune analisi diagnostiche sono le seguenti: analisi grafiche (la serie dei residui non dovrebbe mostrare alcun tipo di regolarità né valori particolarmente diversi gli uni dagli altri; utili indicazioni si possono trarre dal grafico dei residui o dal diagramma di dispersione dei punti, che dovrebbe dar luogo a una nuvola di punti senza alcuna struttura), autocorrelazione dei residui (la serie dei residui può essere trattata come una serie storica a sé stante per la quale è possibile calcolare le funzioni di autocorrelazione totale e parziale; tali funzioni di autocorrelazione devono poter essere ricondotte a quelle di un *White Noise*, e la statistica finale di *Ljung-Box* deve permetterci di accettare, a un livello di significatività del 5%, l'ipotesi nulla di incorrelazione seriale dei residui), test di normalità dei residui (è utile verificare tale ipotesi perché, nel caso di gaussianità, l'incorrelazione dei residui implica anche la loro indipendenza; ciò significa che il modello lineare è in grado di spiegare l'intera struttura di dipendenza seriale della serie d'esame).

La procedura di Box e Jenkins è quindi una procedura iterativa, in quanto, se si ottiene un modello che non risulta particolarmente soddisfacente, si è obbligati a

ripercorrere le stesse operazioni in modo iterativo, fino ad ottenere un modello che ben si adatta a spiegare la variabilità dei dati.

In questa analisi si è deciso di selezionare un punto di riconsegna per ciascuna tipologia di destinazione d'uso (civile, civile e industriale, industriale), e nelle analisi che seguiranno, si adatterà ai punti selezionati un modello *SARIMA* attraverso la procedura di Box e Jenkins.

### *4.1.1 ARGELATO (destinazione d'uso civile)*

La serie storica del consumo di gas presso il punto di riconsegna di Argelato è composta da 577 osservazioni giornaliere dall'01/10/2009 al 30/04/2011; per stimare il modello, tuttavia, si decide di eliminare le osservazioni relative agli ultimi 20 giorni, sulle quali poi si andranno a effettuare poi le previsioni. L'insieme di stima risulta così composto da 557 osservazioni giornaliere dall'01/10/2009 al 10/04/2011, mentre le previsioni saranno effettuate sul periodo che va dall'11/04/2011 al 30/04/2011.

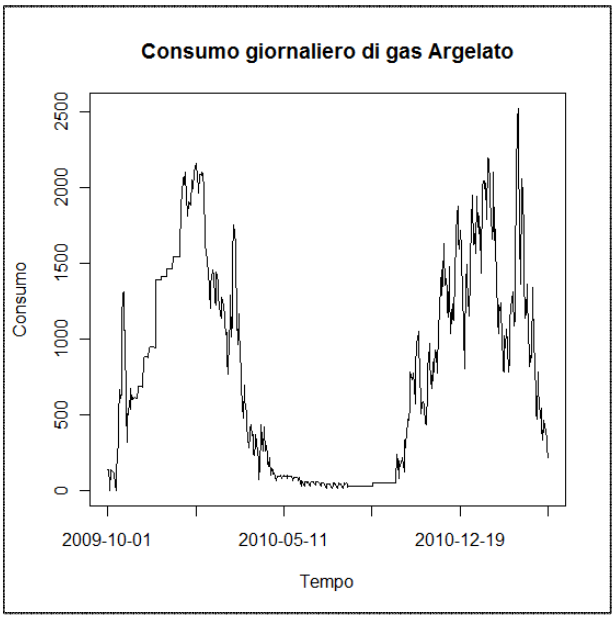



**GRAFICO 4.1: Serie storica consumo di gas Argelato**

La serie storica giornaliera del consumo di gas ad Argelato sembra una serie non stazionaria né in media, e questo è confermato dal fatto che la funzione di autocorrelazione che decresce molto lentamente, né in varianza.

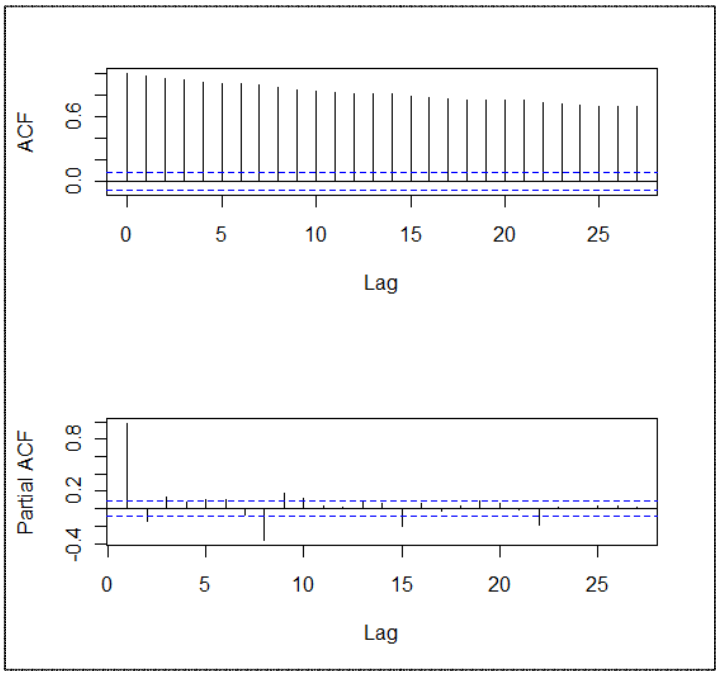



**GRAFICO 4.2: Correlogramma serie storica consumo di gas Argelato**

Per rendere la serie stazionaria in varianza si effettua un'opportuna trasformazione dei dati; vista la presenza di alcune osservazioni pari a 0 la trasformata ideale è, come visto nel capitolo precedente, il seno iperbolico inverso del consumo. Per rendere la serie stazionaria in media ed evitare problemi di sovra-differenziazione, occorre prima di tutto valutare se il trend evidenziato dal grafico è di tipo deterministico o stocastico, attraverso un opportuno test *ADF* (*Augmented Dickey-Fuller Test*) che verifica la presenza di radici unitarie per l'equazione caratteristica del processo.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.247      0.089   2.776    0.006 **
z.lag.1       -0.037      0.013  -2.835    0.005 **
z.diff.lag    -0.292      0.041  -7.165 2.51e-12 ***

Residual standard error: 0.473 on 552 degrees of freedom
Multiple R-squared: 0.109
F-statistic:  33.8 on 2 and 552 DF,  p-value: 1.421e-14

Value of test-statistic is: -2.835 4.020

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.43 -2.86 -2.57
phi1  6.43  4.59  3.78
```

**TABELLA 4.1: Test *ADF* serie storica consumo di gas Argelato**

Il test *ADF* conferma che, a un livello di significatività fissato del 5% la serie risulta non stazionaria e che quindi il trend è di tipo stocastico; per rendere la serie stazionaria in media è quindi possibile procedere con una differenziazione di lag pari a 1.

Si valuta ora il correlogramma della serie resa stazionaria in media e in varianza.

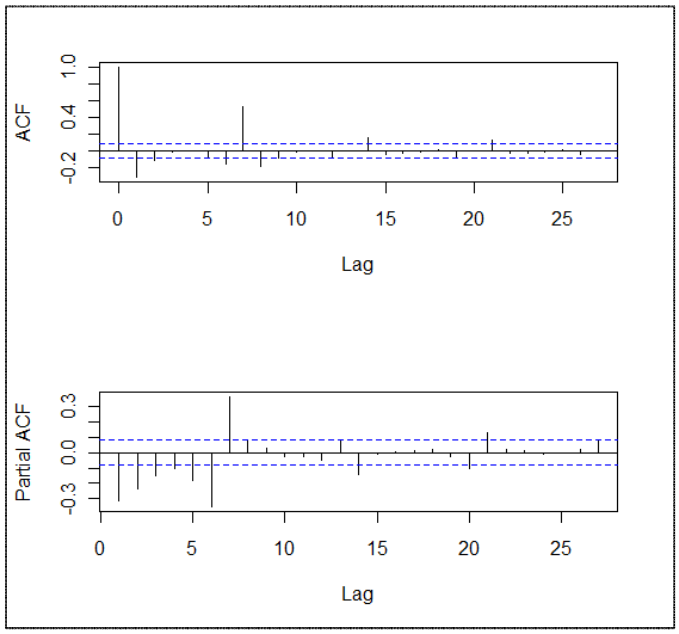



**GRAFICO 4.3: Correlogramma serie storica stazionaria consumo di gas Argelato**

La serie sembra evidenziare una componente stagionale al ritardo 7 e ai suoi multipli; è possibile provare ad eliminarla tramite una differenza di *lag* pari a 7, ottenendo il seguente correlogramma:

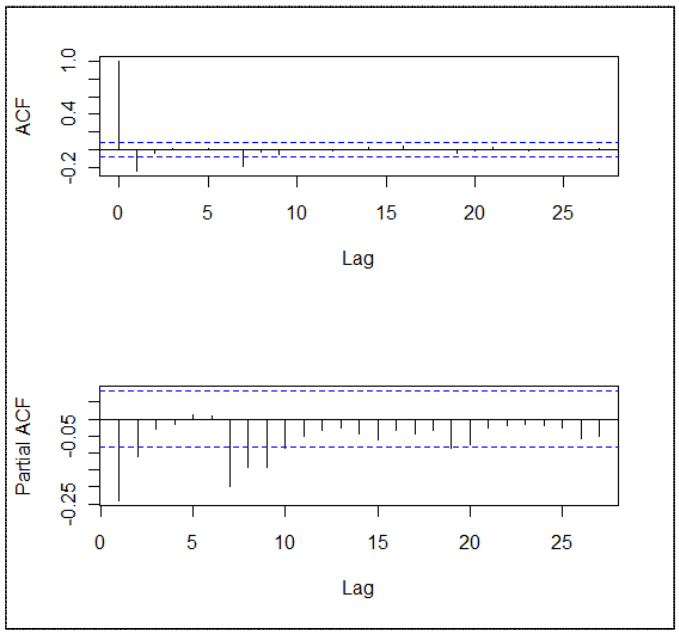



**GRAFICO 4.4: Correlogramma serie storica stazionaria e destagionalizzata consumo di gas Argelato**

Osservando il correlogramma della serie ottenuta è possibile individuare la presenza di una componente auto-regressiva di ordine 1, una componente a media mobile di ordine 2 e una componente auto-regressiva stagionale di ordine 1. Il modello che si prova a stimare è quindi un *SARIMA (1,1,2)*x*(1,1,0)*$_7$ , senza costante; le stime che si ottengono sono le seguenti:

| | coefficiente | errore std. | z | p-value | |
|---|---|---|---|---|---|
| phi_1 | 0.814 | 0.044 | 19.48 | 1.71e-084 | *** |
| Phi_1 | -0.211 | 0.043 | -4.870 | 1.12e-06 | *** |
| theta_1 | -1.187 | 0.062 | -19.13 | 1.46e-081 | *** |
| theta_2 | 0.203 | 0.058 | 3.519 | 0.0004 | *** |

| | | | |
|---|---|---|---|
| Media var. dipendente | -0.001 | SQM var. dipendente | 0.381 |
| Media innovazioni | -0.015 | SQM innovazioni | 0.345 |
| Log-verosimiglianza | -195.851 | Criterio di Akaike | 401.701 |
| Criterio di Schwarz | 423.242 | Hannan-Quinn | 410.119 |

Note: SQM = scarto quadratico medio; E.S. = errore standard

**TABELLA 4.2: Modello consumo di gas Argelato - *SARIMA (1,1,2)*x*(1,1,0)*$_7$**

Le stime dei parametri risultano tutte significativamente diverse da 0 a un livello di significatività fissato pari al 5%.

Il modello stimato per la serie storica del consumo giornaliero di gas ad Argelato è quindi il seguente:

$$(\mathbf{1-0,81}\boldsymbol{B})(\mathbf{1+0,21}\boldsymbol{B}^{\mathbf{7}})(\mathbf{1}-\boldsymbol{B})(\mathbf{1}-\boldsymbol{B}^{\mathbf{7}})\boldsymbol{cons}_{\boldsymbol{ihs;t}} = (\mathbf{1+1,19}\boldsymbol{B}-\mathbf{0,20}\boldsymbol{B}^{\mathbf{2}})\boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

Occorre ora valutare la bontà del modello stimato; per fare ciò bisogna verificare la qualità dei residui del modello, e verificare se il loro comportamento è simile a quello di un *White Noise*.

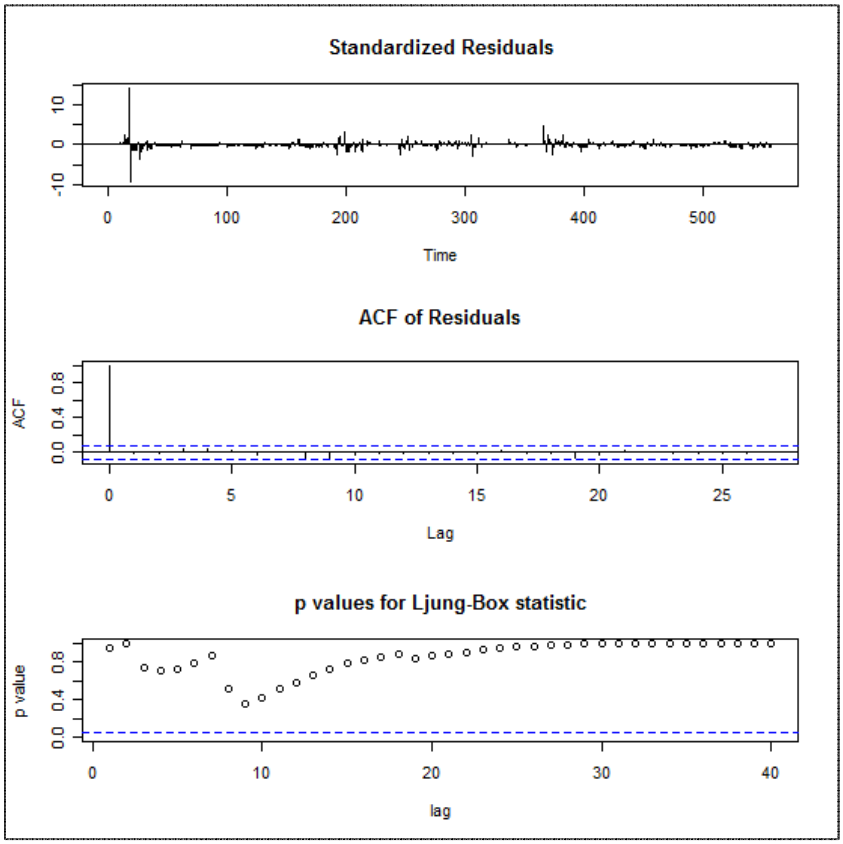



**GRAFICO 4.5: Residui modello consumo di gas Argelato - *SARIMA (1,1,2)x(1,1,0)*$_7$**

Osservando la funzione di autocorrelazione dei residui nel Grafico 4.5 (*ACF of Residuals*), è possibile affermare che i residui risultano tra loro incorrelati, in quanto tale funzione è riconducibile alla funzione di autocorrelazione teorica di un *White Noise*. Anche la statistica di *Ljung-Box* conferma questa affermazione, in quanto è possibile accettare l'ipotesi nulla di incorrelazione seriale dei residui con un *p-value* pressoché pari a 1.

Possiamo quindi affermare che il modello stimato è, tutto sommato, un buon modello ed è quindi possibile utilizzarlo per ottenere le previsioni giornaliere per il periodo 11/04/2011-30/04/2011. Si ottengono, in questo modo, le previsioni del seno iperbolico inverso del consumo giornaliero di gas, che dovranno poi essere riconvertite in metri cubi attraverso la trasformata seno iperbolico. Occorre comunque tenere in considerazione che le previsioni possono essere considerate attendibili solo per un numero ridotto di passi in avanti a causa del continuo aumento dell'errore standard di previsione.

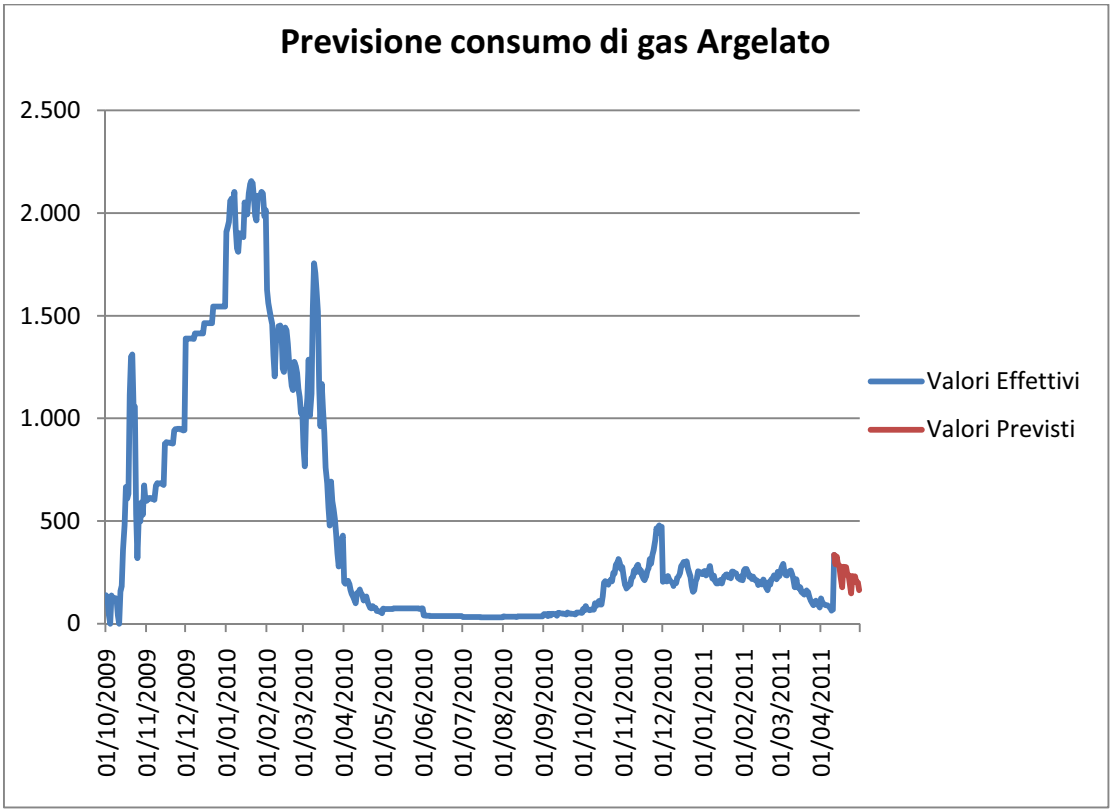



**GRAFICO 4.6: Previsioni consumo di gas Argelato - *SARIMA (1,1,2)x(1,1,0)*$_7$**

| DATA | CONSUMO |
|---|---|
| 11/04/2011 | 335 |
| 12/04/2011 | 290 |
| 13/04/2011 | 328 |
| 14/04/2011 | 286 |
| 15/04/2011 | 284 |
| 16/04/2011 | 231 |
| 17/04/2011 | 178 |
| 18/04/2011 | 277 |
| 19/04/2011 | 240 |
| 20/04/2011 | 275 |

| DATA | CONSUMO |
|---|---|
| 21/04/2011 | 241 |
| 22/04/2011 | 237 |
| 23/04/2011 | 194 |
| 24/04/2011 | 149 |
| 25/04/2011 | 233 |
| 26/04/2011 | 202 |
| 27/04/2011 | 231 |
| 28/04/2011 | 203 |
| 29/04/2011 | 200 |
| 30/04/2011 | 164 |

**TABELLA 4.3: Previsioni consumo di gas Argelato - *SARIMA (1,1,2)x(1,1,0)*$_7$**

### *4.1.2 TOCCO DA CASAURIA (destinazione d'uso civile e industriale)*

La serie storica del consumo di gas presso il punto di riconsegna di Tocco da Casauria è composta da 577 osservazioni giornaliere dall'01/10/2009 al 30/04/2011; per stimare il modello, tuttavia, si decide di eliminare le osservazioni relative agli ultimi 20 giorni, sulle quali poi si andranno a effettuare le previsioni. L'insieme di stima risulta così composto da 557 osservazioni giornaliere dall'01/10/2009 al 10/04/2011, mentre le previsioni saranno effettuate sul periodo che va dall'11/04/2011 al 30/04/2011.

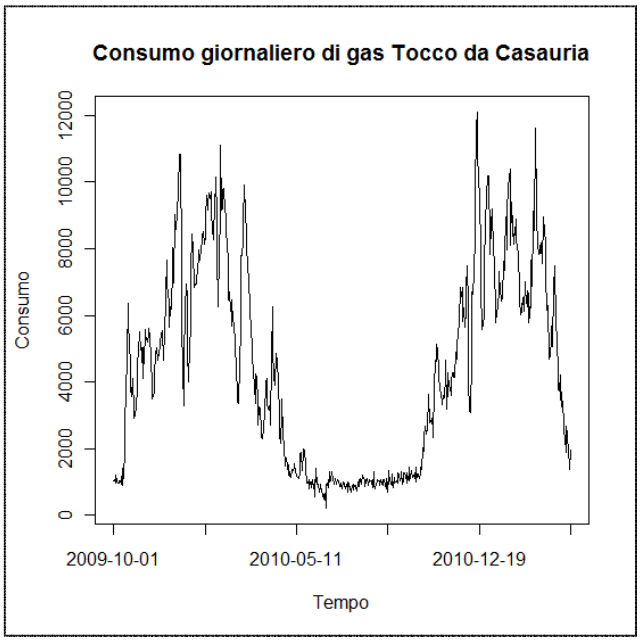



**GRAFICO 4.7: Serie storica consumo di gas Tocco da Casauria**

La serie storica giornaliera del consumo di gas a Tocco di Casauria sembra una serie non stazionaria né in media, e questo è confermato dal fatto che la funzione di autocorrelazione che decresce molto lentamente, né in varianza.

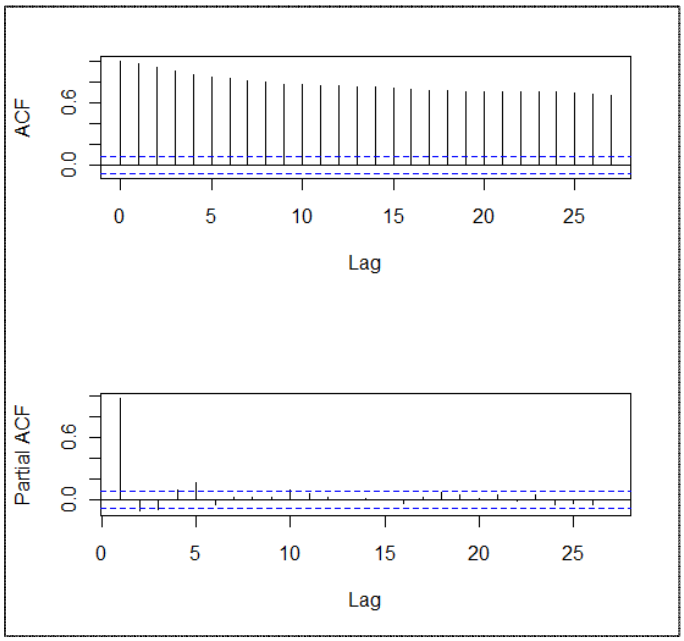



**GRAFICO 4.8: Correlogramma serie storica consumo di gas Tocco da Casauria**

Per rendere la serie stazionaria in varianza si effettua un'opportuna trasformazione dei dati, utilizzando la trasformata seno iperbolico inverso del consumo giornaliero di gas. Per rendere la serie stazionaria in media ed evitare problemi di sovra differenziazione, occorre prima di tutto valutare se il trend evidenziato dal grafico è di tipo deterministico o stocastico, attraverso un opportuno test *ADF* (*Augmented Dickey-Fuller Test*), che verifichi la presenza di eventuali radici unitarie per l'equazione caratteristica del processo.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.165      0.079   2.083   0.0377 *
z.lag.1        -0.019      0.009  -2.078   0.0382 *
z.diff.lag     -0.289      0.041  -7.075 4.57e-12 ***

Residual standard error: 0.183 on 552 degrees of freedom
Multiple R-squared: 0.095
F-statistic: 29.09 on 2 and 552 DF,  p-value: 9.75e-13

Value of test-statistic is: -2.078 2.171

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.43 -2.86 -2.57
phi1  6.43  4.59  3.78
```

**TABELLA 4.4: Test *ADF* serie storica consumo di gas Tocco da Casauria**

Il test *ADF* conferma che, a un livello di significatività fissato del 5% la serie risulta non stazionaria e che quindi il trend è di tipo stocastico; per rendere la serie stazionaria in media è quindi possibile procedere con una differenziazione di ordine 1.

Si valuta ora il correlogramma della serie resa stazionaria in media e in varianza.

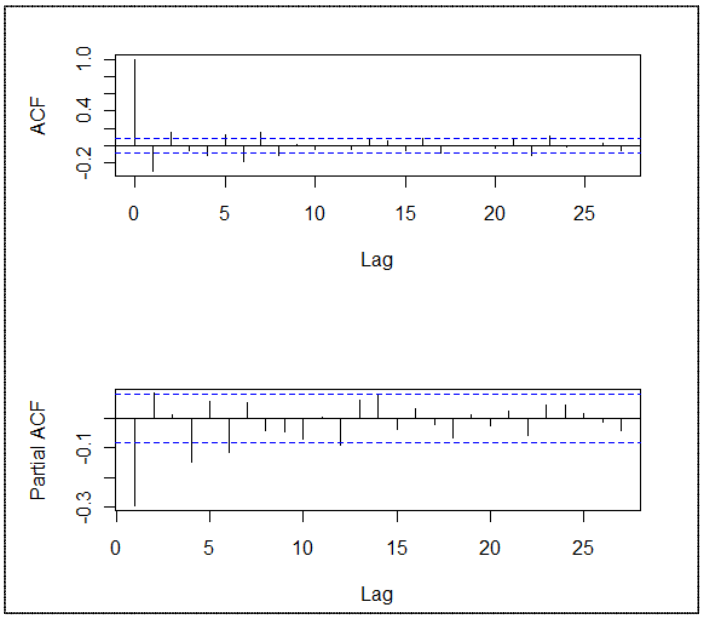



**GRAFICO 4.9: Correlogramma serie storica stazionaria consumo di gas Tocco**

La serie non sembra presentare una componente stagionale ben definita; osservando il correlogramma è possibile individuare la presenza di una componente auto-regressiva di ordine 3 e di una componente a media mobile di ordine 1. Il modello che si prova a stimare è quindi un *ARIMA (3,1,1)*, senza costante; le stime che si ottengono sono le seguenti:

| | coefficiente | errore std. | z | p-value | |
|---|---|---|---|---|---|
| phi_1 | -1.156 | 0.0682 | -16.96 | 1.75e-064 | *** |
| phi_2 | -0.145 | 0.067 | -2.144 | 0.032 | ** |
| phi_3 | 0.145 | 0.045 | 3.260 | 0.001 | *** |
| theta_1 | 0.902 | 0.0572 | 15.78 | 4.59e-056 | *** |

| | | | |
|---|---|---|---|
| Media var. dipendente | 0.001 | SQM var. dipendente | 0.192 |
| Media innovazioni | 0.001 | SQM innovazioni | 0.179 |
| Log-verosimiglianza | 165.672 | Criterio di Akaike | -321.344 |
| Criterio di Schwarz | -299.740 | Hannan-Quinn | -312.906 |

Note: SQM = scarto quadratico medio; E.S. = errore standard

**TABELLA 4.5: Modello consumo di gas Tocco da Casauria - *ARIMA (3,1,1)***

Le stime dei parametri risultano tutte significativamente diverse da 0 a un livello di significatività fissato pari al 5%.

Il modello stimato per la serie storica del consumo giornaliero di gas a Tocco di Casauria è quindi il seguente:

$$(\mathbf{1} + \mathbf{1,16}\boldsymbol{B} + \mathbf{0,14}\boldsymbol{B^2} - \mathbf{0,15}\boldsymbol{B^3})(\mathbf{1} - \boldsymbol{B})\boldsymbol{cons}_{\boldsymbol{ihs;t}} = (\mathbf{1} - \mathbf{0,90}\boldsymbol{B})\boldsymbol{\varepsilon_t}$$

Occorre ora valutare la bontà del modello stimato; per fare ciò bisogna verificare la qualità dei residui ottenuti, e verificare se il loro comportamento è simile a quello di un *White Noise*.

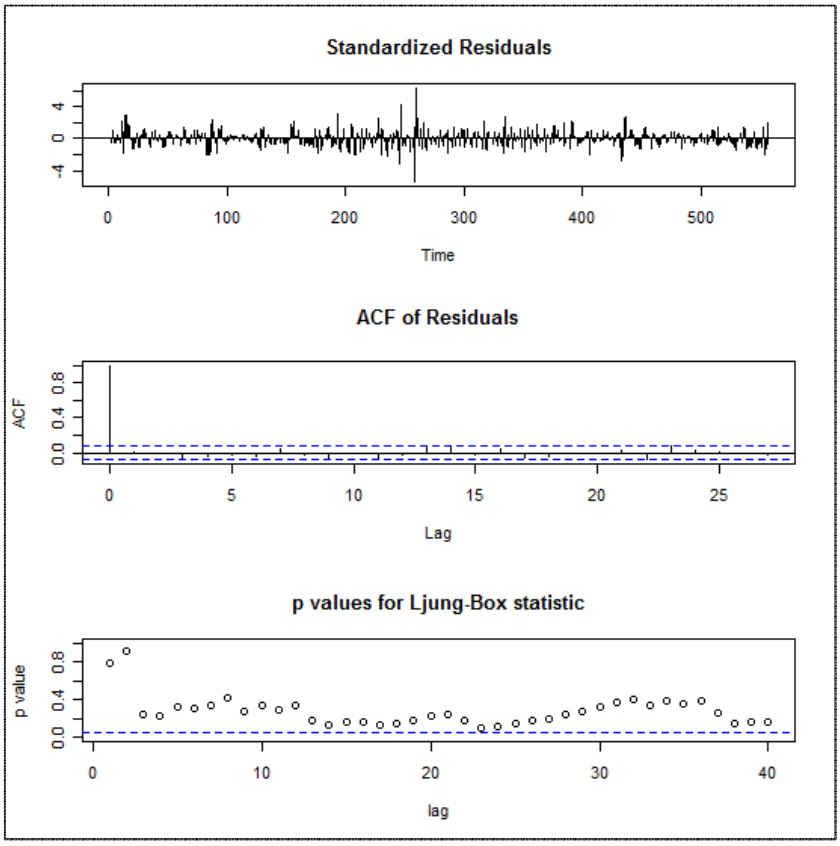



**GRAFICO 4.10: Residui modello consumo di gas**

**Tocco da Casauria - *ARIMA (3,1,1)***

Osservando la funzione di autocorrelazione dei residui nel Grafico 4.10 (*ACF of Residuals*), è possibile affermare che i residui risultano tra loro incorrelati, in quanto tale funzione è riconducibile alla funzione di autocorrelazione di un *White Noise*.

Anche la statistica di *Ljung-Box* conferma questa affermazione, in quanto è possibile accettare l'ipotesi nulla di incorrelazione seriale dei residui con un *p-value* di poco inferiore a 0,2.

Possiamo quindi affermare che il modello stimato è, tutto sommato, un buon modello ed è quindi possibile utilizzarlo per ottenere le previsioni giornaliere per il periodo 11/04/2011-30/04/2011. Si ottengono in questo modo le previsioni del seno iperbolico inverso del consumo giornaliero di gas, che dovranno poi essere riconvertite in metri cubi attraverso la trasformata seno iperbolico. Occorre comunque tenere in considerazione che le previsioni possono essere considerate attendibili solo per un numero ridotto di passi in avanti a causa del continuo aumento dell'errore standard di previsione.

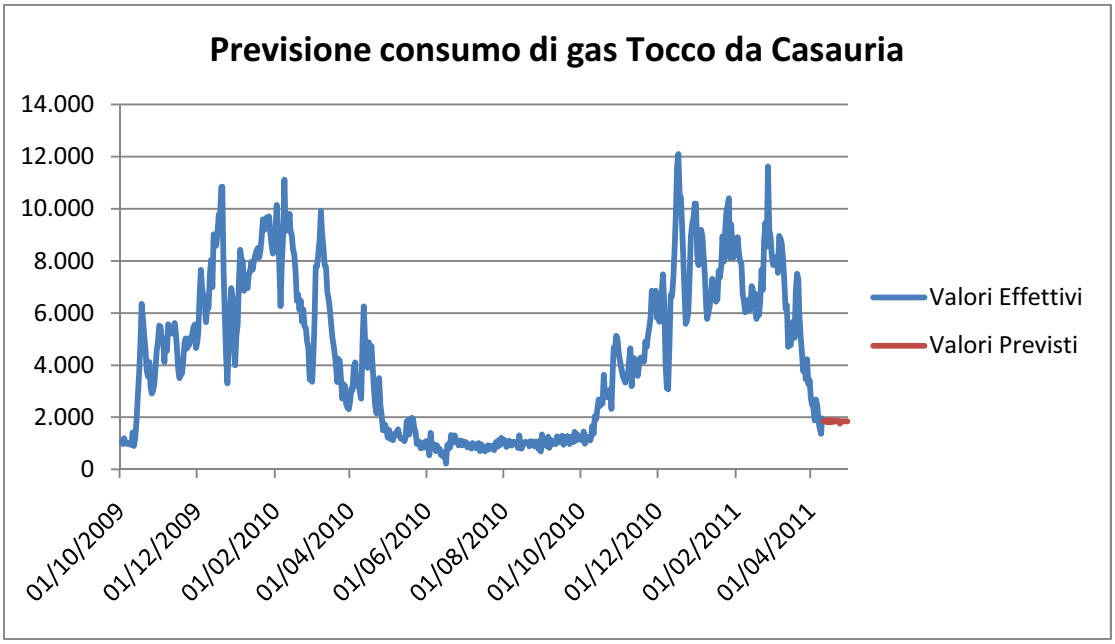



**GRAFICO 4.11: Previsioni consumo di gas Tocco da Casauria - *ARIMA (3,1,1)***

| DATA | CONSUMO |
|---|---|
| 11/04/2011 | 1.847 |
| 12/04/2011 | 1.858 |
| 13/04/2011 | 1.870 |
| 14/04/2011 | 1.813 |
| 15/04/2011 | 1.892 |
| 16/04/2011 | 1.811 |
| 17/04/2011 | 1.885 |
| 18/04/2011 | 1.823 |
| 19/04/2011 | 1.872 |
| 20/04/2011 | 1.834 |

| DATA | CONSUMO |
|---|---|
| 21/04/2011 | 1.861 |
| 22/04/2011 | 1.843 |
| 23/04/2011 | 1.855 |
| 24/04/2011 | 1.762 |
| 25/04/2011 | 1.852 |
| 26/04/2011 | 1.850 |
| 27/04/2011 | 1.850 |
| 28/04/2011 | 1.850 |
| 29/04/2011 | 1.850 |
| 30/04/2011 | 1.851 |

**TABELLA 4.6: Previsioni consumo di gas Tocco da Casauria - *ARIMA (3,1,1)***

### *4.1.3 CARTIERA DI FERRARA (destinazione d'uso industriale)*

La serie storica del consumo di gas presso il punto di riconsegna presso la Cartiera di Ferrara è composta da 365 osservazioni giornaliere dall'01/10/2009 al 30/09/2010; per stimare il modello, tuttavia, si decide di eliminare le osservazioni relative agli ultimi 20 giorni, sulle quali poi si andranno a effettuare poi le previsioni. L'insieme di stima risulta così composto da 345 osservazioni giornaliere dall'01/10/2009 al 10/09/2010, mentre le previsioni saranno effettuate sul periodo che va dall'11/09/2010 al 30/09/2010.

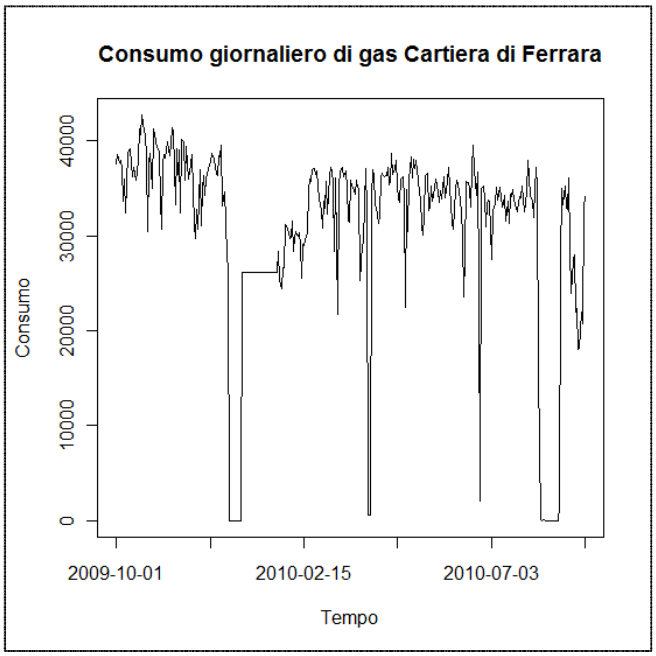



**GRAFICO 4.12: Serie storica consumo di gas Cartiera di Ferrara**

La serie storica giornaliera del consumo di gas presso la Cartiera di Ferrara sembra una serie abbastanza stazionaria in media (poiché il consumo giornaliero di gas è grosso modo costante durante l'Anno Termico, a causa dei processi industriali da svolgere con la necessaria continuità), e questo è confermato dal fatto che la funzione di autocorrelazione decresce abbastanza velocemente e in modo esponenziale, ma non stazionaria in varianza.

Per rendere la serie stazionaria in varianza si effettua un'opportuna trasformazione dei dati; vista la presenza di alcune osservazioni pari a 0 la trasformata ideale è, come visto nel capitolo precedente, il seno iperbolico inverso del consumo.

Si valuta ora il correlogramma della serie resa stazionaria in varianza.

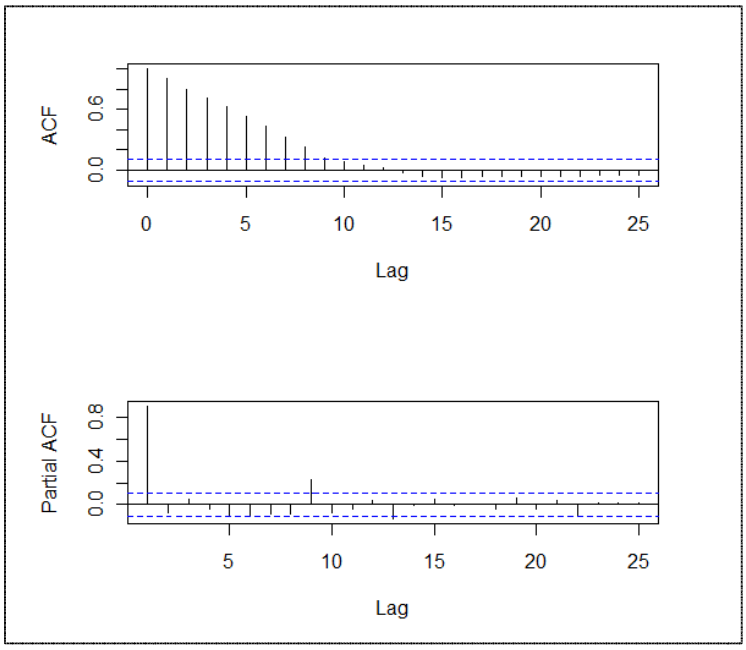



**GRAFICO 4.13: Correlogramma serie storica consumo di gas Cartiera di Ferrara**

La serie non sembra presentare alcuna componente stagionale; osservando il correlogramma è possibile individuare la presenza di una componente auto-regressiva di ordine 1. Il modello che si prova a stimare è quindi un *AR (1)*, con costante; le stime che si ottengono sono le seguenti:

| | coefficiente | errore std. | z | p-value | |
|---|---|---|---|---|---|
| const | 10.410 | 0.523 | 19.91 | 3.66e-088 | *** |
| phi_1 | 0.929 | 0.020 | 47.29 | 0.0000 | *** |

| | | | |
|---|---|---|---|
| Media var. dipendente | 10.369 | SQM var. dipendente | 2.642 |
| Media innovazioni | -0.002 | SQM innovazioni | 1.0549 |
| Log-verosimiglianza | -509.324 | Criterio di Akaike | 1026.648 |
| Criterio di Schwarz | 1042.022 | Hannan-Quinn | 1032.771 |

Note: SQM = scarto quadratico medio; E.S. = errore

**TABELLA 4.7: Modello consumo di gas Cartiera di Ferrara – *AR(1)***

Le stime dei parametri risultano tutte significativamente diverse da 0 a un livello di significatività fissato pari al 5%.

Il modello stimato per la serie storica del consumo giornaliero di gas presso la Cartiera di Ferrara è quindi il seguente:

$$(\mathbf{1 - 0,93}\boldsymbol{B})\boldsymbol{cons}_{\boldsymbol{ihs};\boldsymbol{t}} = \mathbf{10,41} + \boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

Occorre ora valutare la bontà del modello stimato; per fare ciò bisogna verificare la qualità dei residui ottenuti, e verificare se il loro comportamento è simile a quello di un *White Noise*.

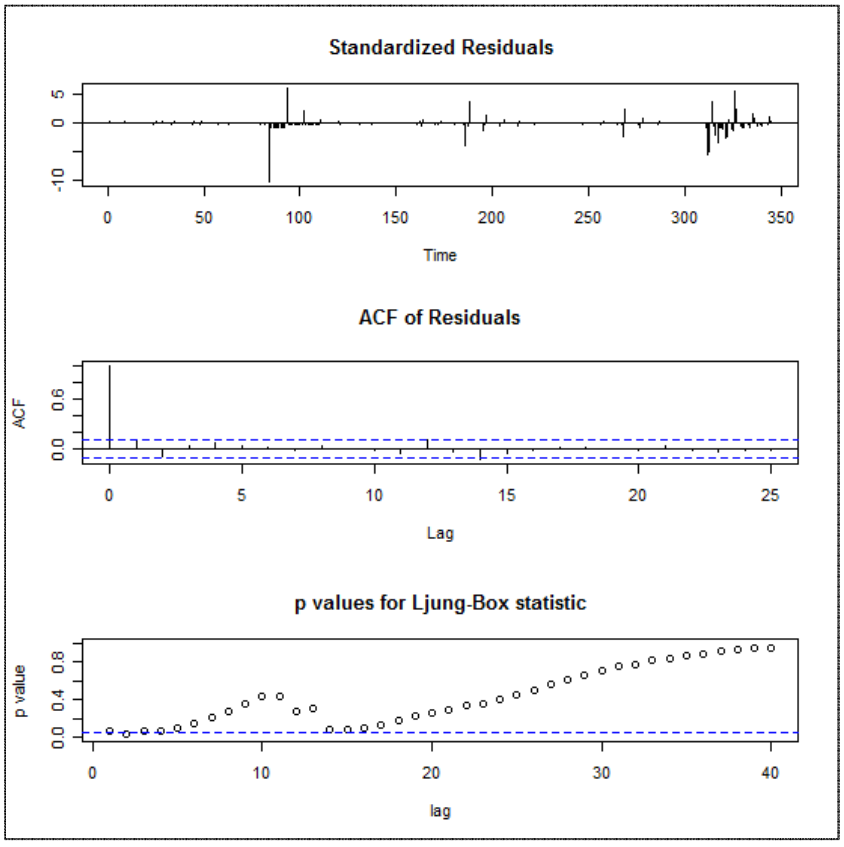



**GRAFICO 4.14: Residui modello consumo di gas Cartiera di Ferrara – *AR(1)***

Osservando la funzione di autocorrelazione dei residui nel Grafico 4.14 (*ACF of Residuals*), è possibile affermare che i residui risultano tra loro incorrelati, in quanto tale funzione è riconducibile alla funzione di autocorrelazione di un *White Noise*.

Anche la statistica di *Ljung-Box* conferma questa affermazione, in quanto è possibile accettare l'ipotesi nulla di incorrelazione seriale dei residui con un *p-value* pressoché pari a 1.

Possiamo quindi affermare che il modello stimato è, tutto sommato, un buon modello ed è quindi possibile utilizzarlo per ottenere le previsioni giornaliere per il periodo 11/09/2010-30/09/2010. Si ottengono in questo modo le previsioni del seno iperbolico inverso del consumo giornaliero di gas, che dovranno poi essere riconvertite in metri cubi attraverso la trasformata seno iperbolico. Occorre comunque tenere in considerazione che le previsioni possono essere considerate attendibili solo per un numero ridotto di passi in avanti a causa del continuo aumento dell'errore standard di previsione.

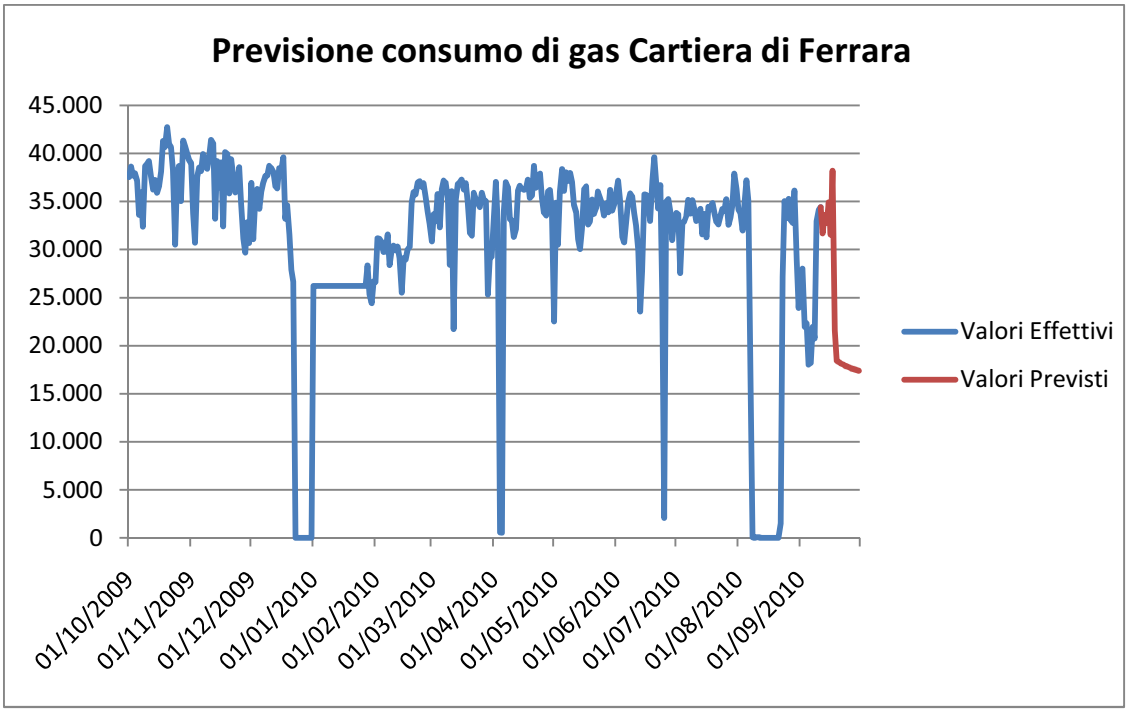



**GRAFICO 4.15: Previsioni consumo di gas Cartiera di Ferrara – *AR(1)***

| DATA | CONSUMO |
|---|---|
| 11/04/2011 | 34.425 |
| 12/04/2011 | 31.677 |
| 13/04/2011 | 33.630 |
| 14/04/2011 | 32.766 |
| 15/04/2011 | 34.902 |
| 16/04/2011 | 31.539 |
| 17/04/2011 | 38.219 |
| 18/04/2011 | 21.593 |
| 19/04/2011 | 18.423 |
| 20/04/2011 | 18.287 |

| DATA | CONSUMO |
|---|---|
| 21/04/2011 | 18.162 |
| 22/04/2011 | 18.046 |
| 23/04/2011 | 17.939 |
| 24/04/2011 | 17.841 |
| 25/04/2011 | 17.749 |
| 26/04/2011 | 17.665 |
| 27/04/2011 | 17.587 |
| 28/04/2011 | 17.515 |
| 29/04/2011 | 17.448 |
| 30/04/2011 | 17.386 |

**TABELLA 4.8: Previsioni consumo di gas Cartiera di Ferrara – *AR(1)***

I modelli *ARIMA* per l'analisi univariata delle serie storiche sono molto utili, in quanto tengono conto della correlazione seriale del consumo giornaliero di gas; l'onere del lavoro, tuttavia, è molto più elevato, in quanto, se volessimo ottenere previsioni per tutti i 279 punti di riconsegna considerati, occorerebbe stimare 279 modelli diversi, uno per ciascun punto di riconsegna.

## 4.2 I MODELLI A FUNZIONE DI TRASFERIMENTO

I modelli *ARIMA* discussi in precedenza rappresentano una classe di modelli generali, che possono essere utilizzati nella modellazione e nella previsione di serie storiche; l'assunzione implicita in questi modelli è che le condizioni di partenza, sotto cui vengono raccolti i dati per l'analisi e su cui vengono poi stimati i modelli, restano sempre le stesse. Se queste condizioni variano nel tempo, tuttavia, i modelli *ARIMA* possono essere migliorati, introducendo una determinata serie storica di input che rifletta i cambiamenti nelle condizioni iniziali del processo. Questa classe di modelli è detta **"a funzione di trasferimento" (Transfer Function Model)**; questi modelli possono essere visti

come dei modelli di regressione caratterizzati da serie storica di output, input e termine di errore serialmente correlati.
I modelli a funzione di trasferimento rappresentano quindi un sistema dinamico, che cerca di identificare una relazione esistente tra una serie storica di output $Y_t$ e una di input $X_t$, nel nostro caso, rispettivamente, il consumo giornaliero di gas e la temperatura media registrata il giorno della rilevazione del consumo. Lo scopo delle funzioni di trasferimento è di stimare la funzione $v(B)$ (a cui sarà poi legato un termine d'errore), basandosi sulle informazioni che si possono trovare nelle due serie storiche (*Montgomery-Jennings-Kulahci, 2008*).

Il sistema è rappresentato dalla seguente relazione:

$$Y_t = v_0 X_t + v_1 X_{t-1} + v_2 X_{t-2} + \cdots = v(B) X_t$$

con B che rappresenta l'operatore ritardo:

$$v(B) = v_0 + v_1 B + v_2 B^2 + \cdots = \sum_{j=0}^{\infty} v_j B^j$$

In pratica, si ritiene che la serie storica di output $Y_t$ sia influenzata per un numero infinito di istanti temporali da un input $X_t$, che ne modifica in maniera significativa l'andamento.
Per semplificare il sistema, riducendo il numero di coefficienti, si utilizza la seguente rappresentazione:

$$(1 - \delta_1 B - \delta_2 B^2 - \cdots - \delta_r B^r) Y_t = (\omega_0 - \omega_1 B - \omega_2 B^2 - \cdots - \omega_s B^s) X_{t-b}$$

$$\delta(B) Y_t = \omega(B) X_{t-b}$$

con:
$\delta(B)$ polinomio in B di grado $r$ (l'ordine di decadimento della funzione);

$\omega(B)$ polinomio in B di grado $s$ (l'ordine della regressione che comprime l'input);

$b$ parametro di ritardo, indica l'istante in cui l'effetto di $X_t$ si manifesta su $Y_t$.

Da questa formulazione si può ritornare, tramite un semplice calcolo, a:

$$Y_t = \frac{\omega(B)B^b}{\delta(B)} X_t = v(B)X_t$$

I pesi dei singoli parametri vengono calcolati eguagliando i coefficienti nell'equazione:

$$(1 - \delta_1 B - \cdots - \delta_r B^r)(v_0 + v_1 B + v_2 B^2 + \cdots) = (\omega_0 - \omega_1 B - \cdots - \omega_s B^s)B^b$$

dove:

$$\begin{aligned}
&v_j = 0 && per\ j = b \\
&v_j = \delta_1 v_{j-1} + \delta_2 v_{j-2} + \cdots + \delta_r v_{j-r} + \omega_0 && per\ j = b \\
&v_j = \delta_1 v_{j-1} + \delta_2 v_{j-2} + \cdots + \delta_r v_{j-r} - \omega_{j-b} && per\ j = b+1, \dots, b+s \\
&v_j = \delta_1 v_{j-1} + \delta_2 v_{j-2} + \cdots + \delta_r v_{j-r} && per\ j > b+s
\end{aligned}$$

Dopo aver descritto le caratteristiche principali dei modelli a funzione di trasferimento,

si passa ora alla descrizione del processo di identificazione e stima del modello, ricordando che il risultato a cui si deve arrivare (dopo aver stimato i relativi parametri), è il seguente:

$$Y_t = v(B)X_t + N_t = \frac{\omega(B)B^b}{\delta(B)} X_t + N_t, \quad N_t = \frac{\theta(B)}{\phi(B)} Z_t, \ \ con \ \ Z_t \sim WN(0, \sigma^2)$$

Le fasi da compiere per l'identificazione e la stima del modello sono le seguenti:

1) STIMA DELLA CORRELAZIONE INCROCIATA: dalla stima della correlazione incrociata tra le due variabili, si deve riuscire a capire quali siano i ruoli delle due serie storiche, in base alla posizione del grafico rispetto allo zero. Per meglio comprendere come sia possibile identificare le due componenti, si riporta un esempio grafico (Grafico 4.16) di una *CCF* (*Cross-Correlation Function*) fra due serie, da cui si deduce che $X_t$ è l'input e $Y_t$ l'output, visto che il grafico risulta a sinistra rispetto lo zero (se fosse stato a destra, sarebbe stato l'inverso); in tutti i passaggi successivi, si assume che la variabile $X_t$ sia l'input e $Y_t$ sia l'output.

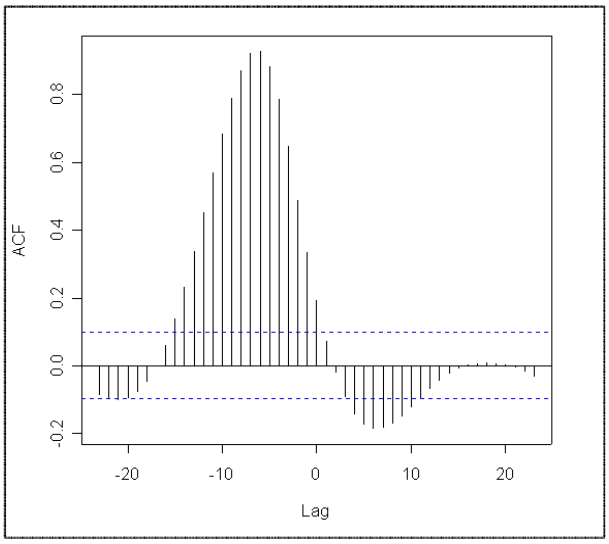



**GRAFICO 4.16: Esempio di cross-correlazione tra due serie storiche**

2) IDENTIFICAZIONE DEL MODELLO PER $X_t$ E SBIANCAMENTO DELLA SERIE DI INPUT: si deve trovare la serie che ha generato il modello per la serie di input, utilizzando la procedura di Box e Jenkins come descritto nel paragrafo precedente, in modo da trovare i parametri che costituiscono l'equazione:

$$\phi_x(B)X_t = \theta_x(B)\alpha_t, \quad con \ \alpha_t \sim WN(0, \sigma_\alpha^2)$$

Dopo aver identificato la serie generatrice, per sbiancarla è necessario compiere una semplice sostituzione nell'equazione, riscrivendola in funzione dell'errore, tramite la quale si ottiene la funzione di input sbiancata:

$$\alpha_t = \frac{\phi_x(B)}{\theta_x(B)} X_t$$

3) APPLICAZIONE DEL FILTRO ALLA SERIE $Y_t$: per poter valutare l'influenza dell'input su $Y_t$, lo stesso filtro deve essere applicato all'output, per ottenere la serie filtrata:

$$\beta_t = \frac{\phi_x(B)}{\theta_x(B)} Y_t$$

4) STIMA DEGLI IMPULSI $v_k$: servendosi della funzione di cross-correlazione tra le serie $\alpha_t$ e $\beta_t$, si possono stimare i valori degli impulsi, che permettono di identificare, al passo successivo, i valori di $b, r, s$. Essi sono ottenuti tramite la seguente relazione:

$$\hat{v}_k = \frac{\hat{\sigma}_\beta}{\hat{\sigma}_\alpha} \hat{\sigma}_{\alpha\beta}(k)$$

5) IDENTIFICAZIONE DEI VALORI DI $b, r, s$: dopo aver calcolato i valori degli impulsi, è possibile stimare i valori di $b, r, s$, attraverso le relazioni precedenti. L'identificazione di questi termini permette di ottenere una stima preliminare dei valori $\omega_j$e $\delta_j$ campionari, fino ad arrivare all'espressione finale:

$$\hat{v}(B) = \frac{\hat{\omega}_s(B)}{\hat{\delta}_r(B)} B^b$$

6) IDENTIFICAZIONE DEL MODELLO PER IL TERMINE DI ERRORE: si hanno ora a disposizione tutti i valori stimati dei parametri che legano la

variabile di input a quella di output, ed è possibile procedere con l'identificazione del modello per il termine d'errore $N_t$, ultima componente da sottoporre alla stima.

$$= Y_t - \hat{v}(B)X_t = Y_t - \frac{\hat{\omega}_s(B)}{\hat{\delta}_r(B)}B^b X_t$$

La serie $\widehat{N}_t$ che si ottiene viene poi identificata tramite un opportuno modello della classe *ARIMA*, con la procedura già vista in precedenza, ottenendo infine:

$$\phi(B)N_t = \theta(B)Z_t$$

7) STIMA E VERIFICA DEL MODELLO: avendo ora a disposizione tutte le componenti del modello, è possibile scriverlo in forma completa:

$$Y_t = \frac{\omega(B)}{\delta(B)}X_{t-b} + \frac{\theta(B)}{\phi(B)}Z_t$$

Per valutare la correttezza e l'adeguatezza del modello, è necessario sottoporlo a verifica, in modo da poter eventualmente apportare delle correzioni e migliorarlo. Le ipotesi fondamentali su cui si basa la costruzione di un modello a funzioni di trasferimento sono che la variabile di input $X_t$ e il termine d'errore $N_t$ siano indipendenti tra loro, così come $Z_t$ (distribuito come WN) e $\alpha_t$. Di conseguenza, le stime ottenute devono rispettare tali vincoli, che possono essere verificati tramite l'analisi dei residui del modello, che, in caso di inadeguatezza, diventano autocorrelati e cross-correlati con $X_t$ e con $\alpha_t$.

Dopo aver descritto come si identificano e si stimano i modelli a funzione di trasferimento, si riportano le analisi e i risultati ottenuti per i tre punti di

riconsegna precedentemente considerati (Argelato, Tocco da Casauria, Cartiera di Ferrara), con le relative valutazioni che tale modello permette di effettuare. In particolare, vengono analizzati i legami fra le due variabili, per determinare se realmente la temperatura media registrata il giorno della rilevazione del consumo di gas rappresenta un input per il sistema, e, in caso affermativo, vengono valutati gli istanti temporali in cui si verificano gli effetti e la loro durata.
L'ipotesi che si fa, ovviamente, è che sia la temperatura media ad influenzare il consumo di gas; per questo motivo, nelle analisi, si considererà la serie storica della temperatura media come serie di input, e la serie storica del consumo giornaliero di gas come serie di output.

### *4.2.1 ARGELATO (destinazione d'uso civile)*

Il primo passaggio da compiere per stimare un modello a funzione di trasferimento per il punto di riconsegna è adattare un modello alla serie storica delle temperature medie registrate presso il comune di Argelato (Bologna).

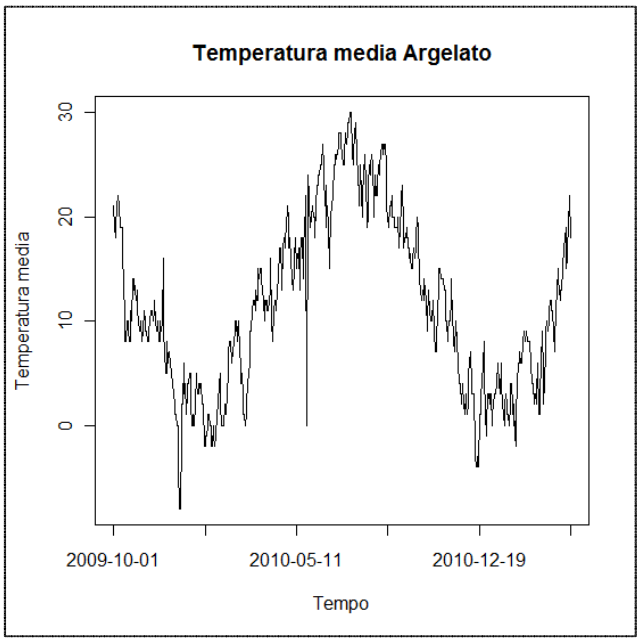



**GRAFICO 4.17: Serie storica temperatura media Argelato**

La serie storica sembra stazionaria in varianza, ma non in media; occorre verificare tramite un opportuno test *ADF* se il trend della serie è di tipo stocastico o deterministico.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.356      0.161   2.207   0.028  *
z.lag.1       -0.031      0.011  -2.736   0.006  **
z.diff.lag    -0.159      0.042  -3.768   0.0001 ***

Residual standard error: 2.229 on 552 degrees of freedom
Multiple R-squared: 0.043
F-statistic: 12.48 on 2 and 552 DF,  p-value: 4.978e-06

Value of test-statistic is: -2.736 3.744

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.43 -2.86 -2.57
phi1  6.43  4.59  3.78
```

**TABELLA 4.9: Test *ADF* serie storica consumo di gas Tocco da Casauria**

A un livello di significatività fissato pari al 5%, si accetta l'ipotesi nulla di non stazionarietà della serie; la serie è quindi caratterizzata da un trend di tipo stocastico, e per renderla stazionaria è sufficiente una differenziazione di lag pari a 1.

Si ottiene quindi la seguente funzione di auto-correlazione:

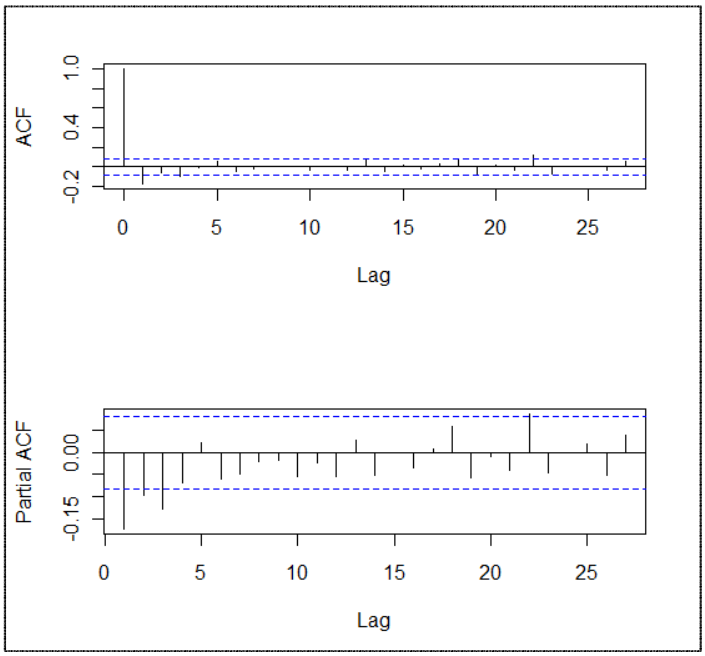



**GRAFICO 4.18: Correlogramma serie storica temperatura media Argelato**

La funzione di autocorrelazione non sembra evidenziare componenti stagionali nella serie; il modello adatto sembra essere un'*ARIMA(1,1,1).*

| | coefficiente | errore std. | z | p-value | |
|---|---|---|---|---|---|
| phi_1 | 0.590 | 0.104 | 5.667 | 1.45e-08 | *** |
| theta_1 | -0.796 | 0.080 | -9.921 | 3.37e-023 | *** |

| | | | |
|---|---|---|---|
| Media var. dipendente | -0.005 | SQM var. dipendente | 2.274 |
| Media innovazioni | -0.009 | SQM innovazioni | 2.202 |
| Log-verosimiglianza | -1227.877 | Criterio di Akaike | 2461.754 |
| Criterio di Schwarz | 2474.716 | Hannan-Quinn | 2466.817 |

Note: SQM = scarto quadratico medio; E.S. = errore standard

**TABELLA 4.10: Modello temperatura media Argelato - *ARIMA (1,1,1)***

Le stime dei parametri risultano tutte significativamente diverse da 0, a un livello di significatività fissato pari al 5%.
Il modello stimato per la serie storica della temperatura media registrata ad Argelato è quindi il seguente:

$$(\mathbf{1-0,59B})(\mathbf{1-B})\boldsymbol{temp}_{\boldsymbol{media;t}} = (\mathbf{1+0,80B})\boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

Occorre ora valutare la bontà del modello stimato; per fare ciò bisogna verificare la qualità dei residui ottenuti, e verificare se il loro comportamento è simile a quello di un *White Noise*.

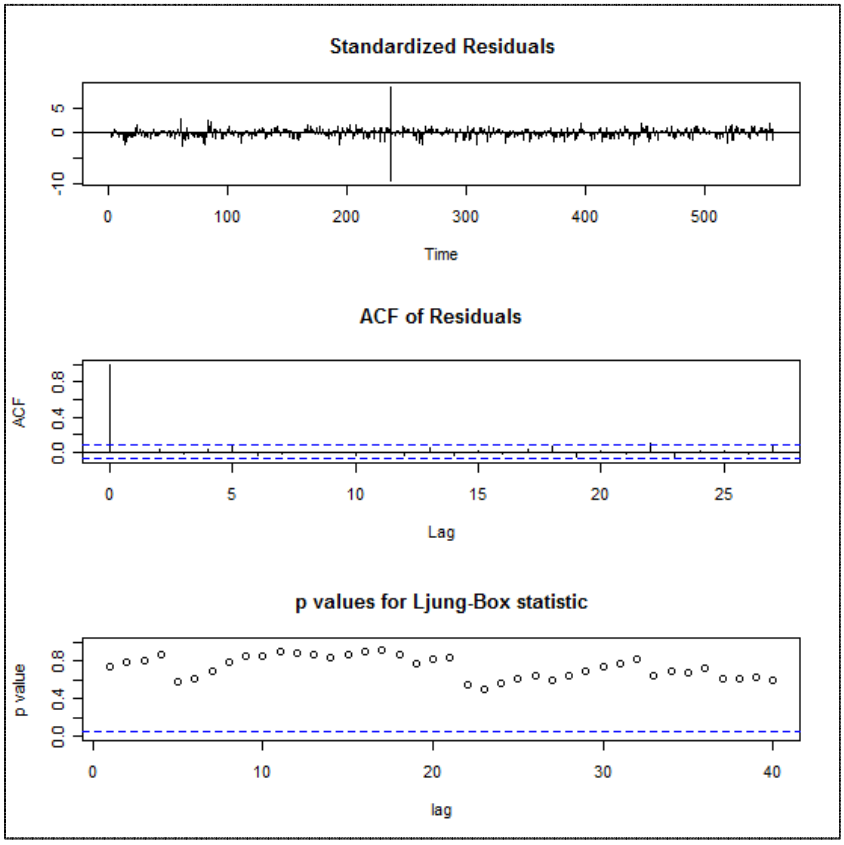



**GRAFICO 4.19: Residui modello temperatura media Argelato - *ARIMA (1,1,1)***

Osservando la funzione di autocorrelazione dei residui nel Grafico 4.19 (*ACF of Residuals*), è possibile affermare che i residui risultano tra loro incorrelati, in quanto tale funzione è riconducibile alla funzione di autocorrelazione di un *White Noise*.
Anche la statistica di *Ljung-Box* conferma questa affermazione, in quanto è possibile accettare l'ipotesi nulla di incorrelazione seriale dei residui con un *p-value* pari a circa 0,6.

Ora che è stato identificato il modello per la serie storica della temperatura media, utilizziamo tale risultato come filtro per la serie storica del consumo di gas:

$$(\mathbf{1-0,59}\boldsymbol{B})(\mathbf{1}-\boldsymbol{B})\boldsymbol{cons}_{\boldsymbol{ihs};\boldsymbol{t}} = (\mathbf{1+0,80}\boldsymbol{B})\boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

Si stima ora la funzione della correlazione incrociata tra la serie storica dei residui del modello stimato per la temperatura media e la serie del consumo di gas filtrata, ottenendo il Grafico 4.20:

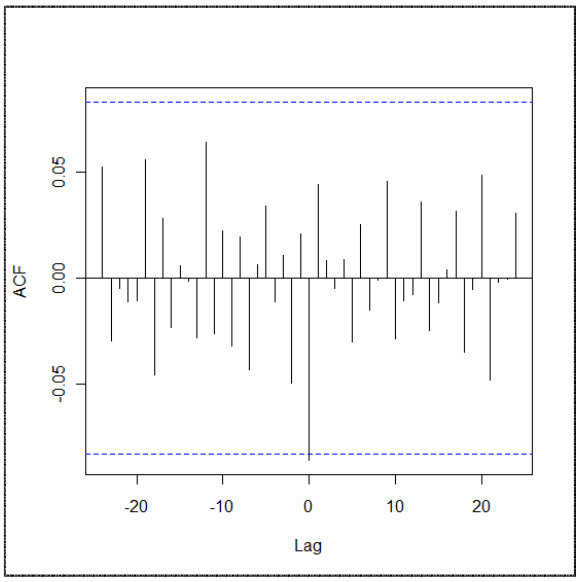



**GRAFICO 4.20: Argelato: correlazione incrociata tra la serie storica dei residui del modello stimato per la temperatura media e la serie storica del consumo di gas filtrata**

Il grafico della correlazione incrociata tra le due serie evidenzia correlazione contemporanea tra le serie; la temperatura media registrata influenza il consumo di gas del giorno di rifermento, e solamente quello.
Partendo da queste informazioni, e sfruttando la relazione tra gli impulsi e la funzione della correlazione incrociata, si può iniziare a stimare la funzione di trasferimento: la cross-correlazione è significativamente diversa da zero a partire dal ritardo 0, (quindi $\boldsymbol{b = 0}$), decresce immediatamente ($\boldsymbol{s = 0}$) e in maniera esponenziale ($\boldsymbol{r = 1}$).

Inserendo nel modello i valori dei parametri, stimati con l'aiuto del grafico ($\boldsymbol{b = 0, s = 0, r = 1}$), e adattando un modello appropriato alla componente d'errore, si ottengono i seguenti risultati:

| Parametro | Stima | Err. standard | Valore t | p-value | Ritardo |
|---|---|---|---|---|---|
| MA1,1 | 1.251 | 0.065 | 19.34 | <0.0001 | 1 |
| MA1,2 | -0.254 | 0.064 | -3.97 | <0.0001 | 2 |
| AR1,1 | 0.789 | 0.042 | 18.56 | <0.0001 | 1 |
| AR2,1 | -0.196 | 0.044 | -4.50 | <0.0001 | 7 |
| NUM1 | -0.024 | 0.004 | -6.80 | <0.0001 | 0 |
| DEN1,1 | 0.764 | 0.050 | 15.34 | <0.0001 | 1 |
| | | | | | |
| Stima varianza | 0.109 | | AIC | 348.699 | |
| Stima errore std | 0.331 | | SBC | 374.536 | |

**TABELLA 4.11: Modello a funzione di trasferimento consumo di gas Argelato**

Le stime dei parametri sono tutte significativamente diverse da 0, a un livello di significatività fissato pari al 5%; per valutare la bontà di adattamento del modello, si verifica se il comportamento dei residui è riconducibile a un *White Noise*.

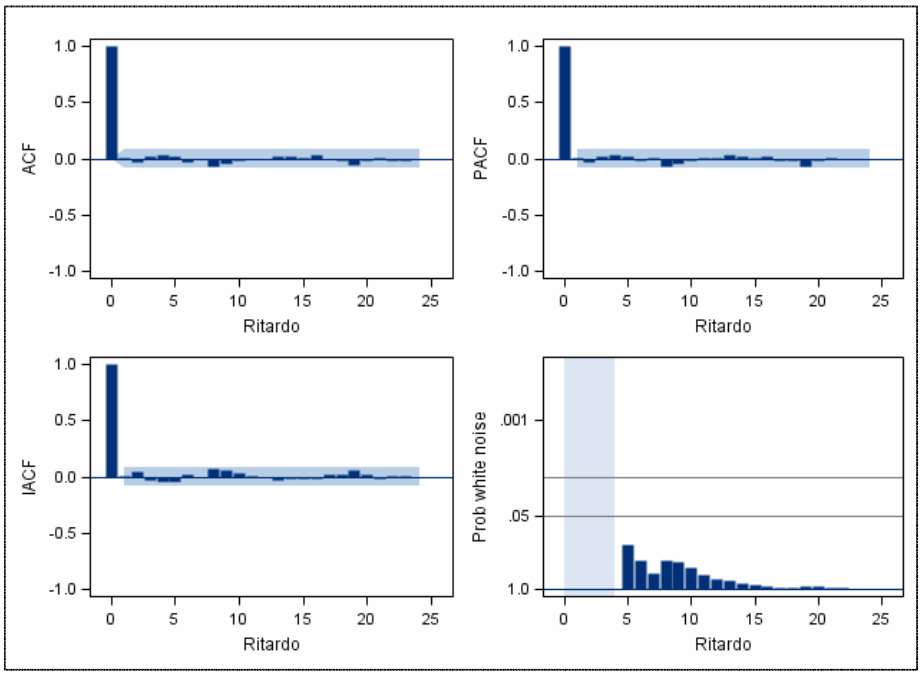



**GRAFICO 4.21: Residui modello a funzione di trasferimento consumo di gas Argelato**

I residui risultano fra loro incorrelati e la statistica di *Ljung-Box* porta sempre ad accettare l'ipotesi nulla di incorrelazione seriale.

| Fino al ritardo | Chi-Quadro | DF | p-value |
|---|---|---|---|
| 6 | 2.40 | 2 | 0.3008 |
| 12 | 5.96 | 8 | 0.6522 |
| 18 | 6.98 | 14 | 0.9356 |
| 24 | 9.43 | 20 | 0.9774 |
| 30 | 9.87 | 26 | 0.9982 |
| 36 | 10.72 | 32 | 0.9998 |
| 42 | 12.37 | 38 | 1.0000 |
| 48 | 17.90 | 44 | 0.9998 |

**TABELLA 4.12: Verifica incorrelazione residui modello a funzione di trasferimento consumo di gas Argelato**

Il modello a funzione di trasferimento stimato è quindi il seguente:

$$cons_{ihs;t} = \frac{-0,02}{1-0,76B} temp_{media;t} + \frac{1-1,25B+0,25B^2}{(1-0,79B)(1+0,20B^7)(1-B)(1-B^7)} \varepsilon_t$$

Da questi risultati si può affermare che, ad Argelato, l'effetto della temperatura media si ripercuote sul consumo di gas nello stesso giorno, e con una correlazione negativa; questo significa che, nello stesso giorno, un aumento della temperatura media registrata comporta una diminuzione nel consumo del gas e, viceversa, una sua diminuzione porta ad un aumento dei consumi.

Infine, possiamo affermare che il modello stimato è sicuramente un buon modello ed è quindi possibile utilizzarlo per ottenere le previsioni giornaliere per il periodo 11/04/2011-30/04/2011. Si ottengono in questo modo le previsioni del seno iperbolico inverso del consumo giornaliero di gas, che dovranno poi essere riconvertite in metri cubi attraverso la trasformata seno iperbolico. Occorre comunque tenere in considerazione che le previsioni possono essere considerate attendibili solo per un numero ridotto di passi in avanti, a causa del continuo aumento dell'errore standard di previsione, e che per la previsione del consumo di gas servono ulteriori informazioni relative alla temperatura media registrata nei giorni di previsione, in quanto è una variabile inserita all'interno del modello a funzione di trasferimento.

Si ottengono quindi i seguenti risultati:

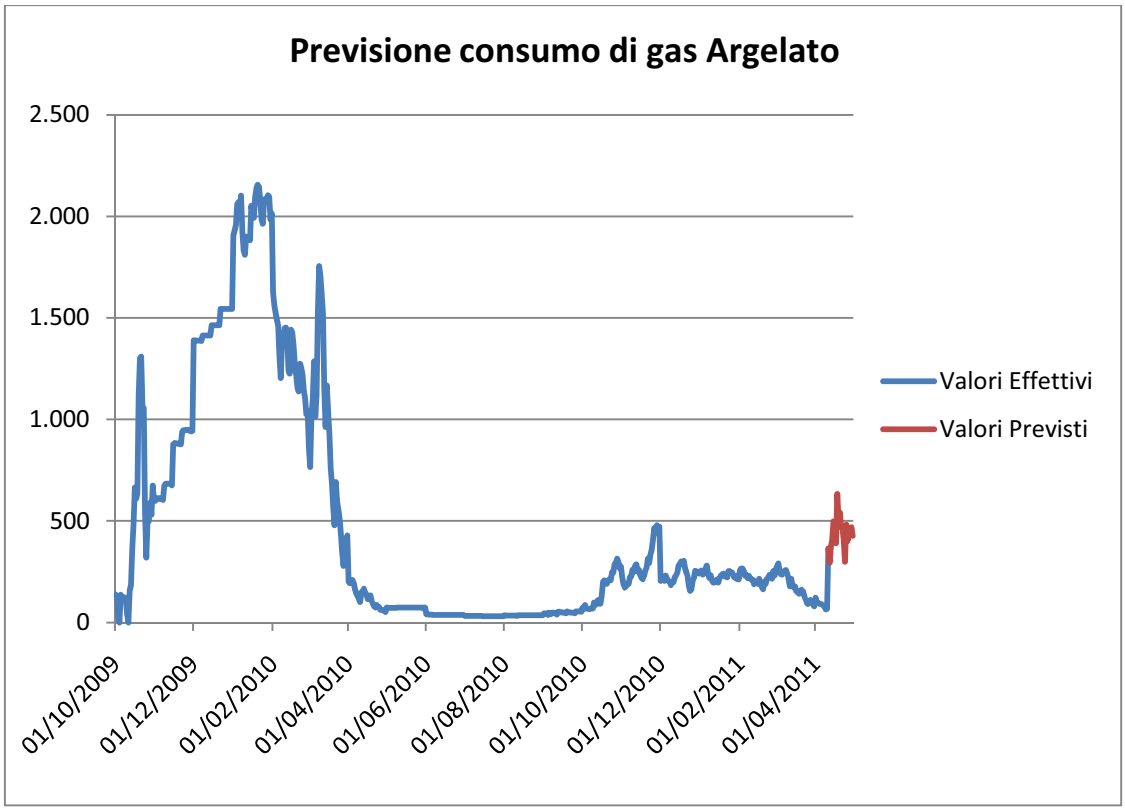



**GRAFICO 4.22: Previsioni modello a funzione di trasferimento consumo di gas Argelato**

| DATA | CONSUMO |
|---|---|
| 11/04/2011 | 364 |
| 12/04/2011 | 295 |
| 13/04/2011 | 373 |
| 14/04/2011 | 404 |
| 15/04/2011 | 497 |
| 16/04/2011 | 487 |
| 17/04/2011 | 391 |
| 18/04/2011 | 632 |
| 19/04/2011 | 506 |
| 20/04/2011 | 542 |

| DATA | CONSUMO |
|---|---|
| 21/04/2011 | 468 |
| 22/04/2011 | 468 |
| 23/04/2011 | 405 |
| 24/04/2011 | 300 |
| 25/04/2011 | 482 |
| 26/04/2011 | 402 |
| 27/04/2011 | 455 |
| 28/04/2011 | 440 |
| 29/04/2011 | 470 |
| 30/04/2011 | 426 |

**TABELLA 4.13: Previsioni modello a funzione di trasferimento consumo di gas Argelato**

### *4.2.2 TOCCO DA CASAURIA (destinazione d'uso civile e industriale)*

Il primo passaggio da compiere per stimare un modello a funzione di trasferimento per il punto di riconsegna è adattare un modello alla serie storica delle temperature medie registrate presso il comune di Tocco da Casauria (Pescara).

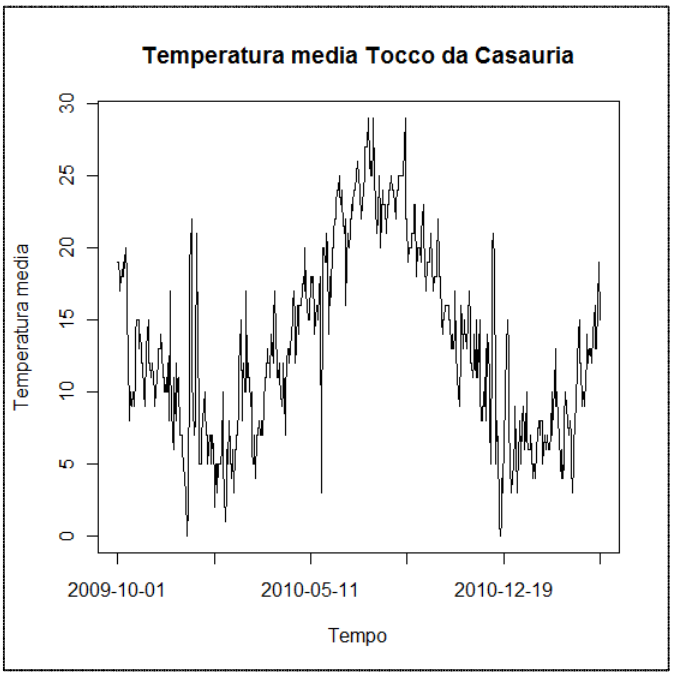



**GRAFICO 4.23: Serie storica temperatura media Tocco da Casauria**

La serie storica sembra stazionaria in varianza, ma non in media; occorre verificare tramite un opportuno test *ADF* se il trend della serie è di tipo stocastico o deterministico.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.356      0.161   2.207   0.028  *
z.lag.1       -0.031      0.011  -2.736   0.006  **
z.diff.lag    -0.159      0.042  -3.768   0.0002 ***

Residual standard error: 2.229 on 552 degrees of freedom
Multiple R-squared: 0.0433
F-statistic: 12.48 on 2 and 552 DF,  p-value: 4.978e-06

Value of test-statistic is: -2.736 3.744

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.43 -2.86 -2.57
phi1  6.43  4.59  3.78
```

**TABELLA 4.14: Test *ADF* serie storica temperatura media Tocco da Casauria**

Il test *ADF* conferma che, a un livello di significatività fissato del 5% la serie risulta non stazionaria e che quindi il trend è di tipo stocastico; per rendere la serie stazionaria in media è quindi possibile procedere con una differenziazione di ordine 1.

Si valuta ora il correlogramma della serie resa stazionaria in media e in varianza.

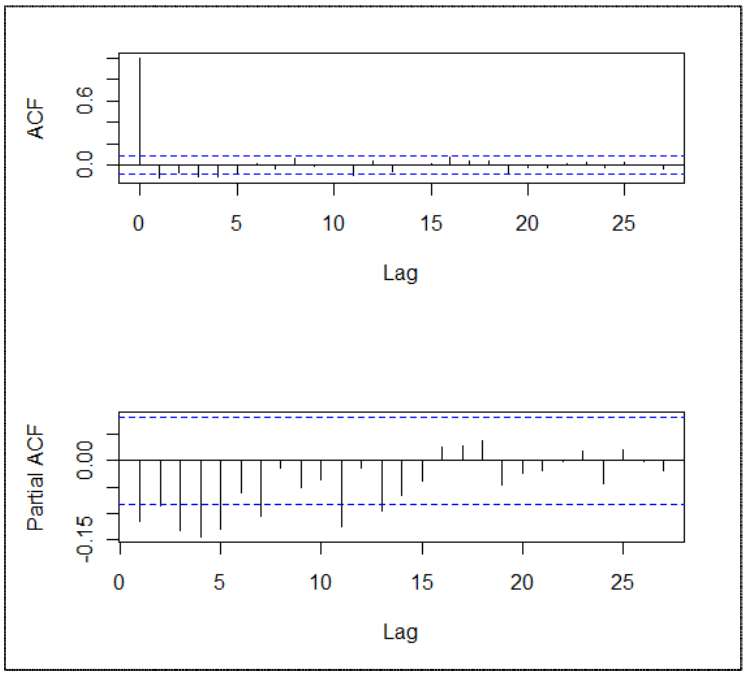



**GRAFICO 4.24: Correlogramma serie storica temperatura media Tocco da Casauria**

La funzione di autocorrelazione non sembra evidenziare componenti stagionali nella serie; il modello adatto sembra essere un'*ARIMA(1,1,1).*

| | coefficiente | errore std. | z | p-value | |
|---|---|---|---|---|---|
| phi_1 | 0.668 | 0.044 | 15.23 | 2.32e-052 | *** |
| theta_1 | -0.907 | 0.022 | -41.22 | 0.0000 | *** |

| | | | |
|---|---|---|---|
| Media var. dipendente | -0.007 | SQM var. dipendente | 2.430 |
| Media innovazioni | -0.027 | SQM innovazioni | 2.312 |
| Log-verosimiglianza | -1255.069 | Criterio di Akaike | 2516.138 |
| Criterio di Schwarz | 2529.100 | Hannan-Quinn | 2521.201 |

Note: SQM = scarto quadratico medio; E.S. = errore standard

**TABELLA 4.15: Modello temperatura media Tocco da Casauria - *ARIMA (1,1,1)***

Le stime dei parametri risultano tutte significativamente diverse da 0, a un livello di significatività fissato pari al 5%.
Il modello stimato per la serie storica della temperatura media registrata a Tocco da Casauria è quindi il seguente:

$$(\mathbf{1-0,67B})(\mathbf{1-B})\boldsymbol{temp}_{\boldsymbol{media;t}} = (\mathbf{1+0,91B})\boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

Occorre ora valutare la bontà del modello stimato; per fare ciò bisogna verificare la qualità dei residui ottenuti, e verificare se il loro comportamento è simile a quello di un *White Noise*.

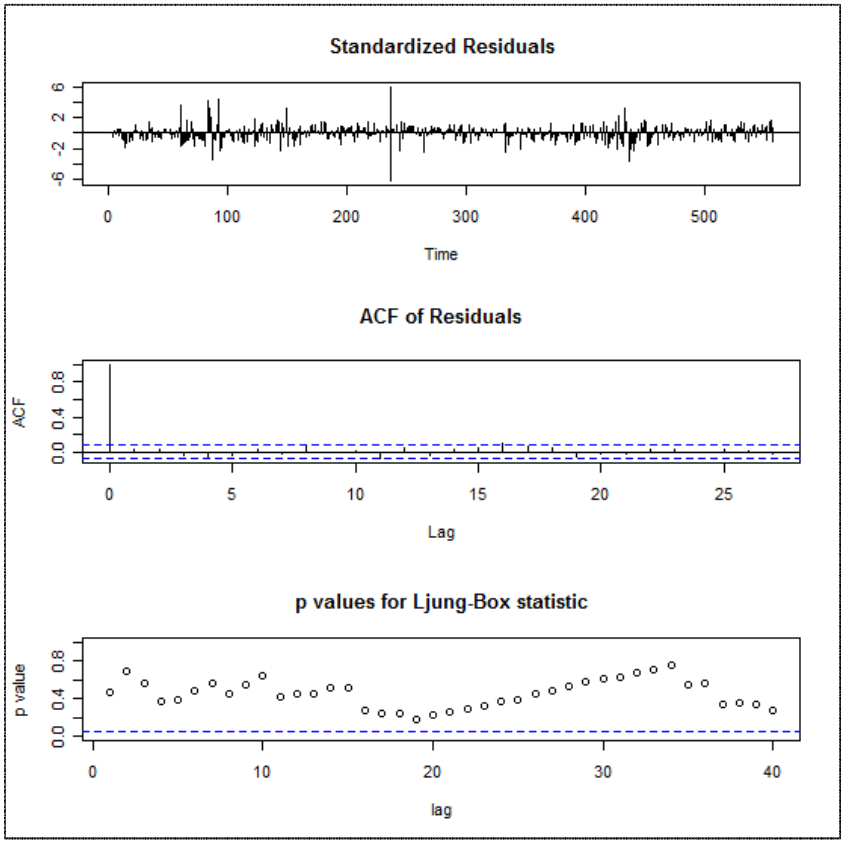



**GRAFICO 4.25: Residui modello temperatura media Tocco da Casauria - *ARIMA (1,1,1)***

Osservando la funzione di autocorrelazione dei residui nel Grafico 4.25 (*ACF of Residuals*), è possibile affermare che i residui risultano tra loro incorrelati, in quanto tale funzione è riconducibile alla funzione di autocorrelazione di un *White Noise*. Anche la statistica di *Ljung-Box* conferma questa affermazione, in quanto è possibile accettare l'ipotesi nulla di incorrelazione seriale dei residui con un *p-value* pari a circa 0,3.

Ora che è stato identificato il modello per la serie storica della temperatura media, utilizziamo tale risultato come filtro per la serie storica del consumo di gas:

$$(\mathbf{1-0,67}\boldsymbol{B})(\mathbf{1}-\boldsymbol{B})\boldsymbol{cons}_{\boldsymbol{ihs;t}} = (\mathbf{1+0,91}\boldsymbol{B})\boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

Si stima ora la funzione della correlazione incrociata tra la serie storica dei residui del modello stimato per la temperatura media e la serie del consumo di gas filtrata.

Si ottiene il seguente grafico:

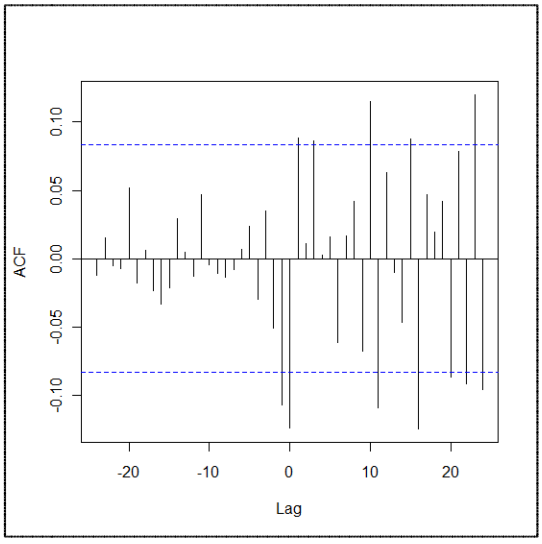



**GRAFICO 4.26: Tocco da Casauria: correlazione incrociata tra la serie storica dei residui del modello stimato per la temperatura media e la serie storica del consumo di gas filtrata**

Osservando il grafico della correlazione incrociata tra le serie filtrate (Grafico 4.26), è possibile osservare come la temperatura media non rappresenta distintamente l'input del sistema, in quanto sono presenti valori al di fuori delle bande di confidenza sia a destra che a sinistra dello 0; per questo motivo la stima del modello a funzione di trasferimento per il consumo di gas a Tocco da Casauria non si può calcolare, visto che risultano violate le ipotesi di base per la costruzione di tale classe di modelli, in quanto il consumo di gas non può essere assunto come output univoco del sistema.

### *4.2.3 CARTIERA DI FERRARA (destinazione d'uso industriale)*

Come svolto in precedenza, la prima operazione da compiere per stimare un modello a funzione di trasferimento per il punto di riconsegna è adattare un modello alla serie storica delle temperature medie registrate presso il comune di Ferrara.

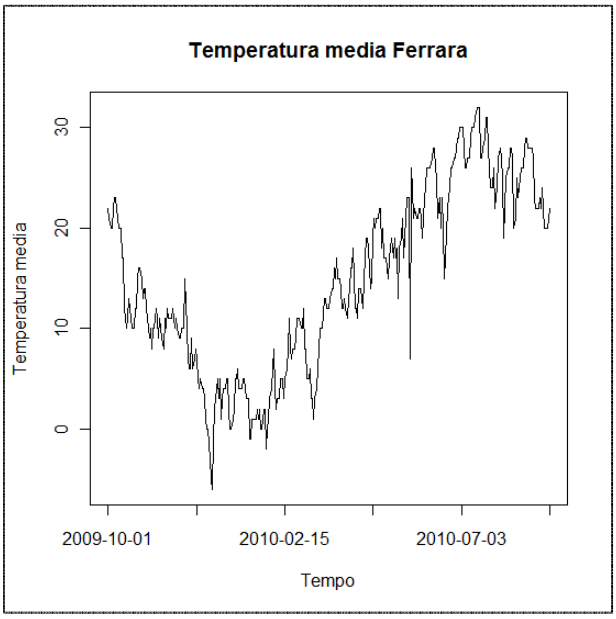



**GRAFICO 4.27: Serie storica temperatura media Ferrara**

La serie storica sembra stazionaria in varianza, ma non in media; occorre verificare tramite un opportuno test *ADF* se il trend della serie è di tipo stocastico o deterministico.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.392      0.236   1.659    0.098 .
z.lag.1       -0.026      0.014  -1.934    0.054 .
z.diff.lag    -0.190      0.053  -3.573   0.0004 ***

Residual standard error: 2.283 on 340 degrees of freedom
Multiple R-squared: 0.052
F-statistic: 9.282 on 2 and 340 DF,  p-value: 0.0001

Value of test-statistic is: -1.934 1.871

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
```

**TABELLA 4.16: Test *ADF* serie storica temperatura media Ferrara**

Il test *ADF* conferma che, a un livello di significatività fissato del 5% la serie risulta non stazionaria e che quindi il trend è di tipo stocastico; per rendere la serie stazionaria in media è quindi possibile procedere con una differenziazione di ordine 1.

Si valuta ora il correlogramma della serie resa stazionaria in media e in varianza.

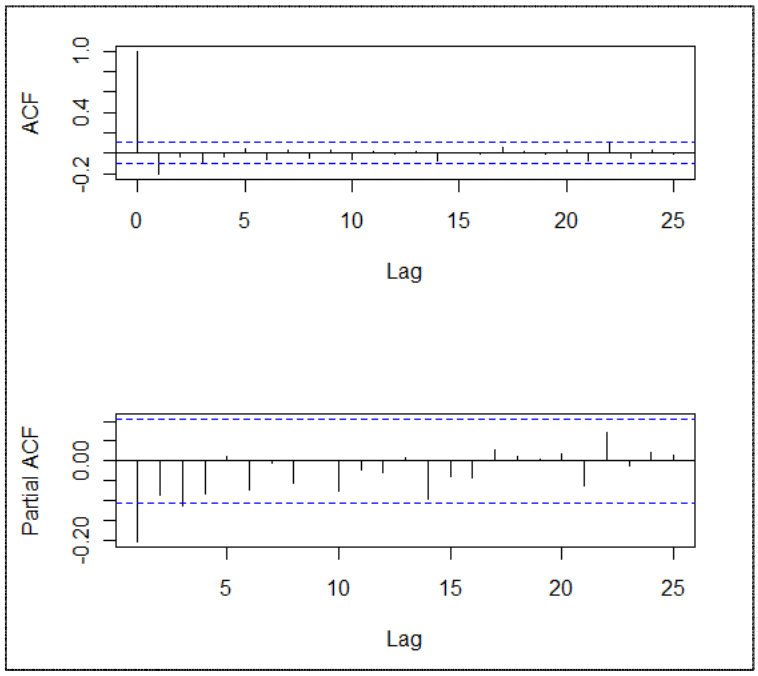



**GRAFICO 4.28: Correlogramma serie storica temperatura media Ferrara**

La funzione di autocorrelazione non sembra evidenziare componenti stagionali nella serie; il modello adatto sembra essere un'*ARIMA(1,1,1)*.

| | coefficiente | errore std. | z | p-value | |
|---|---|---|---|---|---|
| phi_1 | 0.583 | 0.141 | 4.138 | 3.50e-05 | *** |
| theta_1 | -0.804 | 0.108 | -7.426 | 1.12e-013 | *** |

| | | | |
|---|---|---|---|
| Media var. dipendente | 0.000 | SQM var. dipendente | 2.334 |
| Media innovazioni | -0.000 | SQM innovazioni | 2.249 |
| Log-verosimiglianza | -767.111 | Criterio di Akaike | 1540.222 |
| Criterio di Schwarz | 1551.744 | Hannan-Quinn | 1544.811 |

Note: SQM = scarto quadratico medio; E.S. = errore standard

**TABELLA 4.17: Modello temperatura media Ferrara - *ARIMA (1,1,1)***

Le stime dei parametri risultano tutte significativamente diverse da 0, a un livello di significatività fissato pari al 5%.
Il modello stimato per la serie storica della temperatura media registrata a Ferrara è quindi il seguente:

$$(\mathbf{1-0,58}\boldsymbol{B})(\mathbf{1-}\boldsymbol{B})\boldsymbol{temp}_{\boldsymbol{media;t}} = (\mathbf{1+0,80}\boldsymbol{B})\boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

Occorre ora valutare la bontà del modello stimato; per fare ciò bisogna verificare la qualità dei residui ottenuti, e verificare se il loro comportamento è simile a quello di un *White Noise*.

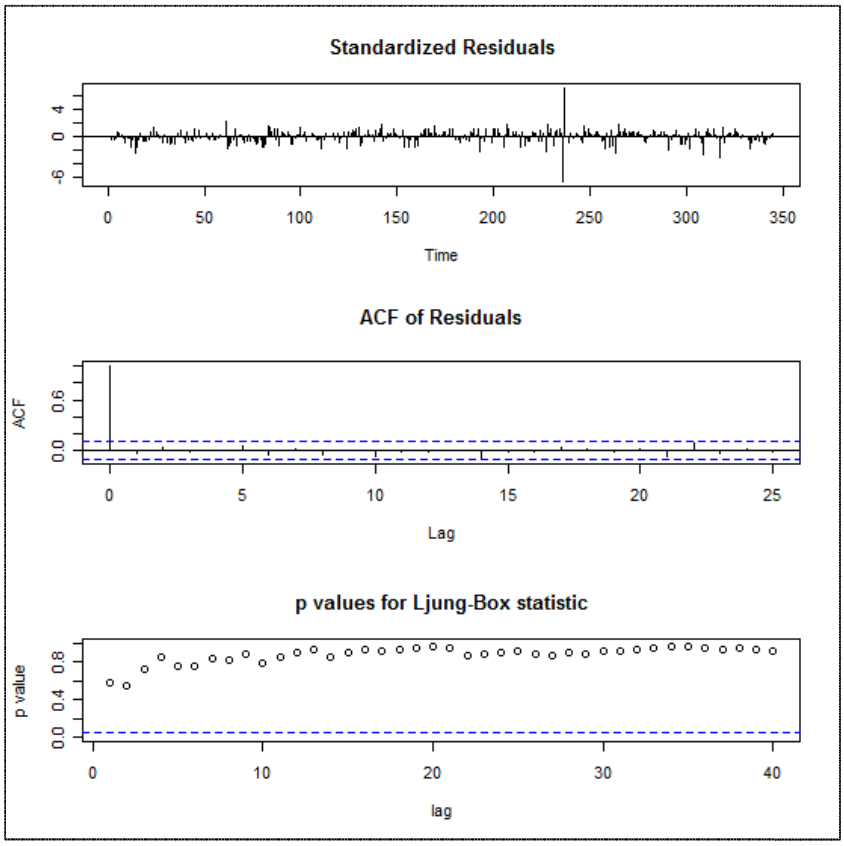



**GRAFICO 4.29: Residui modello temperatura media Ferrara - *ARIMA (1,1,1)***

Osservando la funzione di autocorrelazione dei residui nel Grafico 4.29 (*ACF of Residuals*), è possibile affermare che i residui risultano tra loro incorrelati, in quanto tale funzione è riconducibile alla funzione di autocorrelazione di un *White Noise*. Anche la statistica di *Ljung-Box* conferma questa affermazione, in quanto è possibile accettare l'ipotesi nulla di incorrelazione seriale dei residui con un *p-value* pari a circa 0,9.

Ora che è stato identificato il modello per la serie storica della temperatura media, utilizziamo tale risultato come filtro per la serie storica del consumo di gas:

$$(\mathbf{1-0,58}\boldsymbol{B})(\mathbf{1}-\boldsymbol{B})\boldsymbol{cons}_{\boldsymbol{ihs};\boldsymbol{t}} = (\mathbf{1+0,80}\boldsymbol{B})\boldsymbol{\varepsilon}_{\boldsymbol{t}}$$

Si stima ora la funzione della correlazione incrociata tra la serie storica dei residui del modello stimato per la temperatura media e la serie del consumo di gas filtrata.

Si ottiene il seguente grafico:

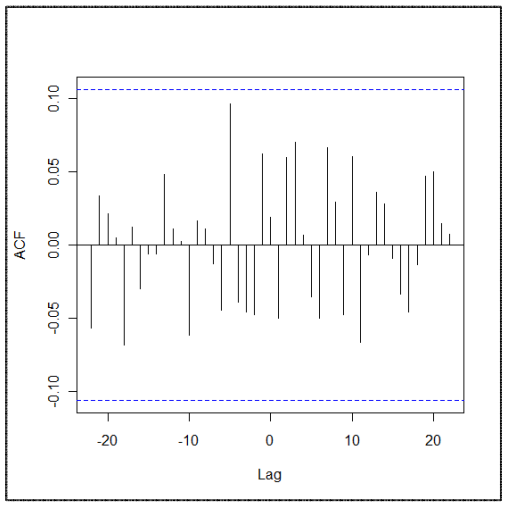



**GRAFICO 4.30: Cartiera di Ferrara: correlazione incrociata tra la serie storica dei residui del modello stimato per la temperatura media e la serie storica del consumo di gas filtrata**

Osservando il grafico della correlazione incrociata tra le serie filtrate, è possibile osservare come non sia possibile attribuire un ruolo ben definito alle variabili temperatura media e consumo di gas; sebbene la logica voglia che la temperatura media sia l'input e il consumo di gas l'output, dal Grafico 4.30 si evidenzia, invece, che non esiste un legame tra le due variabili, in quanto nessun valore risulta rilevante sia a destra che a sinistra dello 0. Questo conferma che per i punti ad uso industriale, come appunto quello presso la Cartiera di Ferrara, non sembra esserci alcuna relazione tra il consumo giornaliero di gas e la temperatura media registrata; questo è probabilmente dovuto al fatto che le industrie necessitano di un quantitativo di gas pressoché costante durante tutto l'Anno Termico, indipendentemente dalle condizioni climatiche, al fine di svolgere efficientemente e con continuità i processi industriali.
In base a queste osservazioni, possiamo quindi concludere che non è possibile condurre uno studio basato sulle funzioni di trasferimento, visto che risultano violate le ipotesi di base per la costruzione di tale classe di modelli.

Come per i modelli *ARIMA*, anche i modelli a funzione di trasferimento presentano un onere lavorativo molto elevato, in quanto sono modelli il cui processo di identificazione risulta abbastanza complesso, e anche perché ciascun punto di riconsegna presenta il proprio modello di riferimento; se volessimo quindi ottenere le previsioni per i 279 punti di riconsegna presi in considerazione, occorrerebbe stimare 279 modelli a funzione di trasferimento; come abbiamo visto nei 3 punti selezionati, inoltre, non sempre è possibile ottenere un modello di questa classe, in quanto a volte i dati non supportano le relazioni ipotizzate.

Per diminuire l'onere del lavoro, si potrebbe pensare di lavorare con modelli di analisi multivariata delle serie storiche, cercando di individuare eventuali relazioni tra più punti di riconsegna in base alla loro destinazione d'uso e in base alla loro collocazione geografica.

## 4.3 L'ANALISI MULTIVARIATA

L'analisi multivariata rappresenta un importante evoluzione rispetto all'analisi univariata, in quanto consente di prendere contemporaneamente in considerazione più punti di riconsegna; in questo modo, si modella non solo l'autocorrelazione presente all'interno di ciascuna serie storica, ma anche l'eventuale correlazione tra le serie storiche presenti nel modello.

I modelli che si utilizzeranno per l'analisi sono detti *VAR* (*Vectorial Auto-Regressive*), e hanno la seguente struttura:

$$y_t = \phi_0 + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + a_t$$

dove $y_t = (y_{1t}, \dots, y_{Kt})$ è un vettore Kx1 di variabili casuali, $\phi_j$, $j = 1, \dots, p$ sono matrici KxK di coefficienti, $\phi_0$ è un vettore Kx1 di costanti (intercette) e $a_t \sim WN(0, \Sigma)$, con Σ non singolare.

Le fasi per la stima di un modello *VAR* sono le stesse del caso univariato: identificazione, stima e controllo diagnostico; dopo aver identificato il modello adeguato e aver stimato i coefficienti, infatti, occorre verificare che i residui ottenuti siano soddisfacenti e, in caso contrario, proseguire con la procedura iterativa e rivedere le operazioni di identificazione e stima.

### *4.3.1 La stima del modello*

In questa analisi, si considerano tre punti di riconsegna ad uso civile della provincia di Verona, situati a Caldiero, Castelnuovo del Garda e Colognola ai Colli, le cui osservazioni vanno dall'01/04/2010 al 10/04/2011.

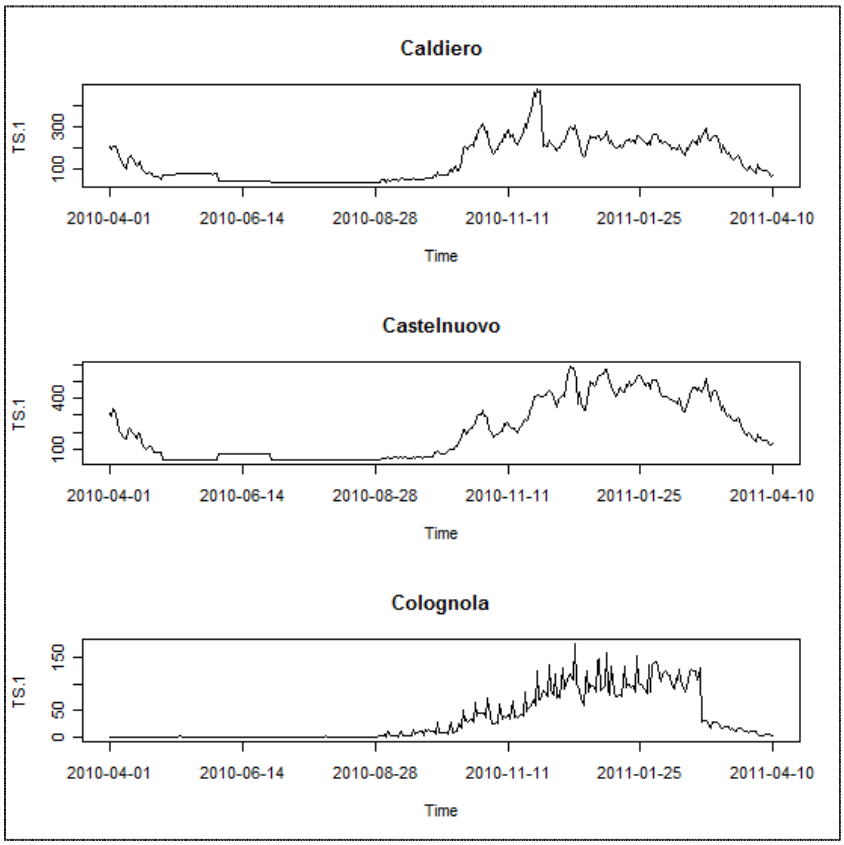



**GRAFICO 4.31: Serie storiche consumo di gas Caldiero, Castelnuovo e Colognola**

Le serie sembrano evidenziare andamenti analoghi nel consumo giornaliero di gas.

Tutte tre le serie, infatti, sembrano caratterizzate da un consumo significativo nei mesi invernali e pressoché nullo nei mesi estivi, presentando il classico andamento dei punti di riconsegna caratterizzati da una destinazione d'uso civile.

Come visto in precedenza, inoltre, le condizioni climatiche influenzano il consumo di gas, soprattutto per quel che riguarda i punti di riconsegna a destinazione d'uso esclusivamente civile. I tre punti presi in considerazione rientrano in questa categoria, e, quindi, inserire in un modello *VAR* le serie storiche delle temperature medie registrate nelle tre località in cui sono situati potrebbe migliorare la capacità previsiva del modello. Poiché i tre punti considerati sono limitrofi (si trovano tutti nella provincia di Verona), si può

assumere che la temperatura media registrata sia la stessa in tutte e tre le località; si inserirà nel modello *VAR*, quindi, un'unica serie storica, contenente la temperatura media registrata nella provincia di Verona.

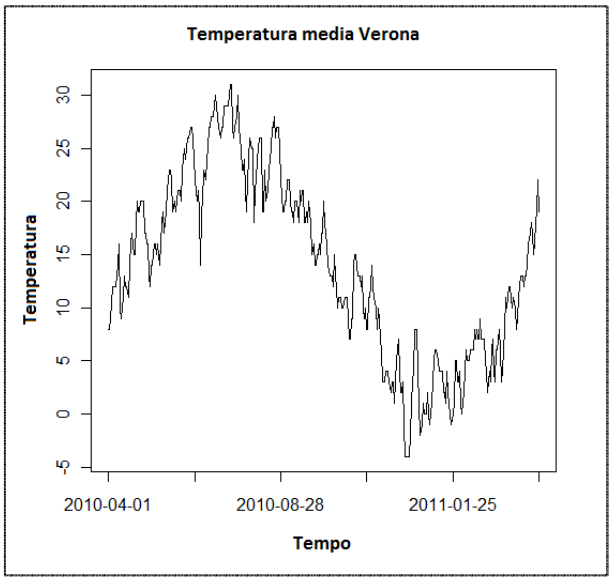



**GRAFICO 4.32: Serie storica temperatura media Verona**

Si verifica ora se sono presenti relazioni di feedback tra i tre punti e tra i punti e la temperatura registrata nella provincia di Verona, attraverso l'analisi delle funzioni di autocorrelazione e correlazione incrociata delle serie.

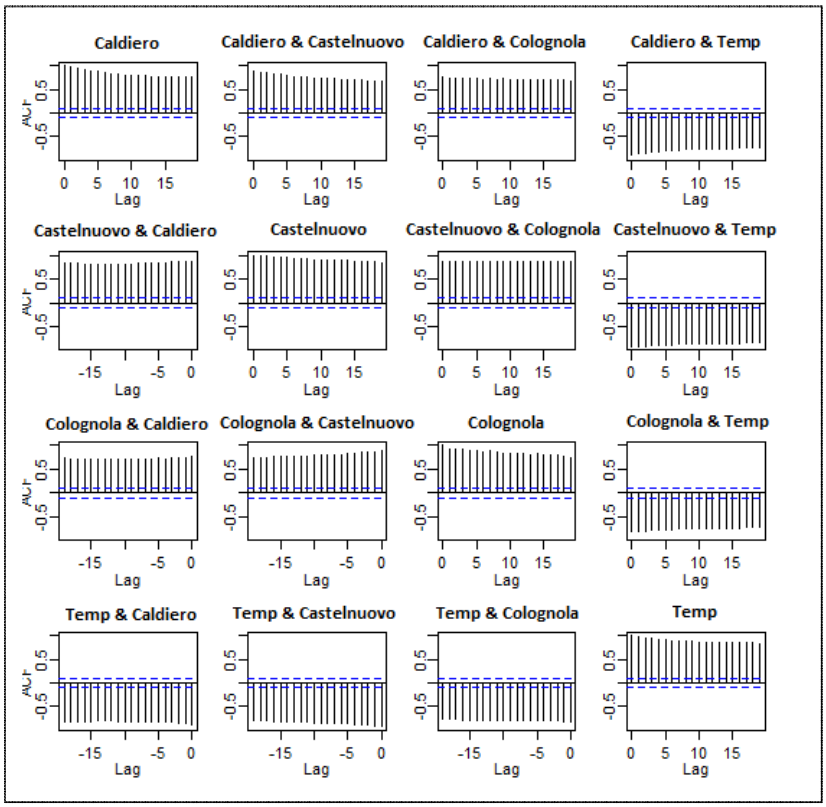



**GRAFICO 4.33: Correlazioni tra le componenti del *VAR***

Le funzioni di autocorrelazione e correlazione incrociata non consentono di cogliere andamenti e relazioni delle serie, in quanto le serie non risultano stazionarie né in media né in varianza. Per rendere le serie stazionarie in varianza, è sufficiente considerare le serie del seno iperbolico inverso del consumo giornaliero di gas, mentre per rendere le serie stazionarie in media occorre prima verificare se sono caratterizzate da un trend stocastico o da un trend deterministico, attraverso un opportuno test *ADF* che verifichi la presenza di eventuali radici unitarie per l'equazione caratteristica del processo.

Il test *ADF* per la serie storica del consumo giornaliero di gas a Caldiero da i seguenti risultati:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.047      0.038   1.231    0.219
z.lag.1        -0.010      0.007  -1.326    0.186
z.diff.lag     -0.065     0.0519  -1.254    0.211

Residual standard error: 0.1118 on 370 degrees of freedom
Multiple R-squared: 0.010
F-statistic:  1.78 on 2 and 370 DF,  p-value: 0.170

Value of test-statistic is: -1.326 1.015

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
```

**TABELLA 4.18: Test *ADF* serie storica consumo di gas Caldiero**

Il test *ADF* conferma che la serie storica del consumo giornaliero di gas a Caldiero è caratterizzata da un trend stocastico; per rendere la serie stazionaria, è quindi sufficiente calcolare la serie delle differenze prime.

Si procede effettuando il test *ADF* sulla serie storica del consumo giornaliero di gas a Castelnuovo del Garda, e ottenendo i seguenti risultati:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.032      0.029   1.118    0.264
z.lag.1        -0.006      0.005  -1.203    0.230
z.diff.lag      0.119      0.052   2.305    0.022 *

Residual standard error: 0.100 on 370 degrees of freedom
Multiple R-squared: 0.017
F-statistic: 3.279 on 2 and 370 DF,  p-value: 0.039
```

```
Value of test-statistic is: -1.203 0.788

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
```

**TABELLA 4.19: Test *ADF* serie storica consumo di gas Castelnuovo**

Il test *ADF* conferma che la serie storica del consumo giornaliero di gas a Castelnuovo del Garda è caratterizzata da un trend stocastico; per rendere la serie stazionaria, è quindi sufficiente calcolare la serie delle differenze prime.

Si calcola, successivamente, il test *ADF* sulla serie storica del consumo giornaliero di gas a Colognola ai Colli, ottenendo i seguenti risultati:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.054      0.035   1.530    0.127
z.lag.1       -0.017      0.010  -1.603    0.110
z.diff.lag    -0.516      0.044 -11.614   <2e-16 ***

Residual standard error: 0.406 on 370 degrees of freedom
Multiple R-squared: 0.279
F-statistic: 71.52 on 2 and 370 DF,  p-value: <2e-16

Value of test-statistic is: -1.603 1.368

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
```

**TABELLA 4.20: Test *ADF* serie storica consumo di gas Colognola**

Il test *ADF* conferma che la serie storica del consumo giornaliero di gas a Colognola ai Colli è caratterizzata da un trend stocastico; per rendere la serie stazionaria, è quindi sufficiente calcolare la serie delle differenze prime.

Si calcola, infine, il test *ADF* sulla serie storica della temperatura media registrata nella provincia di Verona, ottenendo i seguenti risultati:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.356      0.161   2.207    0.028 *
z.lag.1       -0.031      0.011  -2.736    0.006 **
z.diff.lag    -0.159      0.042  -3.768   0.0002 ***

Residual standard error: 2.229 on 552 degrees of freedom
Multiple R-squared: 0.043
F-statistic: 12.48 on 2 and 552 DF,  p-value: 4.978e-06

Value of test-statistic is: -2.736 3.744

Critical values for test statistics:
      1pct  5pct 10pct
tau2 -3.43 -2.86 -2.57
phi1  6.43  4.59  3.78
```

**TABELLA 4.21: Test *ADF* serie storica temperatura Verona**

Le serie storiche del consumo giornaliero di gas a Caldiero e Colognola ai Colli, inoltre, sono caratterizzate da una componente stagionale settimanale; per eliminarla, è sufficiente considerare le rispettive serie storiche delle differenze di ordine 7.

Le funzioni di autocorrelazione e correlazione incrociata delle serie storiche, destagionalizzate, e rese stazionarie in media e in varianza, presentano la seguente struttura

:

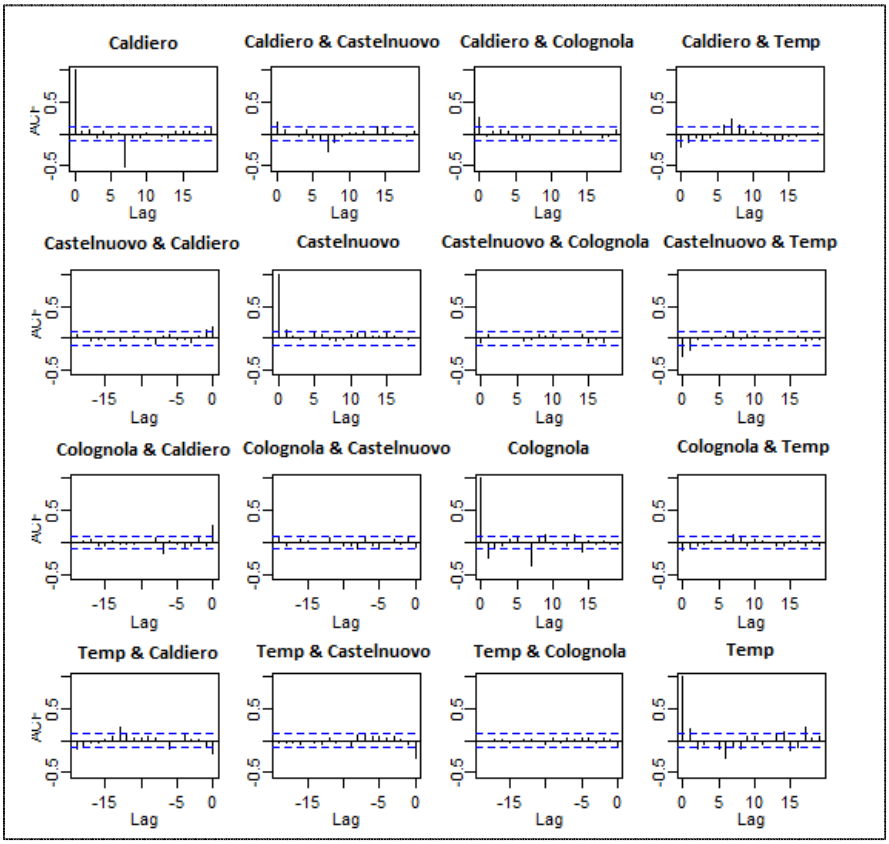



**GRAFICO 4.34: Correlazioni tra le componenti del *VAR* rese stazionarie e destagionalizzate**

Osservando il Grafico 4.34 sembra possibile affermare che:

- il consumo di gas a Caldiero e Castelnuovo del Garda sembra caratterizzato da correlazione contemporanea e settimanale (Caldiero sembra anticipare Castelnuovo);
- il consumo di gas a Caldiero e Colognola ai Colli sembra caratterizzato da correlazione contemporanea e settimanale (sembra che la serie storica di Caldiero anticipi quella di Colognola ai Colli);

- il consumo di gas a Castelnuovo del Garda e Colognola ai Colli non sembra caratterizzato da alcun tipo di correlazione. Le due serie sembrano essere indipendenti;
- la serie storica della temperatura media sembra essere caratterizzata da correlazione contemporanea con ciascuna delle tre serie relative al consumo giornaliero di gas nei tre punti di riconsegna.

Dopo aver analizzato le funzioni di autocorrelazione e correlazione incrociata, occorre identificare l'ordine del modello *VAR*. Per effettuare questa identificazione ci si serve di alcuni criteri di informazione, che consentono di determinare l'ordine ottimale del modello; i criteri utilizzati sono l'*AIC* (*Akaike Iformation Criteria*), l'*HQ* (*Hannan-Quinn Information Criteria*), l'*SC* (*Schwartz Information Criteria*) e l'*FPE* (*Final Prediction Error*).

| AIC(n) | HQ(n) | SC(n) | FPE(n) |
|---|---|---|---|
| 8 | 7 | 1 | 8 |

| | 1 | 2 | 3 | 4 |
|---|---|---|---|---|
| AIC(n) | -1.246e+01 | -1.246e+01 | -1.244e+01 | -1.239e+01 |
| HQ(n) | -1.239e+01 | -1.232e+01 | -1.224e+01 | -1.211e+01 |
| SC(n) | -1.229e+01 | -1.212e+01 | -1.192e+01 | -1.169e+01 |
| FPE(n) | 3.874e-06 | 3.872e-06 | 3.948e-06 | 4.163e-06 |

| | 5 | 6 | 7 | 8 |
|---|---|---|---|---|
| AIC(n) | -1.237e+01 | -1.242e+01 | -1.291e+01 | -1.295e+01 |
| HQ(n) | -1.202e+01 | -1.201e+01 | -1.243e+01 | -1.240e+01 |
| SC(n) | -1.150e+01 | -1.139e+01 | -1.170e+01 | -1.156e+01 |
| FPE(n) | 4.239e-06 | 4.005e-06 | 2.454e-06 | 2.371e-06 |

**TABELLA 4.22: Identificazione dell'ordine del modello *VAR***

La maggioranza dei criteri di informazione utilizzati suggerisce che l'ordine ottimale del modello è pari a 8; occorre quindi stimare un modello *VAR(8)*.

Poiché l'analisi è finalizzata alla previsione del consumo giornaliero di gas, dai risultati che si otterranno dalla stima del modello, si riporteranno solo le equazioni relative ai tre punti di riconsegna:

```
Estimation results for equation Caldiero_t:

Caldiero_t = Temperatura_t-1 + Caldiero_t-7 + Castelnuovo_t-7

                  Estimate Std. Error t value Pr(>|t|)
Temperatura_t-1      -0.064      0.015  -4.177 3.73e-05 ***
Caldiero_t-7         -0.541      0.043 -12.339  < 2e-16 ***
Castelnuovo_t-7      -0.229      0.061  -3.735   0.0002 ***


Residual standard error: 0.115 on 356 degrees of freedom
Multiple R-Squared: 0.356
F-statistic: 65.84 on 3 and 356 DF,  p-value: < 2.2e-16


Estimation results for equation Castelnuovo_t:

Castelnuovo_t = Temperatura_t-1 + Castelnuovo_t-5

                   Estimate Std. Error t value Pr(>|t|)
Temperatura_t-1      -0.048      0.012  -3.802   0.0001 ***
Castelnuovo_t-5       0.124      0.051   2.417   0.0161 *

Residual standard error: 0.09701 on 357 degrees of freedom
Multiple R-Squared: 0.05193
F-statistic: 9.778 on 2 and 357 DF,  p-value: 7.342e-05


Estimation results for equation Colognola_t:

Colognola_t = Colognola_t-1 + Colognola_t-2 + Colognola_t-7 +
Temperatura_t-7 + Colognola_t-8

                  Estimate Std. Error t value Pr(>|t|)
Colognola_t-1       -0.334      0.051  -6.444 3.80e-10 ***
```

```
Colognola_t-2       -0.145      0.048  -3.009   0.0028 **
Colognola_t-7       -0.372      0.049  -7.608 2.53e-13 ***
Temperatura_t-7      0.096      0.046   2.059   0.0402 *
Colognola_t-8       -0.138      0.052  -2.650   0.0084 **

Residual standard error: 0.353 on 354 degrees of freedom
Multiple R-Squared: 0.229
F-statistic: 21.14 on 5 and 354 DF,  p-value: < 2.2e-16

Covariance matrix of residuals:
               Caldiero    Castelnuovo    Colognola   Temperatura
Caldiero          0.014          0.002        0.010        -0.006
Castelnuovo       0.002          0.010       -0.002        -0.010
Colognola         0.010         -0.002        0.134        -0.011
Temperatura      -0.006         -0.010       -0.011         0.136
```

**TABELLA 4.23: Stima del modello *VAR(8)***

Le equazioni relative ai tre punti di riconsegna sono quindi le seguenti:

Caldiero: $y_{1,t} = -0{,}06y_{4,t-1} - 0{,}54y_{1,t-7} - 0{,}23y_{2,t-7} + a_{1,t}$

Castelnuovo: $y_{2,t} = -0{,}05y_{4,t-1} - 0{,}12y_{2,t-5} + a_{2,t}$

Colognola: $y_{3,t} = -0{,}34y_{3,t-1} - 0{,}15y_{3,t-2} - 0{,}37y_{3,t-7} + 0{,}09y_{4,t-7} - 0{,}14y_{3,t-8} + a_{3,t}$

La matrice di varianza e covarianza dei residui, invece, avrà la seguente struttura:

$$\Sigma = \begin{pmatrix} 0{,}014 & 0{,}002 & 0{,}011 & -0{,}007 \\ 0{,}002 & 0{,}010 & -0{,}003 & -0{,}010 \\ 0{,}011 & -0{,}003 & 0{,}135 & -0{,}011 \\ -0{,}007 & -0{,}010 & -0{,}011 & 0{,}136 \end{pmatrix}$$

Per verificare se questo modello è stato stimato correttamente, occorre verificarne l'adeguatezza attraverso un'analisi approfondita dei residui, sui quali si andrà a controllare l'assenza di autocorrelazione, la presenza o meno di eteroschedasticità condizionale (effetti *ARCH*), e l'eventuale normalità.

Per prima cosa, si analizzano le funzioni di autocorrelazione totale e parziale dei residui di ciascuna serie, e si verifica che in tutte vi sia incorrelazione seriale.

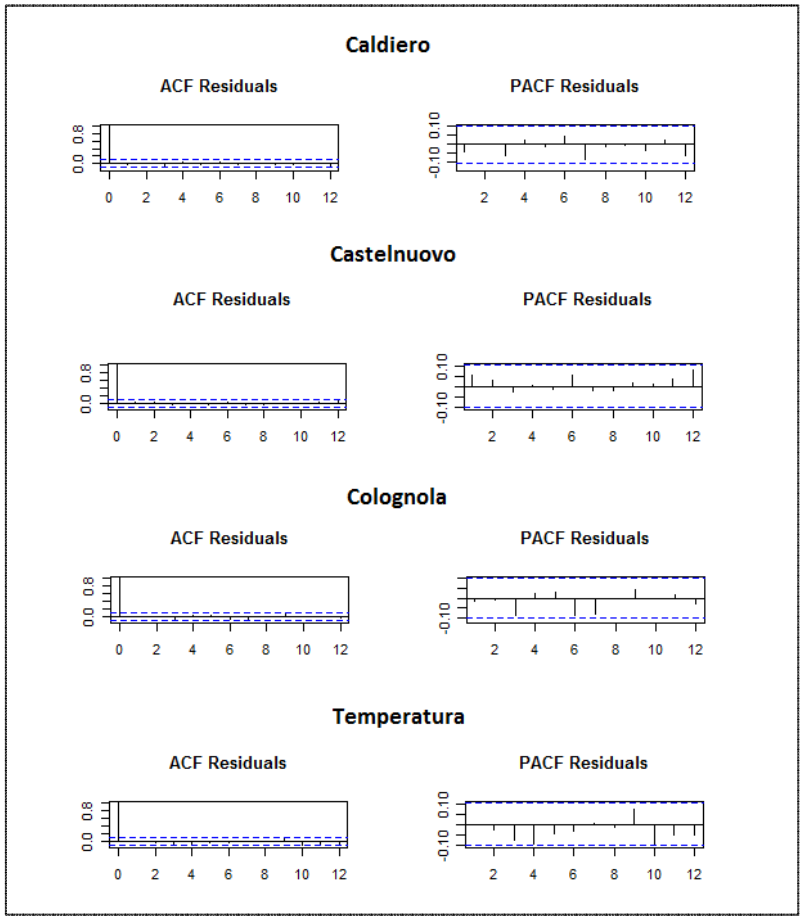



**GRAFICO 4.35: Correlazione residui del modello *VAR(8)***

Osservando le funzioni di autocorrelazione totale e parziale, sembra sia possibile affermare che non sembra esserci correlazione seriale nei residui di ciascuna serie.

Un'analisi di questo tipo, tuttavia, non risulta sufficiente, in quanto occorre anche effettuare un'analisi di tipo multivariato, verificando, in questo caso, che non vi siano tracce di correlazioni incrociate significative tra i residui delle tre serie; per compiere questa analisi ci si può servire della statistica *LM* di *Breusch-Godfrey*:

```
Breusch-Godfrey LM test

Chi-squared = 89.950, df = 80, p-value = 0.209
```

**TABELLA 4.24: Verifica incorrelazione residui modello *VAR(8*)**

L'ipotesi nulla viene accettata con un *p-value* pari a circa 0,21, e questo permette di concludere che non vi sono tracce di presenza di autocorrelazioni e correlazioni incrociate significative nei residui del modello.

Dopo aver verificato l'assenza di correlazione, occorre verificare che non vi siano componenti *ARCH* nei residui del modello:

```
ARCH(Multivariate)

Chi-squared = 437.742, df = 500, p-value = 0.979
```

**TABELLA 4.25: Verifica assenza effetti *ARCH* sui residui modello *VAR(8)***

L'ipotesi nulla viene accettata con un *p-value* pari a circa 1, e questo permette di concludere che non sembra esserci eteroschedasticità condizionale nei residui del modello.

Infine, occorre verificare simmetria, curtosi e normalità sugli errori del modello, attraverso l'utilizzo della versione multivariata del test di *Jarque-Brera*:

```
JB-Test (multivariate)

Chi-squared = 21721.10, df = 8, p-value < 2.2e-16


Skewness only (multivariate)

Chi-squared = Chi-squared = 149.115, df = 4, p-value < 2.2e-16


Kurtosis only (multivariate)

Chi-squared = 21571.99, df = 4, p-value < 2.2e-16
```

**TABELLA 4.26: Verifica normalità residui modello *VAR(8)***

I risultati della versione multivariata del test di *Jarque-Brera* portano a concludere che i residui del modello stimato non sono normali, e questo è probabilmente dovuto alla presenza di alcuni valori anomali nelle tre serie, che portano a una pesantezza eccessiva delle code della distribuzione.

E' possibile quindi concludere che, normalità a parte, i residui del modello sono tutto sommato soddisfacenti, e quindi il modello individuato può essere utilizzato per le previsioni del consumo giornaliero di gas nelle tre serie storiche differenziate e destagionalizzate; una volta calcolate le previsioni sulle serie storiche differenziate, è necessario riconvertire il consumo giornaliero di gas previsto in metri cubi, prima ottenendo i valori per la serie non differenziata, e poi applicando ad essi la trasformata seno iperbolico.

La previsione del consumo giornaliero di gas per il punto di riconsegna situato a Caldiero ha dato i seguenti risultati:

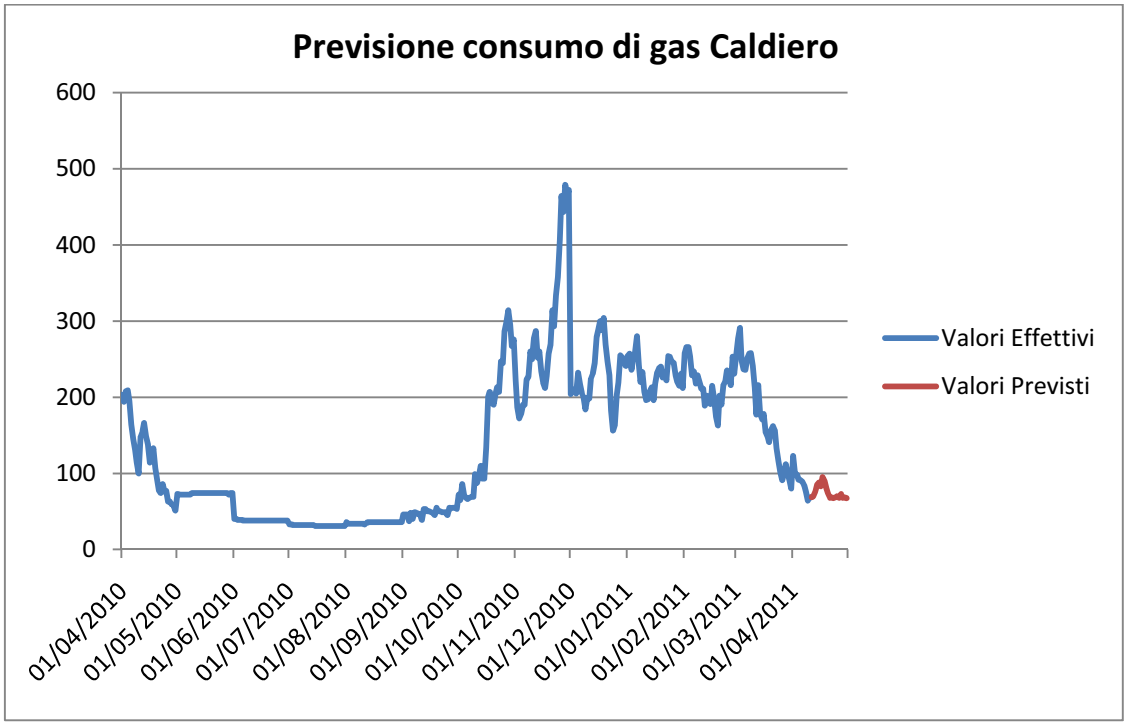



**GRAFICO 4.36: Previsioni consumo di gas Caldiero – *VAR(8)***

| DATA | CONSUMO |
|---|---|
| 11/04/2011 | 69 |
| 12/04/2011 | 70 |
| 13/04/2011 | 76 |
| 14/04/2011 | 85 |
| 15/04/2011 | 88 |
| 16/04/2011 | 83 |
| 17/04/2011 | 95 |
| 18/04/2011 | 90 |
| 19/04/2011 | 80 |
| 20/04/2011 | 74 |

| DATA | CONSUMO |
|---|---|
| 21/04/2011 | 68 |
| 22/04/2011 | 68 |
| 23/04/2011 | 67 |
| 24/04/2011 | 68 |
| 25/04/2011 | 70 |
| 26/04/2011 | 68 |
| 27/04/2011 | 73 |
| 28/04/2011 | 68 |
| 29/04/2011 | 68 |
| 30/04/2011 | 68 |

**TABELLA 4.27: Previsioni consumo di gas Caldiero – *VAR(8)***

La previsione del consumo giornaliero di gas per il punto di riconsegna situato a Castelnuovo del Garda ha dato i seguenti risultati:

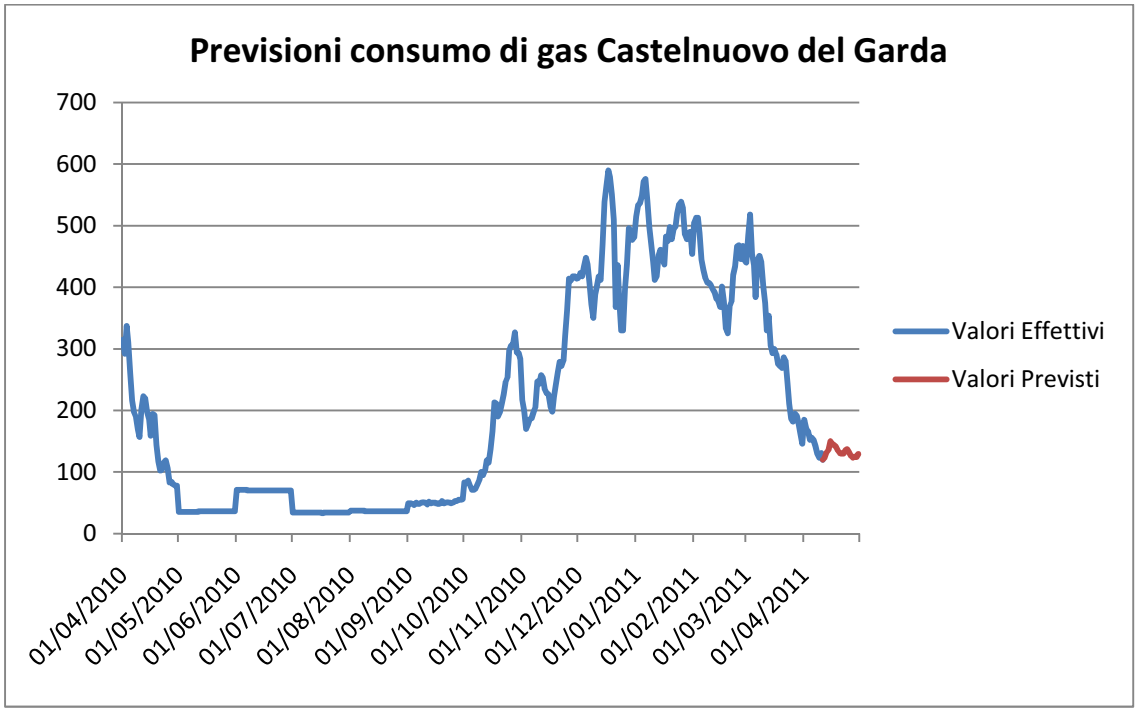



**GRAFICO 4.37: Previsioni consumo di gas Castelnuovo del Garda – *VAR(8)***

| DATA | CONSUMO |
|---|---|
| 11/04/2011 | 120 |
| 12/04/2011 | 123 |
| 13/04/2011 | 132 |
| 14/04/2011 | 136 |
| 15/04/2011 | 150 |
| 16/04/2011 | 146 |
| 17/04/2011 | 144 |
| 18/04/2011 | 141 |
| 19/04/2011 | 135 |
| 20/04/2011 | 131 |

| DATA | CONSUMO |
|---|---|
| 21/04/2011 | 130 |
| 22/04/2011 | 130 |
| 23/04/2011 | 135 |
| 24/04/2011 | 137 |
| 25/04/2011 | 132 |
| 26/04/2011 | 126 |
| 27/04/2011 | 123 |
| 28/04/2011 | 125 |
| 29/04/2011 | 124 |
| 30/04/2011 | 129 |

**TABELLA 4.28: Previsioni consumo di gas Castelnuovo del Garda – *VAR(8)***

Infine, la previsione del consumo giornaliero di gas per il punto di riconsegna situato a Colognola ai Colli ha dato i seguenti risultati:

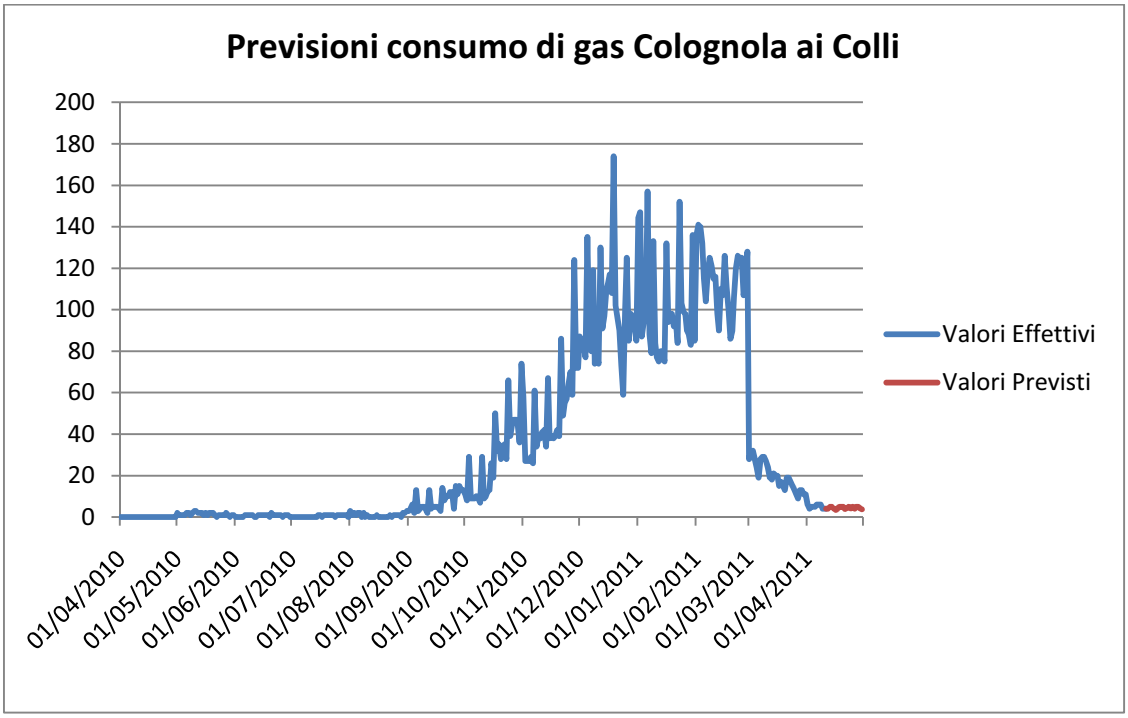



**GRAFICO 4.38: Previsioni consumo di gas Colognola ai Colli – *VAR(8)***

| DATA | CONSUMO |
|---|---|
| 11/04/2011 | 4 |
| 12/04/2011 | 4 |
| 13/04/2011 | 5 |
| 14/04/2011 | 5 |
| 15/04/2011 | 4 |
| 16/04/2011 | 3 |
| 17/04/2011 | 4 |
| 18/04/2011 | 5 |
| 19/04/2011 | 5 |
| 20/04/2011 | 5 |

| DATA | CONSUMO |
|---|---|
| 21/04/2011 | 4 |
| 22/04/2011 | 4 |
| 23/04/2011 | 5 |
| 24/04/2011 | 4 |
| 25/04/2011 | 5 |
| 26/04/2011 | 4 |
| 27/04/2011 | 5 |
| 28/04/2011 | 5 |
| 29/04/2011 | 4 |
| 30/04/2011 | 4 |

**TABELLA 4.29: Previsioni consumo di gas Colognola ai Colli – *VAR(8)***

### *4.3.2 L'analisi strutturale del modello*

Un modello generale *VAR* contiene solitamente molti parametri, che possono risultare di difficile interpretazione, a causa di complesse interazioni e feedback tra le componenti del modello. Pertanto, le proprietà dinamiche di un modello *VAR* si possono spesso sintetizzare attraverso vari tipi di analisi strutturali; le principali sono:

- Analisi di causalità secondo *Granger*: se una variabile, o un gruppo di variabili, $y_1$ è di aiuto nel migliorare le previsioni di un'altra variabile, o gruppo di variabili, $y_2$, allora si dice che $y_1$ causa secondo *Granger* $y_2$.

Nel modello VAR precedentemente stimato, occorre verificare se la temperatura media causa secondo *Granger* le serie storiche del consumo giornaliero di gas:

```
H0: Temperatura do not Granger-cause Caldiero Castelnuovo
Colognola

F-Test = 1.433, df1 = 24, df2 = 1308, p-value = 0.081
```

**TABELLA 4.30: Caldiero causa secondo *Granger* Castelnuovo-Colognola**

A un livello di significatività fissato pari al 5%, si accetta l'ipotesi nulla che la serie storica della temperatura media non causi secondo *Granger* le serie storiche del consumo giornaliero di gas.

- Analisi di causalità istantanea: se una variabile, o un gruppo di variabili, $y_1$ è caratterizzata da correlazione non nulla nei confronti di un'altra variabile, o gruppo di variabili, $y_2$, allora si dice che $y_1$ causa istantaneamente $y_2$.

Nel modello *VAR* precedentemente stimato, occorre verificare la causalità istantanea tra la temperatura media e le serie storiche del consumo giornaliero di gas:

```
H0: No instantaneous causality between: Temperatura and Caldiero
Castelnuovo Colognola

Chi-squared = 30.107, df = 3, p-value = 1.310e-06
```

**TABELLA 4.31: Caldiero causa istantaneamente Castelnuovo-Colognola**

A un livello di significatività fissato pari al 5%, si rifiuta l'ipotesi nulla che la temperatura media non causi istantanemente le serie storiche del consumo giornaliero di gas. Sembra esserci quindi causalità istantanea tra le serie considerate.

- Funzioni di risposta impulsiva (*IRF: Impulse Response Functions*): attraverso l'IRF si esamina la risposta (reazione) nel tempo di una variabile in relazione ad un impulso di un'altra variabile in un sistema dinamico che coinvolge anche altre variabili. Con riferimento al modello *VAR*, si tratta di seguire e misurare l'effetto di uno shock esogeno o innovazione in una delle variabili, su una o più altre variabili. Poiché nel modello precedentemente stimato le componenti del termine d'errore $a_t$ sono tra loro contemporaneamente correlate ($\Sigma$ è non diagonale), è poco probabile che lo shock che interviene su una componente rimanga isolato, anzi è facile, data la correlazione

contemporanea tra componenti, che uno shock in una variabile sia accompagnato da uno shock in un'altra variabile; in questa situazione è preferibile ortogonalizzare gli errori e derivare conseguentemente le funzioni di risposta impulsiva. Con riferimento al modello *VAR* stimato in precedenza, si analizza l'effetto di uno shock sulla temperatura media nei confronti del consumo giornaliero di gas (i valori dello shock sono sulle serie differenziate):

| | Caldiero | Castelnuovo | Colognola |
|---|---|---|---|
| 0 | 0,0000 | 0,0000 | 0,0000 |
| 1 | -0,0195 | -0,0124 | -0,0234 |
| 2 | -0,0019 | -0,0034 | -0,0002 |
| 3 | -0,0052 | -0,0009 | 0,0014 |
| 4 | -0,0010 | 0,0012 | 0,0044 |
| 5 | 0,0049 | 0,0046 | 0,0056 |
| 6 | 0,0104 | 0,0033 | -0,0051 |
| 7 | 0,0120 | 0,0095 | 0,0361 |
| 8 | 0,0147 | -0,0067 | 0,0199 |
| 9 | 0,0040 | 0,0006 | -0,0112 |
| 10 | 0,0017 | 0,0007 | -0,0060 |
| 11 | -0,0016 | -0,0011 | -0,0084 |
| 12 | -0,0080 | -0,0010 | -0,0039 |
| 13 | -0,0117 | -0,0041 | -0,0124 |
| 14 | -0,0095 | 0,0007 | -0,0086 |
| 15 | -0,0081 | -0,0014 | -0,0166 |
| 16 | 0,0003 | 0,0013 | 0,0118 |
| 17 | 0,0009 | 0,0014 | 0,0091 |
| 18 | 0,0029 | 0,0002 | 0,0063 |
| 19 | 0,0056 | 0,0002 | 0,0055 |
| 20 | 0,0073 | -0,0005 | 0,0014 |

**TABELLA 4.32: IRF – Shock sulla componente Temperatura**

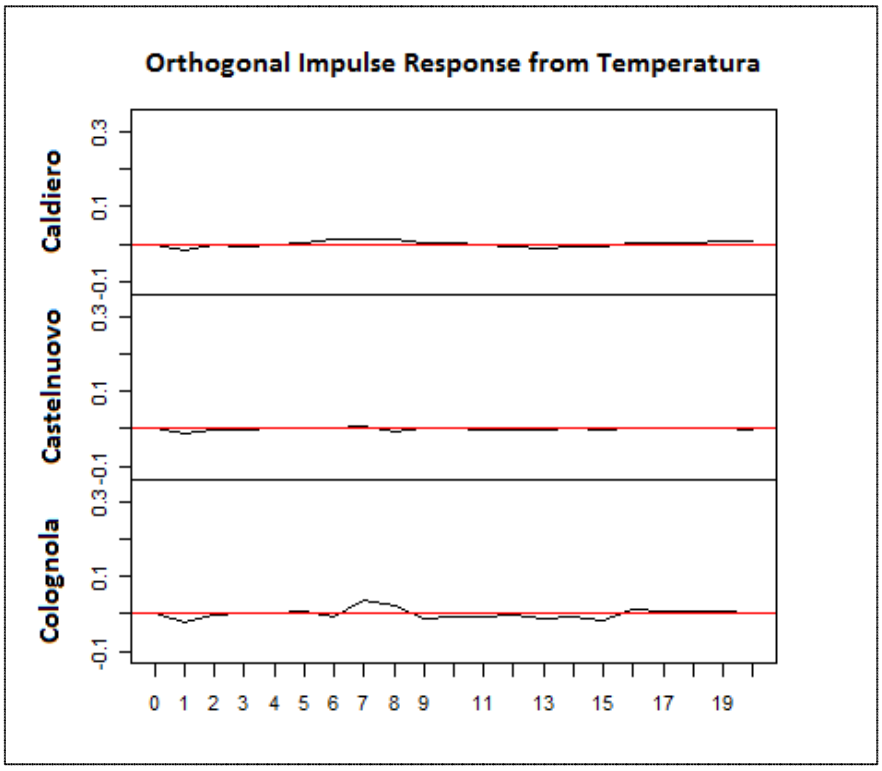



**GRAFICO 4.39: IRF – Shock sulla componente Temperatura**

Uno shock sulla componente temperatura media genera sulle tre componenti relative al consumo giornaliero di gas un impulso, che tende ad oscillare tra valori positivi e negativi (vicini allo zero) nei 20 giorni successivi; inoltre, sembra possibile affermare che l'effetto più importante dello shock sembra evidenziarsi nel punto di riconsegna di Colognola ai Colli.

L'analisi e la previsione del consumo giornaliero di gas attraverso i modelli *VAR* è molto più completa ed esaustiva rispetto all'analisi univariata; innanzitutto, è possibile, tramite un unico modello, analizzare e prevedere contemporaneamente l'andamento di più punti di riconsegna, con un notevole risparmio nell'onere del lavoro.

Inoltre, è possibile studiare le relazioni che intercorrono tra le componenti del modello, tramite l'analisi strutturale: è possibile verificare le relazioni di causalità, istantanea e non (in questo caso la causalità tra la temperatura media e

il consumo giornaliero di gas), e gli effetti di shock o innovazioni su una componente nei confronti di tutte le altre componenti del modello (in questo caso uno shock sulla temperatura media nei confronti del consumo giornaliero di gas).

Infine, per migliorare ulteriormente le analisi, e renderle ancora più complete, si potrebbe studiare il processo da un punto di vista spazio-temporale, e stimare un modello che colga gli effetti principali del tempo e dello spazio, con l'introduzione di una struttura di covarianza che tenga conto anche della collocazione spaziale dei punti di riconsegna.
In questo modo, l'insieme delle osservazioni viene considerato come realizzazione di un processo stocastico a due indici, uno legato al tempo e l'altro allo spazio; si suddivide la variabilità di tale processo in una parte che colga principalmente l'andamento dovuto alle differenze globali, spaziali e temporali (variabilità di grande scala, caratterizzata da processi stocastici non stazionari), e in una parte che invece colga sostanzialmente il "rumore" intorno all'altra componente (variabilità di piccola scala, caratterizzata da un andamento più regolare, stazionario nello spazio e nel tempo). La stima contemporanea di tali processi può avvenire attraverso una formulazione *state-space* del modello indicato, che consenta di ottenere delle stime di massima verosimiglianza attraverso l'applicazione del filtro di *Kalman*.

# CAPITOLO 5: LA VALUTAZIONE DELLE PREVISIONI

Dopo aver analizzato i dati a disposizione, e aver stimato una vasta gamma di modelli, dal *Data Mining* alle Serie Storiche, occorre analizzare i risultati e, soprattutto, le previsioni ottenute, confrontandole attentamente con i valori che poi si sono verificati nella realtà; *Main Consulting*, infatti, ritiene buono un modello che commette un errore di previsione fino a due passi in avanti massimo del 4%-5%. Questa valutazione dell'azienda permette di dare un giudizio più completo e più approfondito alle previsione ottenute.

## 5.1 LA PREVISIONE COL *DATA MINING*

La previsione col *Data Mining*, come visto nel Capitolo 3, si svolge in due fasi: nella prima si classificano i consumi di gas uguali a 0 e quelli maggiori di 0 tramite un modello di tipo *Bagging*, mentre nella seconda, attraverso un'analisi di regressione, si prevedono i valori che la classificazione precedente aveva classificato maggiori di 0, utilizzando un modello *Projection Pursuit*.

Per valutare le previsioni ottenute, si considera il punto di riconsegna situato nel comune di Alessandria, e si valutano le previsione nel periodo che va dall'11 Aprile 2011 al 30 Aprile 2011:

| | *Previsione* | *Reali* | *Errore* | *Errore %* |
|---|---|---|---|---|
| 11/04/2011 | 2.282 | 2.000 | 282 | 14,10% |
| 12/04/2011 | 2.239 | 2.100 | 139 | 6,62% |
| 13/04/2011 | 1.974 | 2.000 | -26 | -1,30% |

| | | | | |
|---|---|---|---|---|
| 14/04/2011 | 2.324 | 2.400 | -76 | -3,17% |
| 15/04/2011 | 2.442 | 2.300 | 142 | 6,17% |
| 16/04/2011 | 0 | 0 | 0 | 0,00% |
| 17/04/2011 | 0 | 261 | -261 | -100,00% |
| 18/04/2011 | 2.591 | 2.200 | 391 | 17,77% |
| 19/04/2011 | 2.506 | 2.345 | 161 | 6,87% |
| 20/04/2011 | 2.450 | 2.243 | 207 | 9,23% |
| 21/04/2011 | 2.418 | 2.145 | 273 | 12,73% |
| 22/04/2011 | 0 | 0 | 0 | 0,00% |
| 23/04/2011 | 0 | 0 | 0 | 0,00% |
| 24/04/2011 | 0 | 270 | -270 | -100,00% |
| 25/04/2011 | 0 | 243 | -243 | -100,00% |
| 26/04/2011 | 2.329 | 2.214 | 115 | 5,19% |
| 27/04/2011 | 2.191 | 2.214 | -23 | -1,04% |
| 28/04/2011 | 2.331 | 2.214 | 117 | 5,28% |
| 29/04/2011 | 2.256 | 2.214 | 42 | 1,90% |
| 30/04/2011 | 2.141 | 2.214 | -73 | -3,30% |

**TABELLA 5.1: Confronto valori reali e valori previsti per il punto di Alessandria**

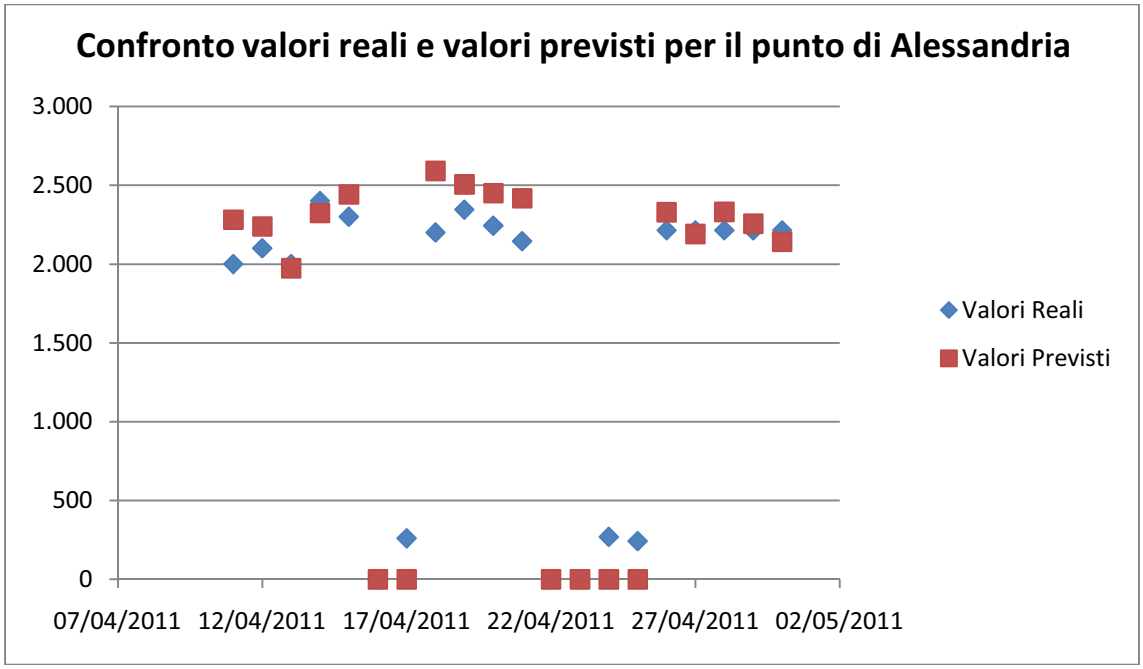



**GRAFICO 5.1: Confronto valori reali e valori previsti per il punto di Alessandria**

Dal Grafico 5.1 è possibile osservare come il modello tende a sovrastimare il valore reale; questo potrebbe essere dovuto alle trasformazioni dei dati, che sono state svolte durante la procedura di stima. L'errore, inoltre, nei primi due giorni dell'orizzonte di previsione, è molto elevato, e questo è probabilmente dovuto al fatto che i modelli del *Data Mining* applicati in questo elaborato non tengono conto della correlazione seriale nel consumo giornaliero di gas.

## 5.2 LA PREVISIONE CON LE SERIE STORICHE

La previsione con le Serie Storiche, come visto nel Capitolo 4, è stata svolta per tenere in considerazione la correlazione seriale del consumo giornaliero di gas. E' ragionevole pensare, infatti, che il consumo di gas in un giorno sia condizionato dal consumo di gas effettuato il giorno precedente. In questo ambito di analisi, si sono stimati modelli di analisi univariata (*SARIMA*), modelli a funzione di trasferimento e modelli di analisi multivariata (*VAR*).

### *5.2.1 I modelli di analisi univariata (SARIMA)*

I modelli di analisi univariata sono stati applicati a tre punti di riconsegna, uno per ciascuna destinazione d'uso: Argelato (destinazione d'uso esclusivamente civile), Tocco da Casauria (destinazione d'uso civile e industriale, Cartiera di Ferrara (destinazione d'uso esclusivamente industriale).

Per il punto di riconsegna ad uso esclusivamente civile, situato nel comune di Argelato, è stato stimato un modello *SARIMA* $(1,1,2)x(1,1,0)_7$. I valori previsti dal modello, per il periodo che va dall'11 Aprile 2011 al 30 Aprile 2011, sono stati i seguenti:

| | *Previsione* | *Reali* | *Errore* | *Errore %* |
|---|---|---|---|---|
| 11/04/2011 | 335 | 347 | -12 | -3,46% |
| 12/04/2011 | 290 | 359 | -69 | -19,22% |
| 13/04/2011 | 328 | 390 | -62 | -15,90% |
| 14/04/2011 | 286 | 495 | -209 | -42,22% |
| 15/04/2011 | 284 | 625 | -341 | -54,56% |
| 16/04/2011 | 231 | 571 | -340 | -59,54% |
| 17/04/2011 | 178 | 457 | -279 | -61,05% |
| 18/04/2011 | 277 | 533 | -256 | -48,03% |
| 19/04/2011 | 240 | 542 | -302 | -55,72% |
| 20/04/2011 | 275 | 435 | -160 | -36,78% |
| 21/04/2011 | 241 | 391 | -150 | -38,36% |
| 22/04/2011 | 237 | 371 | -134 | -36,12% |
| 23/04/2011 | 194 | 341 | -147 | -43,11% |
| 24/04/2011 | 149 | 196 | -47 | -23,98% |
| 25/04/2011 | 233 | 186 | 47 | 25,27% |
| 26/04/2011 | 202 | 404 | -202 | -50,00% |
| 27/04/2011 | 231 | 420 | -189 | -45,00% |
| 28/04/2011 | 203 | 427 | -224 | -52,46% |
| 29/04/2011 | 200 | 413 | -213 | -51,57% |
| 30/04/2011 | 164 | 359 | -195 | -54,32% |

**TABELLA 5.2: Confronto valori reali e valori previsti per il punto di Argelato**

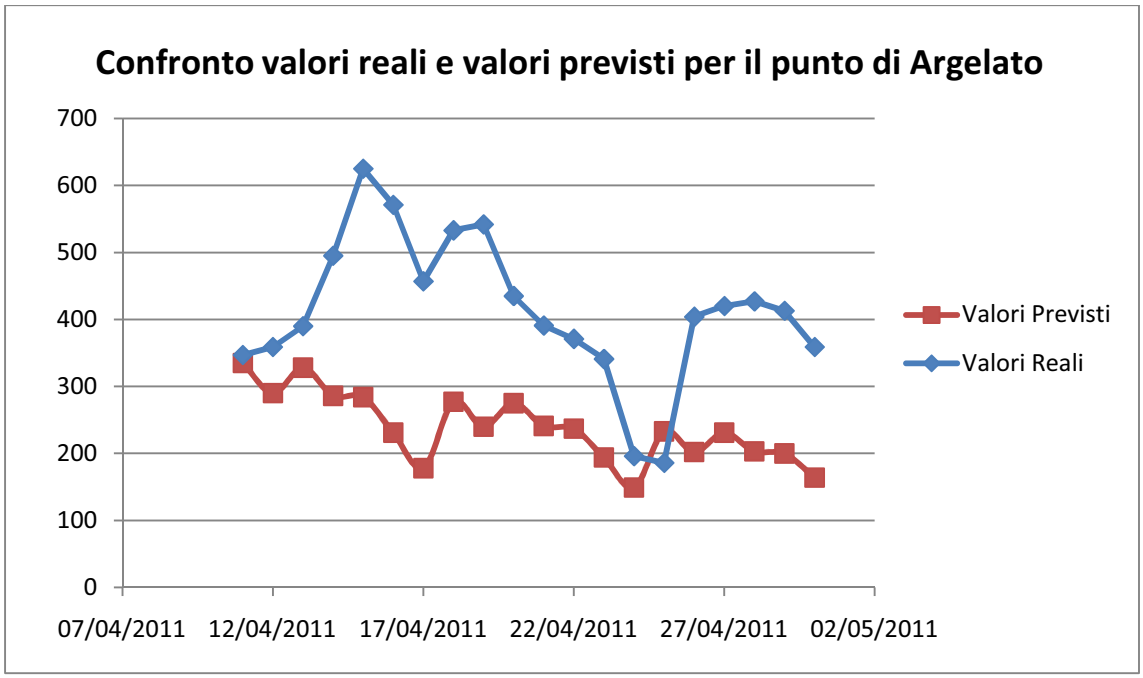



**GRAFICO 5.2: Confronto valori reali e valori previsti per il punto di Argelato**

Dalla Tabella 5.2 e dal Grafico 5.2 è possibile osservare che l'errore nel primo giorno dell'orizzonte di previsione è abbastanza contenuto, ma poi tende ad aumentare e a rimanere piuttosto elevato; il modello, infatti, sembra cogliere abbastanza bene l'andamento dei valori reali, ma tende a sottostimare regolarmente quello che poi sarà il valore che si verifica nella realtà. Questo può essere dovuto all'assenza di ulteriori variabili esplicative nel modello, quali temperatura e altri fattori atmosferici, che potrebbero aver causato l'aumento nel consumo giornaliero di gas, che non è stato colto dal modello stimato.

Per il punto di riconsegna ad uso civile e industriale, situato nel comune di Tocco da Casauria, è stato stimato un modello *ARIMA (3,1,1)*. I valori previsti dal modello, per il periodo che va dall'11 Aprile 2011 al 30 Aprile 2011, sono stati i seguenti:

| | *Previsione* | *Reali* | *Errore* | *Errore %* |
|---|---|---|---|---|
| 11/04/2011 | 1.847 | 1.920 | -73 | -3,80% |
| 12/04/2011 | 1.858 | 1.504 | 354 | 23,54% |
| 13/04/2011 | 1.870 | 2.160 | -290 | -13,43% |
| 14/04/2011 | 1.813 | 2.637 | -824 | -31,25% |
| 15/04/2011 | 1.892 | 3.600 | -1.708 | -47,44% |
| 16/04/2011 | 1.811 | 3.532 | -1721 | -48,73% |
| 17/04/2011 | 1.885 | 4.506 | -2621 | -58,17% |
| 18/04/2011 | 1.823 | 3.476 | -1.653 | -47,55% |
| 19/04/2011 | 1.872 | 2.914 | -1.042 | -35,76% |
| 20/04/2011 | 1.834 | 2.477 | -643 | -25,96% |
| 21/04/2011 | 1.861 | 2.191 | -330 | -15,06% |
| 22/04/2011 | 1.843 | 2.255 | -412 | -18,27% |
| 23/04/2011 | 1.855 | 2.119 | -264 | -12,46% |
| 24/04/2011 | 1.762 | 1.107 | 655 | 59,17% |
| 25/04/2011 | 1.852 | 2.559 | -707 | -27,63% |
| 26/04/2011 | 1.850 | 3.050 | -1.200 | -39,34% |
| 27/04/2011 | 1.850 | 3.041 | -1.191 | -39,16% |
| 28/04/2011 | 1.850 | 2.954 | -1.104 | -37,37% |
| 29/04/2011 | 1.850 | 2.103 | -253 | -12,03% |
| 30/04/2011 | 1.851 | 3.241 | -1.390 | -42,89% |

**TABELLA 5.3: Confronto valori reali e valori previsti per il punto di Tocco da Casauria**

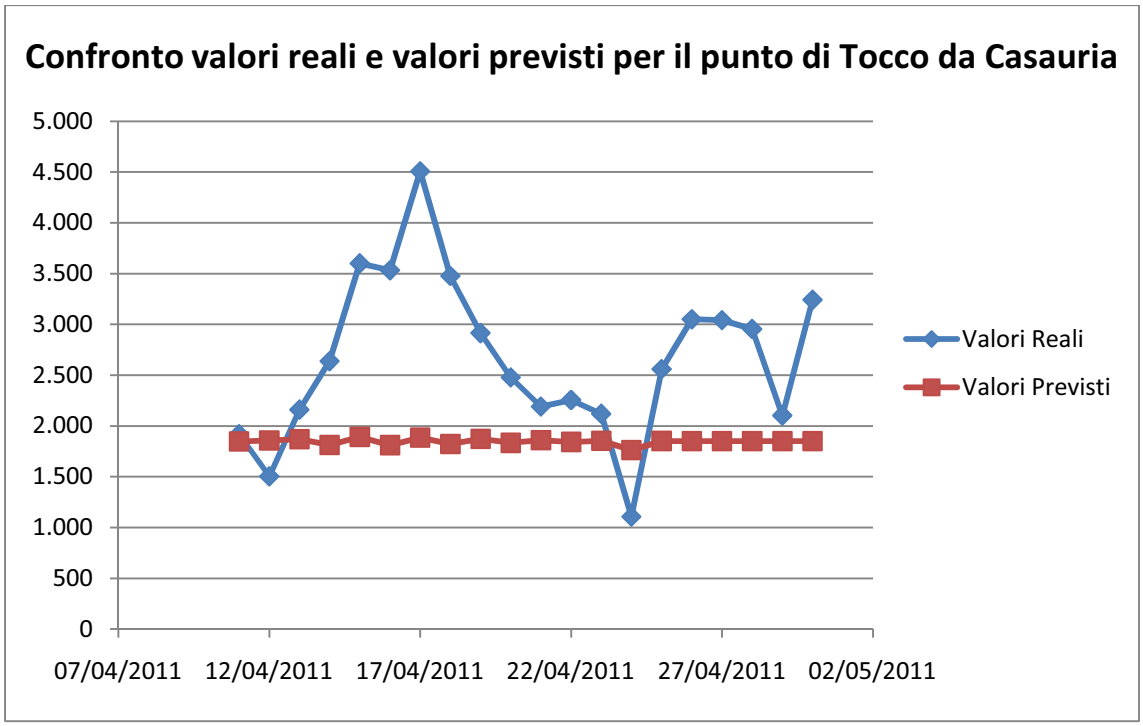



**GRAFICO 5.3: Confronto valori reali e valori previsti per il punto di Tocco da Casauria**

Dalla Tabella 5.3 è possibile osservare che l'errore nel primo giorno dell'orizzonte di previsione è abbastanza contenuto, ma poi tende ad aumentare e a rimanere piuttosto elevato; il modello, inoltre, non riesce a cogliere nemmeno l'andamento delle previsioni, in quanto viene prevista una serie di valori pressoché costante, mentre, come si nota dal Grafico 5.3, l'andamento dei valori reali risulta molto più fluttuante.

Per il punto di riconsegna ad uso esclusivamente industriale, situato presso la Cartiera di Ferrara, è stato stimato un modello *AR(1)*. I valori previsti dal modello, per il periodo che va dall'11 Aprile 2011 al 30 Aprile 2011, sono stati i seguenti:

| | *Previsione* | *Reali* | *Errore* | *Errore %* |
|---|---|---|---|---|
| 11/04/2011 | 34.425 | 34.735 | -310 | -0,89% |
| 12/04/2011 | 31.677 | 33.806 | -2.129 | -6,30% |
| 13/04/2011 | 33.630 | 31.487 | 2.143 | 6,81% |
| 14/04/2011 | 32.766 | 32.747 | 19 | 0,06% |
| 15/04/2011 | 34.902 | 30.294 | 4.608 | 15,21% |
| 16/04/2011 | 31.539 | 33.206 | -1667 | -5,02% |
| 17/04/2011 | 38.219 | 33.753 | 4466 | 13,23% |
| 18/04/2011 | 21.593 | 34.712 | -13.119 | -37,79% |
| 19/04/2011 | 18.423 | 35.899 | -17.476 | -48,68% |
| 20/04/2011 | 18.287 | 34.227 | -15.940 | -46,57% |
| 21/04/2011 | 18.162 | 33.321 | -15.159 | -45,49% |
| 22/04/2011 | 18.046 | 34.199 | -16.153 | -47,23% |
| 23/04/2011 | 17.939 | 33.081 | -15142 | -45,77% |
| 24/04/2011 | 17.841 | 31.634 | -13793 | -43,60% |
| 25/04/2011 | 17.749 | 37.174 | -19425 | -52,25% |
| 26/04/2011 | 17.665 | 34.062 | -16.397 | -48,14% |
| 27/04/2011 | 17.587 | 34.884 | -17.297 | -49,58% |
| 28/04/2011 | 17.515 | 35.800 | -18.285 | -51,08% |
| 29/04/2011 | 17.448 | 35.472 | -18.024 | -50,81% |
| 30/04/2011 | 17.386 | 34.359 | -16.973 | -49,40% |

**TABELLA 5.4: Confronto valori reali e valori previsti per il punto presso la Cartiera di Ferrara**

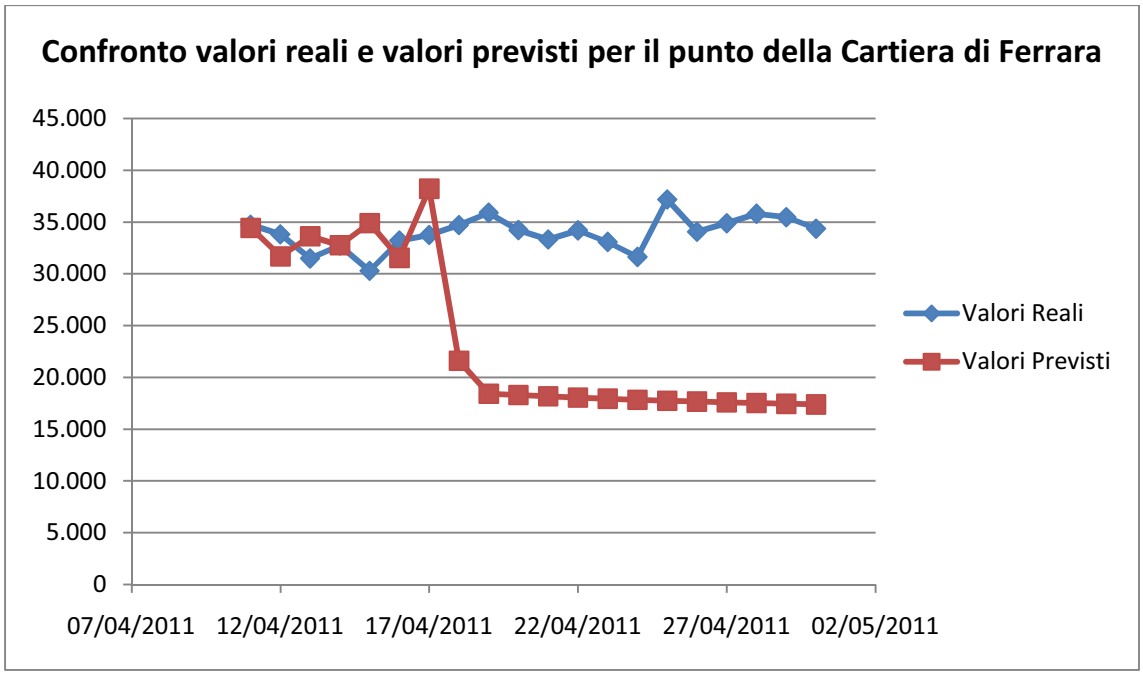



**GRAFICO 5.4: Confronto valori reali e valori previsti per il punto presso la Cartiera di Ferrara**

Dalla Tabella 5.4 è possibile osservare che l'errore nella prima settimana dell'orizzonte di previsione è abbastanza contenuto, mentre poi tende ad aumentare e a rimanere piuttosto elevato; oltre la prima settimana, infatti, il modello non riesce a cogliere nemmeno l'andamento delle previsioni, in quanto viene prevista una serie di valori pressoché costante e inferiore ai valori reali, come si nota dal Grafico 5.4.

### *5.2.2 I modelli a funzione di trasferimento*

Dopo aver stimato i modelli per l'analisi univariata delle serie storiche, si è passati alla stima di un modello a funzione di trasferimento per ciascuno dei tre punti di riconsegna precedentemente considerati (Argelato, Tocco da Casauria, Cartiera di Ferrara); in particolare, sono stati analizzati i legami fra le variabili temperatura media e consumo giornaliero di gas, al fine di determinare se la

temperatura media registrata il giorno della rilevazione del consumo di gas rappresenta un input per il sistema, e, in caso affermativo, di valutare gli istanti temporali in cui si verificano gli effetti e la loro durata.

Per il punto di riconsegna ad uso esclusivamente civile, situato presso il comune di Argelato, i valori previsti dal modello a funzione di trasferimento, per il periodo che va dall'11 Aprile 2011 al 30 Aprile 2011, sono stati i seguenti:

| | *Previsione* | *Reali* | *Errore* | *Errore %* |
|---|---|---|---|---|
| 11/04/2011 | 364 | 347 | 17 | 4,90% |
| 12/04/2011 | 295 | 359 | -64 | -17,83% |
| 13/04/2011 | 373 | 390 | -17 | -4,36% |
| 14/04/2011 | 404 | 495 | -91 | -18,38% |
| 15/04/2011 | 497 | 625 | -128 | -20,48% |
| 16/04/2011 | 487 | 571 | -84 | -14,71% |
| 17/04/2011 | 391 | 457 | -66 | -14,44% |
| 18/04/2011 | 632 | 533 | 99 | 18,57% |
| 19/04/2011 | 506 | 542 | -36 | -6,64% |
| 20/04/2011 | 542 | 435 | 107 | 24,60% |
| 21/04/2011 | 468 | 391 | 77 | 19,69% |
| 22/04/2011 | 468 | 371 | 97 | 26,15% |
| 23/04/2011 | 405 | 341 | 64 | 18,77% |
| 24/04/2011 | 300 | 196 | 104 | 53,06% |
| 25/04/2011 | 482 | 186 | 296 | 159,14% |
| 26/04/2011 | 402 | 404 | -2 | -0,50% |
| 27/04/2011 | 455 | 420 | 35 | 8,33% |
| 28/04/2011 | 440 | 427 | 13 | 3,04% |
| 29/04/2011 | 470 | 413 | 57 | 13,80% |
| 30/04/2011 | 426 | 359 | 67 | 18,66% |

**TABELLA 5.5: Confronto valori reali e valori previsti per il punto di Argelato**

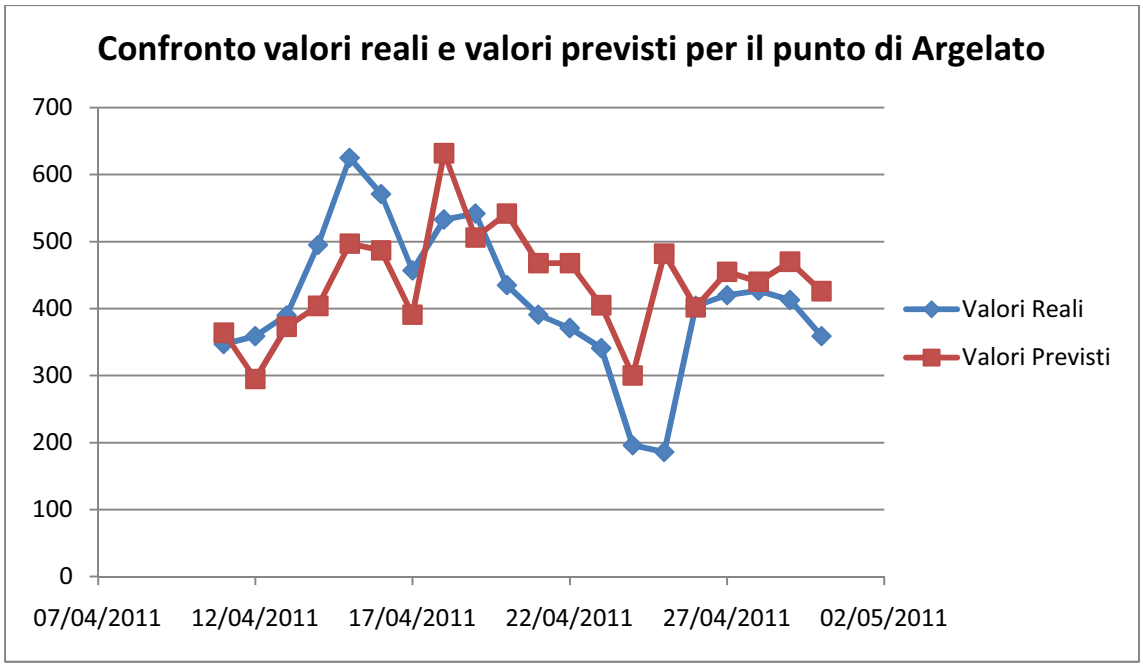



**GRAFICO 5.5: Confronto valori reali e valori previsti per il punto di Argelato**

Dalla Tabella 5.5 è possibile osservare che l'errore nel primo giorno dell'orizzonte di previsione è abbastanza contenuto, mentre poi tende ad aumentare; dal grafico 5.5, si nota come il modello a funzione di trasferimento, a differenza del modello univariato, oltre a cogliere abbastanza bene l'andamento dei valori reali, tende a cogliere con buona approssimazione anche il valore che poi effettivamente si verifica nella realtà. Questo è quasi sicuramente dovuto all'inserimento nel modello di una serie di input (la temperatura media), che, come evidenziato, aiuta a migliorare sensibilmente le previsioni per il punto di riconsegna.

Per il punto di riconsegna ad uso civile e industriale, situato presso il comune di Tocco da Casauria, la temperatura media non rappresenta distintamente l'input del sistema, e per questo motivo la stima del modello a funzione di trasferimento per il consumo di gas non si può calcolare, in quanto risultano violate le ipotesi

di base per la costruzione di tale classe di modelli, poiché il consumo di gas non può essere assunto come output univoco del sistema.
Per il punto di riconsegna ad uso esclusivamente industriale, situato presso la Cartiera di Ferrara, non è possibile attribuire un ruolo ben definito alle variabili temperatura media e consumo di gas; sebbene la logica voglia che la temperatura media sia l'input e il consumo di gas l'output, sembra non esserci alcuna relazione tra il consumo giornaliero di gas e la temperatura media registrata; questo è probabilmente dovuto al fatto che le industrie necessitano di un quantitativo di gas pressoché costante durante tutto l'Anno Termico, indipendentemente dalle condizioni climatiche, al fine di svolgere efficientemente e con continuità i processi industriali. In base a queste osservazioni, si può quindi concludere che non è possibile condurre uno studio basato sulle funzioni di trasferimento, visto che risultano violate le ipotesi di base per la costruzione di tale classe di modelli.

### *5.2.3 I modelli di analisi multivariata (VAR)*

I modelli di analisi multivariata applicati sono modelli di tipo *VAR*; un modello di questo tipo è stato applicato a tre punti di riconsegna, situati nella provincia di Verona: Caldiero, Castelnuovo del Garda e Colognola ai Colli; nel modello, inoltre, è stata inserita la serie storica della temperatura media registrata nella provincia di Verona, al fine di valutare l'impatto di questa variabile sul consumo giornaliero di gas. Per queste variabili è stato stimato un modello *VAR(8)*; lo scopo è quello di prevedere esclusivamente il consumo giornaliero di gas, in quanto le temperature previste saranno acquistate dalle agenzie meteorologiche.

Per il punto di riconsegna, situato nel comune di Caldiero, i valori previsti dal modello, per il periodo che va dall'11 Aprile 2011 al 30 Aprile 2011, sono stati i seguenti:

| | *Previsione* | *Reali* | *Errore* | *Errore %* |
|---|---|---|---|---|
| 11/04/2011 | 69 | 70 | -1 | -1,43% |
| 12/04/2011 | 70 | 74 | -4 | -5,41% |
| 13/04/2011 | 76 | 80 | -4 | -5,00% |
| 14/04/2011 | 85 | 91 | -6 | -6,59% |
| 15/04/2011 | 88 | 97 | -9 | -9,28% |
| 16/04/2011 | 83 | 86 | -3 | -3,49% |
| 17/04/2011 | 95 | 110 | -15 | -13,64% |
| 18/04/2011 | 90 | 80 | 10 | 12,50% |
| 19/04/2011 | 80 | 75 | 5 | 6,67% |
| 20/04/2011 | 74 | 71 | 3 | 4,23% |
| 21/04/2011 | 68 | 71 | -3 | -4,23% |
| 22/04/2011 | 68 | 66 | 2 | 3,03% |
| 23/04/2011 | 67 | 65 | 2 | 3,08% |
| 24/04/2011 | 68 | 78 | -10 | -12,82% |
| 25/04/2011 | 70 | 73 | -3 | -4,11% |
| 26/04/2011 | 68 | 66 | 2 | 3,03% |
| 27/04/2011 | 73 | 66 | 7 | 10,61% |
| 28/04/2011 | 68 | 70 | -2 | -2,86% |
| 29/04/2011 | 68 | 67 | 1 | 1,49% |
| 30/04/2011 | 68 | 65 | 3 | 4,62% |

**TABELLA 5.6: Confronto valori reali e valori previsti per il punto di Caldiero**

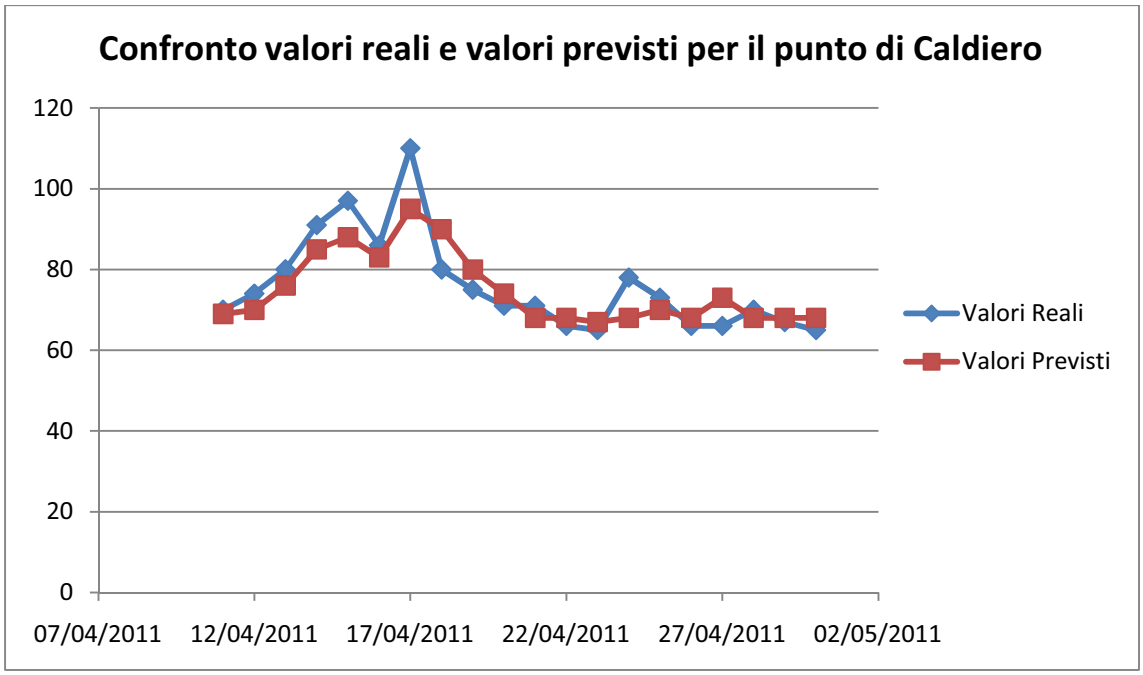



**GRAFICO 5.6: Confronto valori reali e valori previsti per il punto di Caldiero**

Dalla Tabella 5.6 è possibile osservare che l'errore di previsione è sempre abbastanza contenuto; dal grafico 5.6, inoltre, si nota come il modello sembra cogliere abbastanza bene l'andamento dei valori reali, e, con buona approssimazione, il valore che poi effettivamente si verifica nella realtà.

Per il punto di riconsegna, situato nel comune di Castelnuovo del Garda, i valori previsti dal modello, per il periodo che va dall'11 Aprile 2011 al 30 Aprile 2011, sono stati i seguenti:

| | *Previsione* | *Reali* | *Errore* | *Errore %* |
|---|---|---|---|---|
| 11/04/2011 | 120 | 132 | -12 | -9,09% |
| 12/04/2011 | 123 | 118 | 5 | 4,24% |
| 13/04/2011 | 132 | 116 | 16 | 13,79% |
| 14/04/2011 | 136 | 154 | -18 | -11,69% |

| | | | | |
|---|---|---|---|---|
| 15/04/2011 | 150 | 169 | -19 | -11,24% |
| 16/04/2011 | 146 | 161 | -15 | -9,32% |
| 17/04/2011 | 144 | 161 | -17 | -10,56% |
| 18/04/2011 | 141 | 140 | 1 | 0,71% |
| 19/04/2011 | 135 | 134 | 1 | 0,75% |
| 20/04/2011 | 131 | 126 | 5 | 3,97% |
| 21/04/2011 | 130 | 126 | 4 | 3,17% |
| 22/04/2011 | 130 | 125 | 5 | 4,00% |
| 23/04/2011 | 135 | 144 | -9 | -6,25% |
| 24/04/2011 | 137 | 146 | -9 | -6,16% |
| 25/04/2011 | 132 | 136 | -4 | -2,94% |
| 26/04/2011 | 126 | 118 | 8 | 6,78% |
| 27/04/2011 | 123 | 117 | 6 | 5,13% |
| 28/04/2011 | 125 | 129 | -4 | -3,10% |
| 29/04/2011 | 124 | 127 | -3 | -2,36% |
| 30/04/2011 | 129 | 133 | -4 | -3,01% |

**TABELLA 5.7: Confronto valori reali e valori previsti per il punto di Castelnuovo**

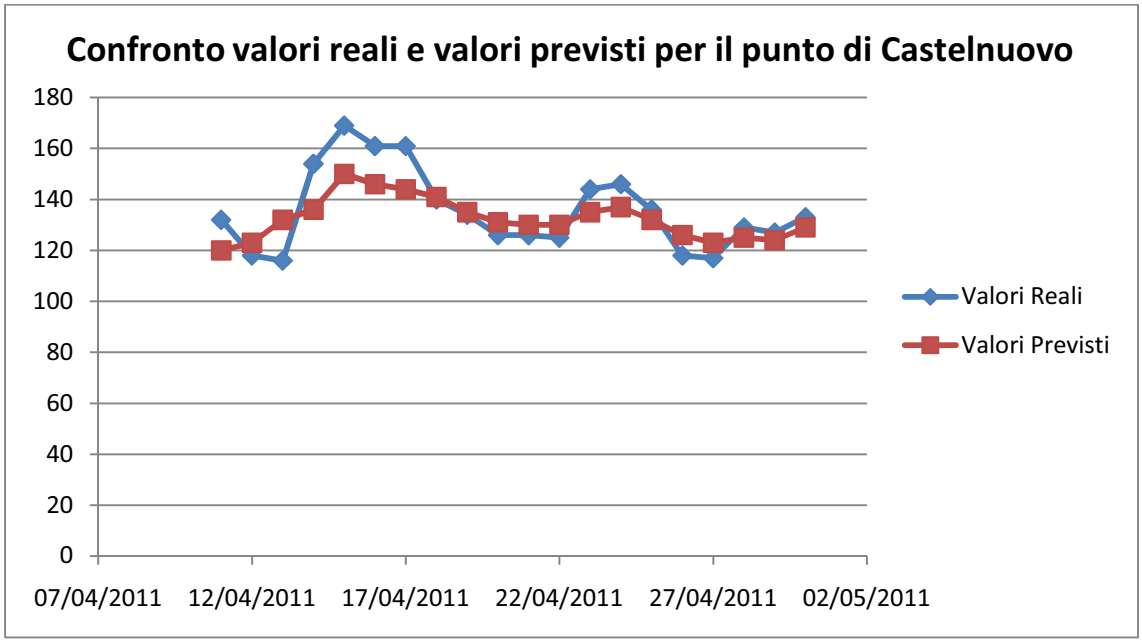



**GRAFICO 5.7: Confronto valori reali e valori previsti per il punto di Castelnuovo**

Dalla Tabella 5.7 è possibile osservare che l'errore di previsione è sempre abbastanza contenuto; dal grafico 5.7, inoltre, si nota come il modello sembra cogliere abbastanza bene l'andamento dei valori reali, e, con buona approssimazione, il valore che poi effettivamente si verifica nella realtà.

Per il punto di riconsegna, situato nel comune di Colognola ai Colli, i valori previsti dal modello, per il periodo che va dall'11 Aprile 2011 al 30 Aprile 2011, sono stati i seguenti:

| | *Previsione* | *Reali* | *Errore* | *Errore %* |
|---|---|---|---|---|
| 11/04/2011 | 4 | 5 | -1 | -20,00% |
| 12/04/2011 | 4 | 6 | -2 | -33,33% |
| 13/04/2011 | 5 | 6 | -1 | -16,67% |
| 14/04/2011 | 5 | 6 | -1 | -16,67% |
| 15/04/2011 | 4 | 6 | -2 | -33,33% |
| 16/04/2011 | 3 | 5 | -2 | -40,00% |
| 17/04/2011 | 4 | 5 | -1 | -20,00% |
| 18/04/2011 | 5 | 5 | 0 | 0,00% |
| 19/04/2011 | 5 | 5 | 0 | 0,00% |
| 20/04/2011 | 5 | 5 | 0 | 0,00% |
| 21/04/2011 | 4 | 4 | 0 | 0,00% |
| 22/04/2011 | 4 | 4 | 0 | 0,00% |
| 23/04/2011 | 5 | 4 | 1 | 25,00% |
| 24/04/2011 | 4 | 4 | 0 | 0,00% |
| 25/04/2011 | 5 | 4 | 1 | 25,00% |
| 26/04/2011 | 4 | 5 | -1 | -20,00% |
| 27/04/2011 | 5 | 5 | 0 | 0,00% |
| 28/04/2011 | 5 | 5 | 0 | 0,00% |
| 29/04/2011 | 4 | 5 | -1 | -20,00% |
| 30/04/2011 | 4 | 4 | 0 | 0,00% |

**TABELLA 5.8: Confronto valori reali e valori previsti per il punto di Colognola**

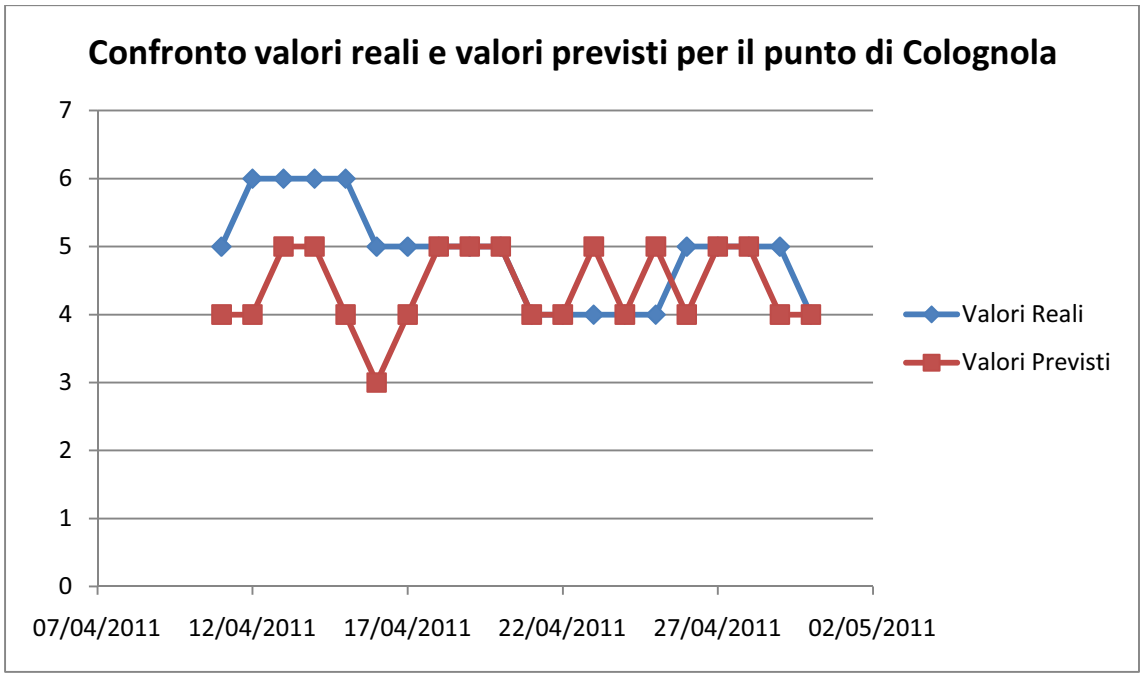



**GRAFICO 5.8: Confronto valori reali e valori previsti per il punto di Colognola**

Dalla Tabella 5.8 è difficile cogliere l'importanza dell'errore di previsione, in quanto il consumo giornaliero di gas presso il punto di riconsegna situato nel comune di Colognola ai Colli risulta molto basso, e quindi l'errore di previsione di un solo metro cubo rappresenta una percentuale d'errore abbastanza elevata; lo stesso discorso vale per l'andamento delle previsioni nel Grafico 5.8.
Intuitivamente, tuttavia, si potrebbe affermare che, tutto sommato, non vi sono grandissimi errori di previsione nel modello e che, quindi, le previsioni sono sostanzialmente accettabili.

Dopo aver analizzato le previsioni del modello *VAR*, è possibile svolgere un'analisi strutturale del modello. Da questa analisi è emerso che vi è causalità istantanea tra la temperatura media e il consumo giornaliero di gas in ciascun punto di riconsegna, e che uno shock sulla temperatura media genera effetti significativi nel consumo giornaliero di gas, soprattutto in quello presso il comune di Colognola ai Colli.

## 5.3 CONCLUSIONI

Concludendo, è quindi possibile affermare che i modelli *VAR*:

- diminuiscono sensibilmente l'onere del lavoro, in quanto con un solo modello è possibile analizzare l'andamento di più punti di riconsegna, con le rispettive temperature medie e qualsiasi altra informazione atmosferica;

- sembrano essere quelli che forniscono le migliori previsioni del consumo giornaliero di gas;

- consentono di svolgere un'analisi strutturale: si può infatti verificare l'eventuale causalità (secondo *Granger* e istantanea) tra la temperatura media e il consumo di gas, e l'effetto di uno shock sulla temperatura nei confronti del consumo giornaliero di gas.

I modelli *VAR* sembrano essere, quindi, la famiglia di modelli più adeguata per la previsione del consumo giornaliero di gas. La stima dei modelli *VAR*, tuttavia, è possibile solo se le serie storiche che lo caratterizzano hanno la stessa lunghezza; in caso non sia così, i modelli utilizzabili per ottenere delle previsioni sono quelli per l'analisi univariata delle serie storiche o, se si vuole un unico modello per tutti i punti di riconsegna, i modelli del *Data Mining*.

Come affermato nel Capitolo 3, tuttavia, i modelli del *Data Mining* non tengono conto della correlazione seriale nel consumo giornaliero di gas. Per tenere conto di questo aspetto, si potrebbe pensare ad un modello di tipo additivo (*GAM*) che abbia la seguente forma:

$$y = f(t) + g(s_1, s_2) + \cdots + \varepsilon$$

dove:

- $f(t)$ è una funzione dall'andamento sufficientemente regolare del tempo, per la quale possono essere utilizzati il *loess* o le *spline di lisciamento* come stimatori non parametrici;

- $g(s_1, s_2)$ è una funzione sufficientemente regolare delle coordinate spaziali del punto di riconsegna (latitudine e longitudine), per la quale possono essere utilizzati il *loess* o le *spline di lisciamento* come stimatori non parametrici;

- $\varepsilon$ è un termine d'errore che contiene la variabilità e l'autocorrelazione (temporale e spaziale), non colta dal modello.

Un modello di questo tipo risulta molto più semplice e molto più intuitivo di un modello *VAR*, e, anche attraverso l'inserimento di altre variabili esplicative, permette di effettuare una previsione, almeno di prima approssimazione, per tutti i punti di riconsegna presi in considerazione, indipendentemente dal loro periodo di osservazione.

# APPENDICE

## APPENDICE 1: ALTRI FATTORI ATMOSFERICI

- **<u>Punto di rugiada</u>**: temperatura alla quale il vapore acqueo presente nell'aria, dopo essersi raffreddato a pressione costante, condensa; in genere ciò avviene in prossimità del suolo.

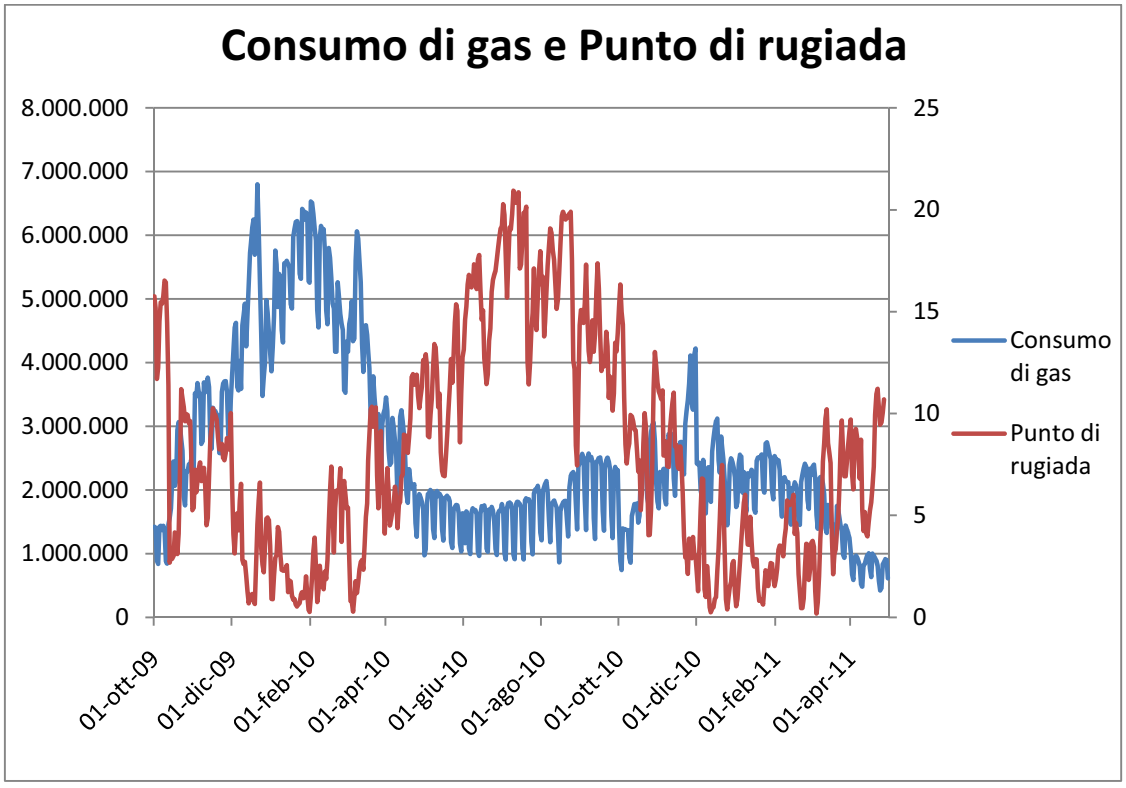



**GRAFICO 1: Consumo medio di gas e punto di rugiada**

Il valore del Punto di Rugiada è strettamente collegato con il valore della temperatura; pertanto anche dal Grafico 1 si nota come la serie storica del consumo giornaliero complessivo di gas sembri essere correlata negativamente

con la seria storica del punto di rugiada medio: all'aumentare del punto di rugiada medio registrato il consumo giornaliero complessivo di gas diminuisce.

- **Umidità**: quantità d'acqua presente nell'aria sotto forma di vapore. Si misura con l'igrometro o lo psicrometro, e si riferisce al rapporto tra il peso del vapore acqueo contenuto in una determinata porzione d'aria e quello che in teoria la stessa porzione potrebbe acquisire senza saturarsi nelle stesse condizioni di temperature e pressione.

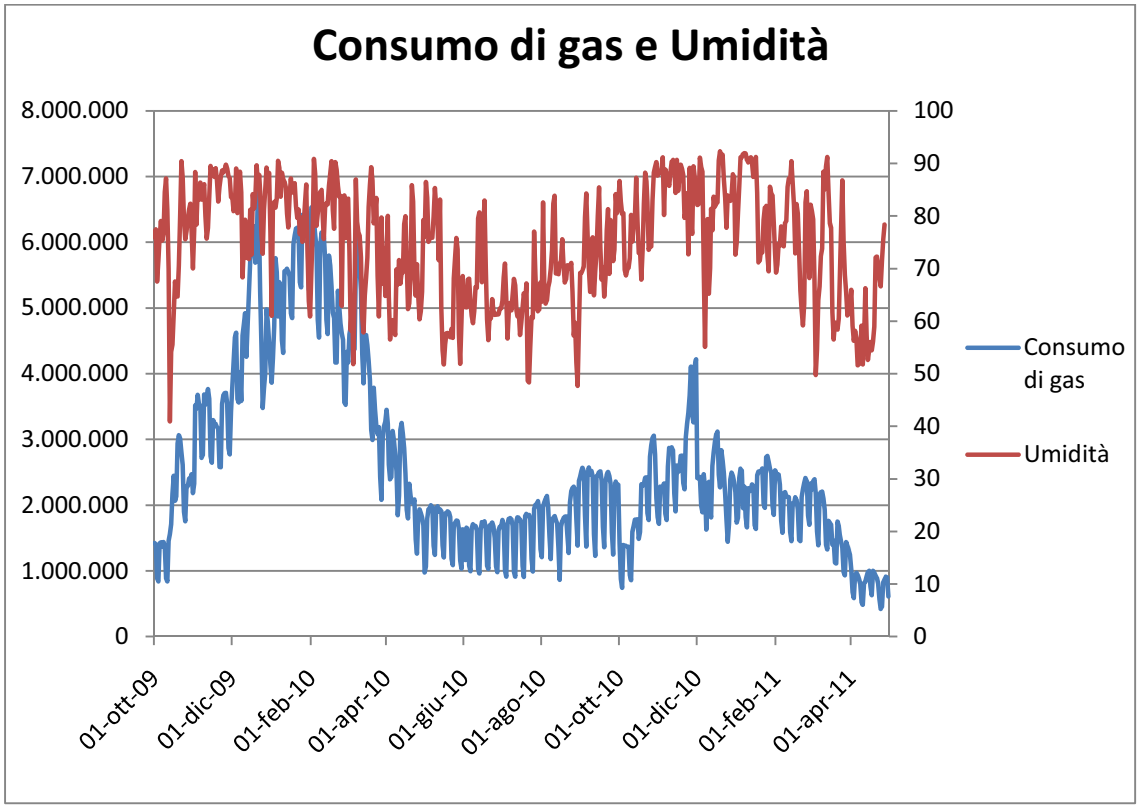



**GRAFICO 2: Consumo medio di gas e umidità**

Dal Grafico 2 si nota come la serie storica del consumo giornaliero complessivo di gas non sembri essere correlata con la serie storica dell'umidità media

registrata; l'impressione, quindi, è che non sembra esserci alcuna relazione tra le due variabili, ma tuttavia non è il caso di escludere una correlazione a priori.

- **Visibilità**: espressa in km.

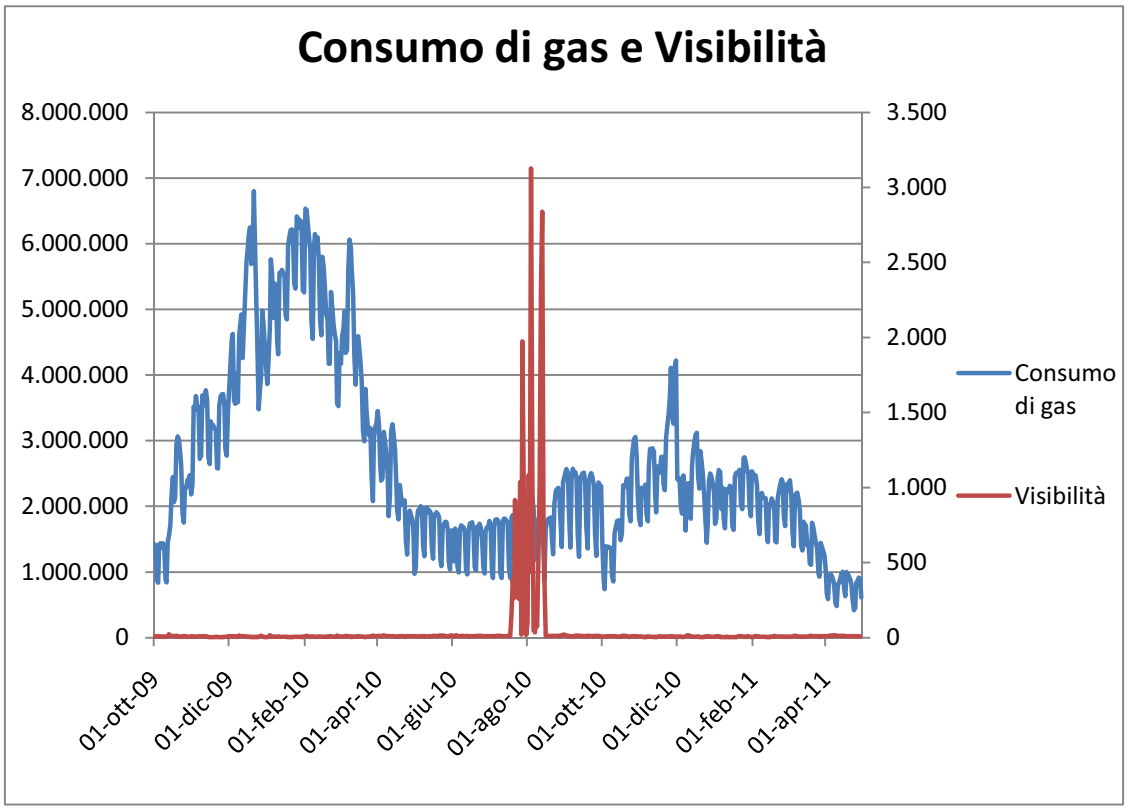



**GRAFICO 3: Consumo medio di gas e visibilità**

Dal Grafico 3 non si riesce a cogliere alcuna relazione tra la serie storica del consumo giornaliero complessivo di gas e la serie storica della visibilità media registrata; una relazione tra le due serie storiche non è da escludere a priori, ma non si riesce a cogliere semplicemente dall'osservazione del grafico precedente.

- **<u>Velocità media del vento</u>**: velocità media in km/h del vento registrata nel corso del giorno.

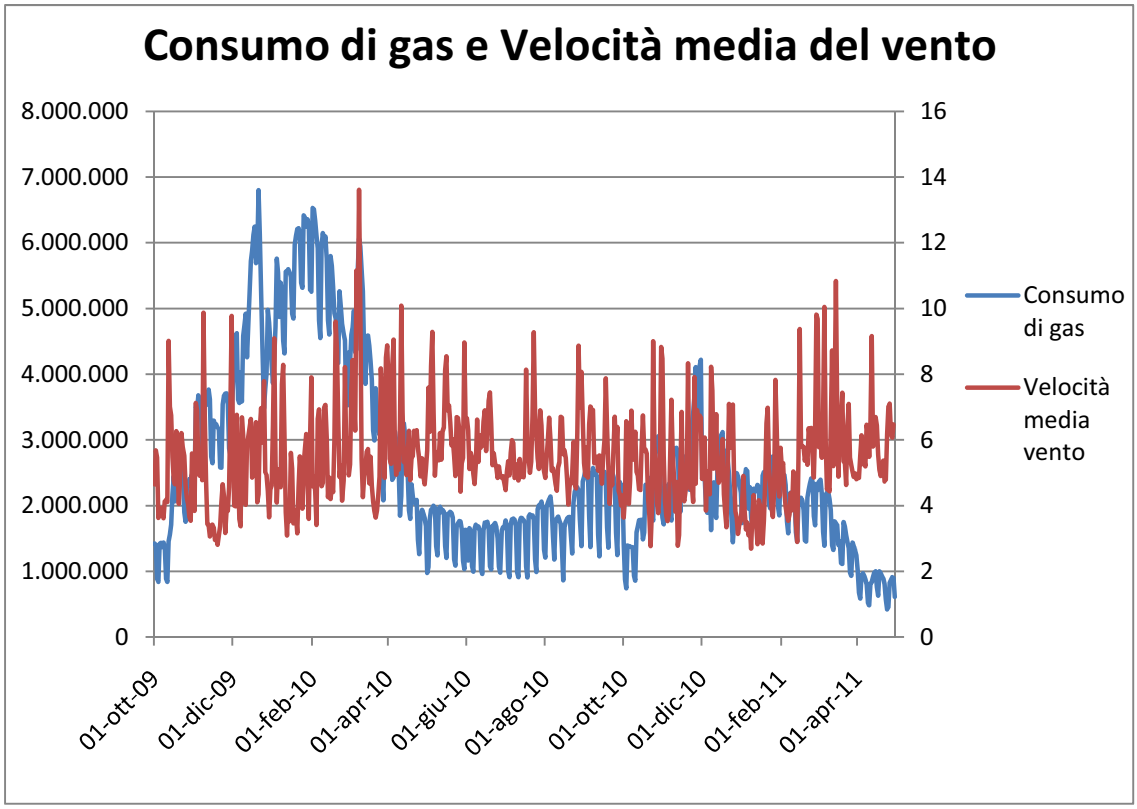



**GRAFICO 4: Consumo medio di gas e velocità media del vento**

Dal Grafico 4 si nota come la serie storica della velocità media del vento sia pressoché stazionaria; dal grafico precedente, tuttavia, non si riesce a cogliere alcuna correlazione con la serie storica del consumo giornaliero complessivo di gas, ma non si può nemmeno escludere a priori.

- **<u>Velocità massima del vento</u>**: picco massimo della velocità in km/h del vento registrato nel corso del giorno.

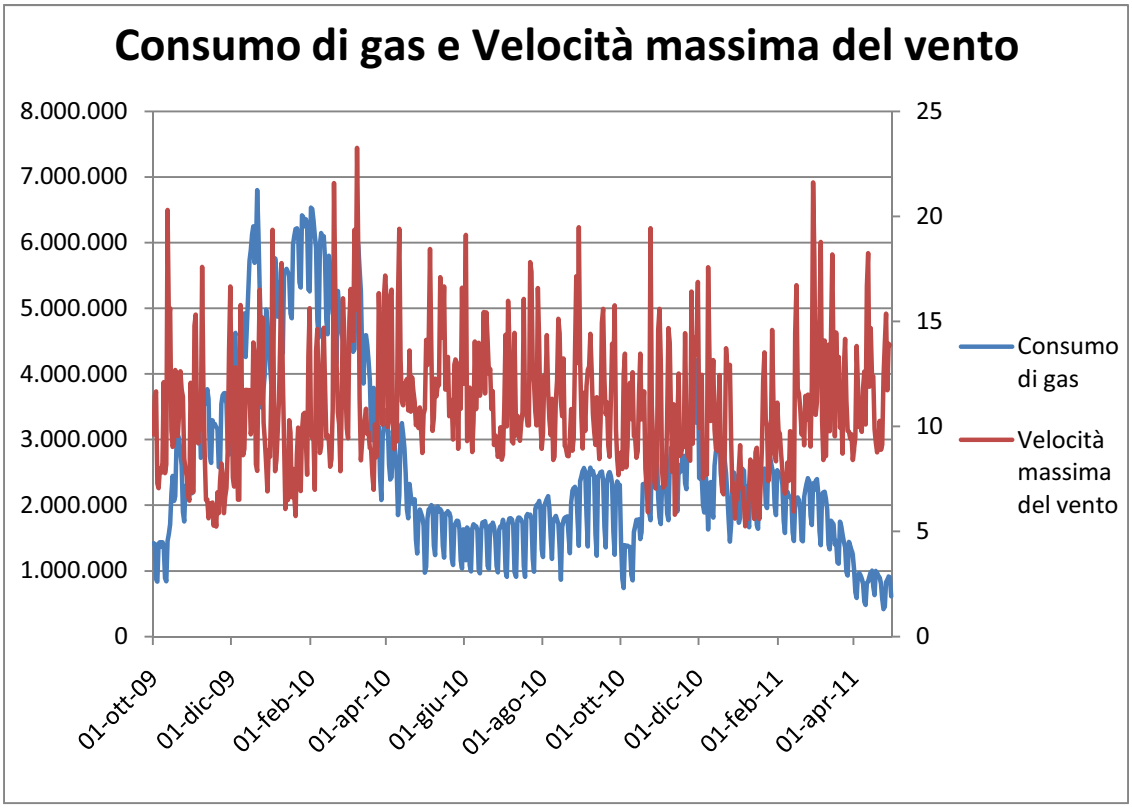



**GRAFICO 5: Consumo medio di gas e velocità massima del vento**

Dal Grafico 5 si nota come la serie storica della velocità massima del vento sia pressoché stazionaria; dal grafico precedente, tuttavia, non si riesce a cogliere alcuna correlazione con la serie storica del consumo giornaliero complessivo di gas, ma non si può nemmeno escludere a priori.

- **Raffica**: velocità in km/h delle raffiche di vento, se registrate.

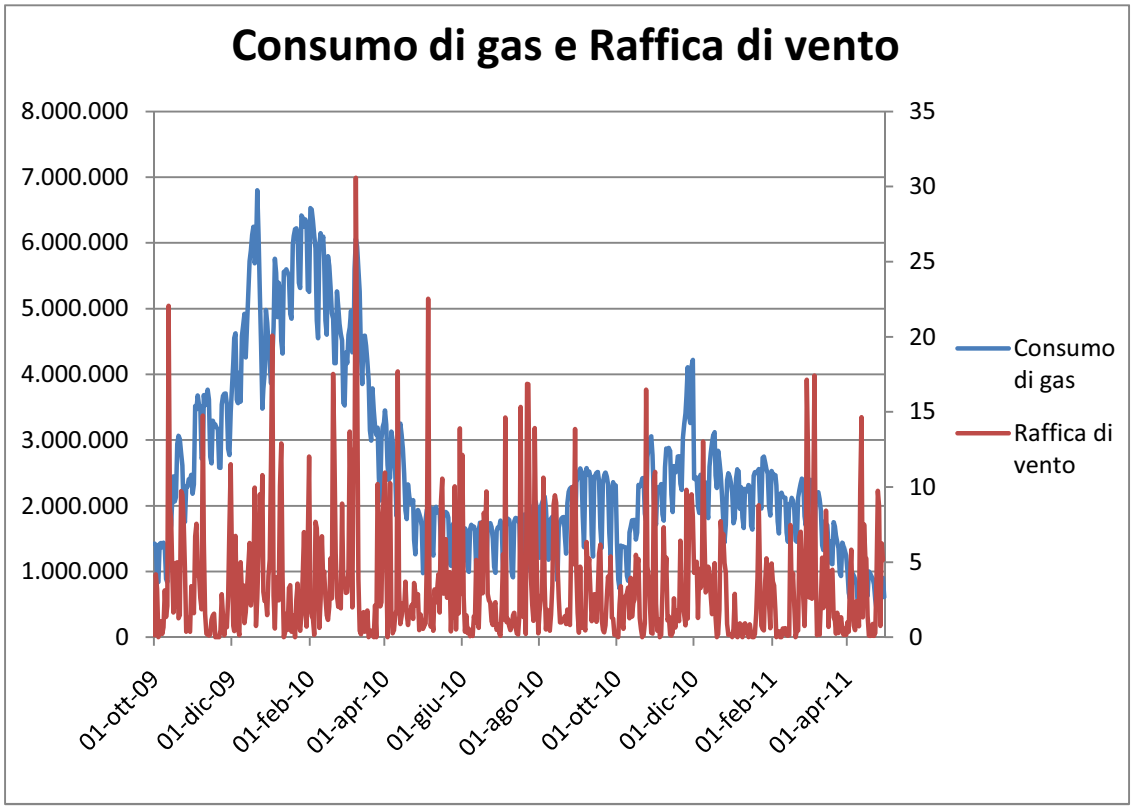



**GRAFICO 6: Consumo medio di gas e raffiche di vento**

Dal Grafico 6 si nota come la serie storica delle raffiche di vento sia pressoché stazionaria; dal grafico precedente, tuttavia, non si riesce a cogliere alcuna correlazione con la serie storica del consumo giornaliero complessivo di gas, ma non si può nemmeno escludere a priori.

- **Pressione**: pressione che l'atmosfera esercita sulla Terra; secondo il sistema internazionale si misura in Pascal, ma più frequentemente viene espressa in millibar o in atmosfere.

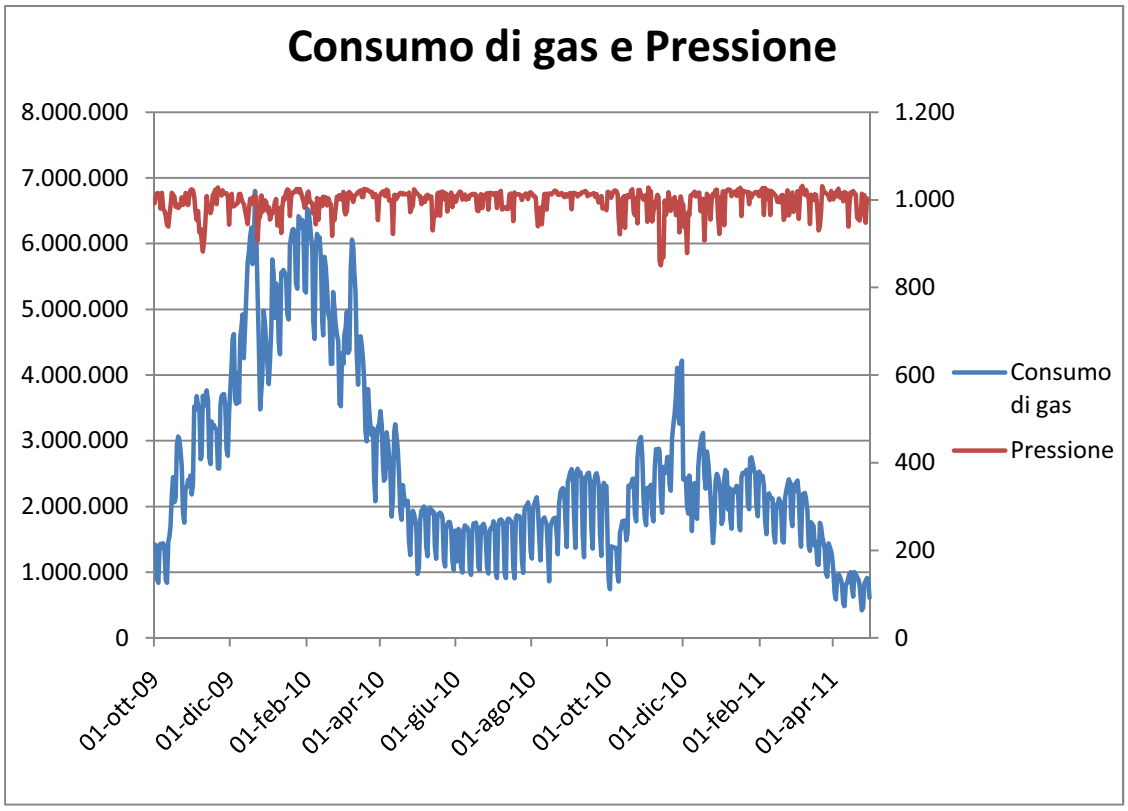



**GRAFICO 7: Consumo medio di gas e pressione**

Dal Grafico 7 si nota come la serie storica della pressione atmosferica sia pressoché stazionaria; dal grafico precedente, tuttavia, non si riesce a cogliere alcuna correlazione con la serie storica del consumo giornaliero complessivo di gas, ma non si può nemmeno escludere a priori.

# BIBLIOGRAFIA


Azzalini A. – Scarpa B., *Analisi dei dati e Data Mining*, 2009, Springer

Pace L. – Salvan A., *Introduzione alla Statistica*, 2001, Cedam

Ricci V., *Principali tecniche di regressione con R*, 2006

Di Fonzo T. – Lisi F., *Serie Storiche Economiche*, 2007, Carocci

Shumway R. – Stoffer D., *Time Series Analysis and Its Applications*, 2006, Springer

Montgomery D. – Jennings C. – Kulahci M., *Introduction to Time Series Analysis and Forecasting*, 2008, Wiley

Cryer J. – Chan K., *Time Series Analysis with applications in R*, 2008, Springer

Kirchgassner G. – Wolters J., *Introduction to Modern Time Series Analysis*, 2007, Springer

Petris G. – Petrone S. – Campagnoli P., *Dynamic Linear Models with R*, 2007, Springer

Brocklebank J. – Dickey D., *SAS for Forecasting Time Series*, 2003, Sas Institute

Crawley M., *The R Book*, 2007, Wiley

Cowpertwait P. – Metcalfe A., *Introductory Time Series with R*, 2009, Springer

Wei W., *Time Series Analysis : Univariate and multivariate methods*, 1989, Addison-Wesley

Lutkepohl H., *New Introduction to Multiple Time Series Analysis*, 2005, Springer

Tsay R., *Analysis of Financial Time Series*, 2010, Wiley

Zivot E. – Wang J., *Modeling Financial Time Series with S-Plus*, 2005

# SITOGRAFIA

*www.snamretegas.it*

*http://cran.r-project.org/*