Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in
Scienze Statistiche

Modellazione e valutazione del *Cost-at-Risk* nel Mercato italiano per il Servizio di Dispacciamento

Relatore: Prof. Francesco Lisi
Dipartimento di Scienze Statistiche

Correlatore: Ing. Federico Quaglia
Terna Rete Italia

Laureando: Anna Rosina
Matricola: 1202666

Anno Accademico 2021/2022

Alla mia famiglia

Abstract

L'obiettivo di questa tesi è quello di modellare il rischio associato ai costi sostenuti da Terna nel Mercato per il Servizio di Dispacciamento, per l'approvvigionamento di riserva di energia.

La valutazione del rischio di incorrere in costi non programmati per la riserva consiste nel modellarne la deviazione dal loro valore atteso. Si lavora su base giornaliera, con costi giornalieri, valutando il rischio con orizzonte di 1 giorno e 30 giorni.

Gli approcci adoperati sono due e, in entrambi, la complessità delle relazioni tra le variabili è modellata in modo additivo non parametrico. Il primo approccio è basato sulla modellazione della media condizionata con GAM accanto a quella dei residui di tale modello. Rappresentando la deviazione dalla media dei costi, i residui sono valutati con una misura di rischio adeguata: il CaR (*Cost-at-Risk*), derivata dal VaR (*Value-at-Risk*), ma applicata ai costi.

I modelli di CaR sono stimati sia sull'orizzonte giornaliero che mensile, ad un livello del 10%. Il CaR a 1 giorno è modellato con la distribuzione *kernel*, con un modello GARCH, con la regressione quantilica e con una sua specificazione, il modello CAViaR (Engle and Manganelli, 2004). Il calcolo del CaR a 30 giorni avviene su base giornaliera, estendendo l'orizzonte temporale con il metodo della simulazione storica filtrata (Barone Adesi, 2015). Il secondo approccio modella direttamente il rischio associato ai costi considerando il quantile al 10% dei costi e applicando a questo una generalizzazione del GAM al quantile condizionato, il *Quantile*-GAM (Fasiolo et al., 2021b).

I diversi modelli sono poi validati tramite tecniche di *backtesting*: il test di Kupiec (1995), o di copertura incondizionata, il test di Christoffersen (1998), di copertura condizionata alla verifica di indipendenza degli sforamenti e il test del quantile dinamico di Engle e Manganelli (2004).

Indice

Abstract v

Elenco delle figure ix

Elenco delle tabelle xiii

Introduzione 1

1 Il mercato elettrico in Italia 3
1.1 La filiera elettrica in Italia 4
1.1.1 La produzione 6
1.1.2 La trasmissione e il dispacciamento 13
1.1.3 La distribuzione 18
1.1.4 Il *metering* 18
1.2 Soggetti del mercato italiano 19
1.2.1 Soggetti non istituzionali 19
1.2.2 Soggetti istituzionali 20
1.3 Mercato elettrico a pronti 25
1.3.1 Principali modelli di MPE 25
1.3.2 Suddivisione in zone di mercato 26
1.3.3 Il Mercato del Giorno Prima 29
1.3.4 Il Mercato Infragiornaliero 35
1.3.5 Il Mercato dei Prodotti Giornalieri 39
1.3.6 Il Mercato per il Servizio di Dispacciamento 40
1.4 Mercato a Termine 63

2 Analisi preliminari 65
2.1 Il dataset 66
2.1.1 I costi di approvvigionamento 67
2.1.2 Le variabili 72

3 Strumenti statistici 91
3.1 Modelli additivi generalizzati . . . 94
3.1.1 Stima del modello GAM . . . 97
3.2 Splines . . . 101
3.2.1 Splines di regressione . . . 103
3.2.2 Splines di lisciamento . . . 104
3.3 Modelli di Costo-a-Rischio . . . 106
3.4 Modelli di CaR basati su distribuzione marginale . . . 109
3.4.1 CaR basato su metodo del nucleo . . . 111
3.5 Modelli di CaR basati su distribuzione condizionata . . . 114
3.5.1 CaR basato su modelli GARCH . . . 115
3.5.2 CaR basato sulla regressione quantilica . . . 121
3.5.3 CaR basato sul modello CAViaR . . . 126
3.6 Modello QGAM per il calcolo del rischio . . . 129
3.6.1 Quantile GAM . . . 130
3.7 Validazione del modello . . . 135
3.7.1 Test di Kupiec . . . 136
3.7.2 Test di Christoffersen . . . 137
3.7.3 Test del quantile dinamico . . . 140

4 Rischio associato ai costi di approvvigionamento 143
4.1 CaR a 1 giorno . . . 146
4.1.1 Stima e validazione del modello . . . 146
4.1.2 Verifica della capacità predittiva . . . 155
4.2 CaR a 30 giorni . . . 159
4.2.1 Stima e validazione del modello . . . 160
4.2.2 Verifica della capacità predittiva . . . 163

Conclusioni 170

Bibliografia 171

Sitografia 173

Elenco delle figure

1.1 Filiera dell'energia elettrica dal punto di vista tecnologico 5
1.2 Filiera dell'energia elettrica dal punto di vista economico 6
1.3 Carico di base e di picco della domanda di energia nel corso di 24 ore 7
1.4 Prezzo medio annuo 2020 dell'energia elettrica e capacità di interconnessione 2020 tra Italia e Paesi della frontiera Nord 12
1.5 Il Centro Nazionale di Controllo di Terna 14
1.6 Interventi sulla Rete di Trasmissione Nazionale programmati da Terna e annunciati nel Piano di Sviluppo 2021 15
1.7 Le due tratte, Est e Ovest, del Tyrrhenian Link 16
1.8 La tratta del collegamento SA.PE.I. 16
1.9 Struttura del mercato elettrico italiano 24
1.10 Configurazione zonale precedente e attuale a confronto 27
1.11 Stato attuale dell'implementazione del *Price Coupling of Regions* . . 31
1.12 Formazione dei prezzi nel Mercato del Giorno Prima 32
1.13 Integrazione tra il Mercato Infragiornaliero e il Mercato per il Servizio di Dispacciamento: scenario precedente 36
1.14 Integrazione tra il Mercato Infragiornaliero e il Mercato per il Servizio di Dispacciamento: scenario attuale 37
1.15 Mappatura tra i prodotti del Mercato italiano per il Servizio di Dispacciamento e i prodotti europei 46
1.16 La regolazione di frequenza in funzione del tempo di intervento . . 47
1.17 Modello di bilanciamento adottato nei Paesi Membri UE 62

2.1 Serie storica dei costi giornalieri sostenuti da Terna in MSD 68
2.2 Boxplot dei costi giornalieri di approvvigionamento per giorno della settimana 69
2.3 Boxplot dei costi giornalieri di approvvigionamento per mese dell'anno e per anno 70
2.4 Serie storica dei costi mensili sostenuti da Terna in MSD 71
2.5 Relazione tra i consuntivi nazionali giornalieri e i costi di approvvigionamento giornalieri, suddivisi per anno 74

2.6 Relazione tra la produzione di energia a livello nazionale da idroelettrico, fotovoltaico ed eolico e i costi di approvvigionamento giornalieri, suddivisi per anno . . . 77
2.7 Relazione tra la riserva secondaria a salire e i costi di approvvigionamento giornalieri, suddivisi per anno . . . 78
2.8 Stima della relazione tra la riserva secondaria a salire e i costi di approvvigionamento giornalieri per anno . . . 79
2.9 Relazione tra la riserva terziaria nazionale a salire e i costi di approvvigionamento giornalieri, suddivisi per anno . . . 81
2.10 Relazione tra la riserva terziaria a salire per Sardegna, Sicilia e Penisola e i costi di approvvigionamento giornalieri, suddivisi per anno . 82
2.11 Relazione tra i vincoli a rete integra e i costi di approvvigionamento giornalieri, suddivisi per anno . . . 83
2.12 Serie storica del prezzo giornaliero del gas naturale sul mercato TTF e del prezzo delle emissioni di CO_2 fino al 30 settembre 2021 . . . 85
2.13 Serie storica del prezzo giornaliero del gas naturale sul mercato TTF fino al 30 settembre 2020 . . . 86
2.14 Relazione tra il prezzo del gas sul TTF e i costi di approvvigionamento giornalieri, suddivisi per anno, fino a settembre 2020 . . . 87
2.15 Serie storica del prezzo giornaliero delle quote di emissione di CO_2 fino al 30 settembre 2020 . . . 88
2.16 Relazione tra il prezzo delle quote di emissione di CO_2 e i costi di approvvigionamento giornalieri, suddivisi per anno, fino a settembre 2020 . . . 89

3.1 La maledizione della dimensionalità . . . 95
3.2 Funzione di perdita *pinball* . . . 131

4.1 Effetto parziale della componente annuale e settimanale sulla media dei costi . . . 148
4.2 Effetto parziale dei costi ritardati $t-1$ e $t-7$ sulla media dei costi . 149
4.3 Effetto parziale dei vincoli a rete integra sulla media dei costi . . . 150
4.4 Effetto parziale delle variabili relative alle fonti rinnovabili sulla media dei costi . . . 151
4.5 Effetto parziale della riserva secondaria e terziaria sulla media dei costi 152
4.6 Serie storica dei costi giornalieri per il periodo di verifica, e modelli CaR a 1 giorno: GAM-K, GAM-GARCH, GAM-QR e GAM-CAViaR . . . 155
4.7 Serie storica dei costi giornalieri per il periodo di verifica, modello sulla media condizionata e modello di CaR a 1 giorno con GAM-GARCH . . . 157

4.8 Sequenza e ampiezza degli sforamenti osservati nel periodo di verifica del CaR a 1 giorno calcolato con modello GAM-GARCH 158
4.9 Distribuzione del CaR a 1 giorno calcolato con modello GAM-GARCH 158
4.10 Serie storica dei costi a 30 giorni per il periodo di verifica, modello sulla media condizionata e modello di CaR a 30 giorni con GAM-GARCH 163
4.11 Sequenza e ampiezza degli sforamenti osservati nel periodo di verifica del CaR a 30 giorni calcolato con GAM-GARCH 165
4.12 Distribuzione del CaR a 30 giorni calcolato con modello GAM-GARCH 165

Elenco delle tabelle

4.1 Funzioni utilizzate nel modello GAM sulla media condizionata . . . 147
4.2 Livello di copertura osservato nell'insieme di validazione e valore del p-value dei test di Kupiec, Christoffersen e del quantile dinamico per il CaR a 1 giorno 154
4.3 Livello di copertura osservato nell'insieme di verifica e valore del p-value dei test di Kupiec, Christoffersen e del quantile dinamico per il CaR a 1 giorno 156
4.4 Statistiche riassuntive della distribuzione del CaR a 1 giorno 158
4.5 Livello di copertura osservato nell'insieme di verifica e valore del p-value dei test di Kupiec, Christoffersen e del quantile dinamico per il CaR a 30 giorno 164
4.6 Statistiche riassuntive della distribuzione del CaR a 30 giorni 165

Introduzione

L'energia elettrica è tutto ciò che garantisce alla nostra società di essere quello che è oggi. Tutto ciò che ha permesso l'evoluzione dello stile di vita dell'uomo, e il miglioramento di molte delle sue attività è dovuto alla capacità sviluppata dall'uomo di gestirla.
Al giorno d'oggi, è così intrinsecamente presente nella nostra vita e funzionale al nostro sostentamento, che ci appare naturale utilizzare l'energia sotto forma di elettricità, secondo i nostri bisogni. Tanto che eventi di occasionali *blackout* hanno sempre creato enormi disagi. Ma accanto a questi, anche una maggiore consapevolezza di ciò che appare scontato, come accendere la luce quando è notte o caricare il cellulare quando è scarico. Quindi, perché i *blackout* possono verificarsi? Ma soprattutto, come è possibile che, vista la grande complessità del sistema, si verifichino così raramente?
Per rispondere a queste domande occorre capire cosa c'è dietro. Ciò che per la nostra società è ormai naturale, in realtà è il frutto di un lavoro estremamente accurato di approvvigionamento delle risorse energetiche, previsione dei bisogni della popolazione e controllo dell'equilibrio complessivo del sistema.

In Italia, a garantire tali servizi, è Terna. In particolare, la predisposizione di riserva è tra le sue principali attività, e in qualità di *Transmission System Operator* italiano, Terna è responsabile della trasmissione di energia elettrica sulla rete nazionale e garante della sua sicurezza.
L'approvvigionamento di riserva e la sua eventuale attivazione rappresentano il mezzo con cui Terna mantiene l'equilibrio tra immissioni e prelievi nel sistema elettrico nazionale, garantendo il funzionamento continuo di ogni sua parte.

Il Mercato per il Servizio di Dispacciamento è la piattaforma su cui Terna opera al fine di acquisire l'energia di riserva. E per fare ciò sostiene dei costi, analizzati in questa tesi.
Il nostro obiettivo è quello di valutare il rischio associato al fatto che questi costi possano discostarsi dal loro valore atteso, operando su dati forniti da Terna. Sono stati utilizzati dati giornalieri, costruendo modelli per il rischio con orizzonte temporale giornaliero e mensile, utilizzando una misura di rischio adeguata ai costi, il Costo-a-Rischio (CaR). La

finalità è capire quale sia la quantità di denaro che Terna deve immobilizzare per far fronte ai costi entro un giorno ed entro un mese.

La tesi si struttura in quattro capitoli. Il primo capitolo descrive il mercato elettrico italiano, soffermandosi sulla filiera elettrica e sulle dinamiche del Mercato del Giorno Prima (MGP), del Mercato Infragiornaliero (MI) e del Mercato per il Servizio di Dispacciamento (MSD).

Il secondo capitolo presenta i dati forniti da Terna e utilizzati nell'analisi, rappresentando, a livello descrittivo, le relazioni tra i costi di approvvigionamento e le altre variabili.

Il terzo capitolo espone dal punto di vista teorico i modelli impiegati per il calcolo del rischio, come i modelli additivi generalizzati (GAM), le *splines*, i modelli di misura del rischio CaR, basati su GARCH e sulla regressione quantilica, i modelli Quantile-GAM, e i test utilizzati per la validazione dei modelli.

Per finire il quarto capitolo, descrive l'applicazione di tali modelli ai nostri dati, per il calcolo del Costo-a-Rischio a 1 giorno e a 30 giorni, individuando il miglior modello sulla base di alcune tecniche di *backtesting* come i test di Kupiec, Christoffersen e del quantile dinamico.

1

Il mercato elettrico in Italia

Nel corso degli anni novanta si è assistito ad un progressivo processo di liberalizzazione del settore elettrico a livello internazionale, con la conseguente creazione dei mercati elettrici. In Italia il mercato elettrico, o Borsa Elettrica, nasce con il *Decreto Legislativo 16 marzo 1999, n. 79* (1999), il "Decreto Bersani", per rispondere all'esigenza di promuovere la competizione delle attività di produzione e vendita all'ingrosso e favorire la massima trasparenza ed efficienza dell'attività di dispacciamento. Il decreto avvia la liberalizzazione del settore elettrico, nell'ambito del recepimento della prima direttiva comunitaria, la *Direttiva 96/92/CE del Parlamento Europeo e del Consiglio del 19 dicembre 2016* (2016), per la creazione di un mercato interno dell'energia. Nasce così l'*Italian Power Exchange* (IPEX).

Dal 1° aprile 2004, giorno del *go-live* della borsa elettrica in Italia, la spinta all'integrazione dei mercati nazionali, ha coinvolto inizialmente i mercati dell'energia, rivolgendosi poi ai mercati di bilanciamento, in base alle linee guida europee CACM (*Capacity Allocation and Congestion Management*) ed EBGL (*Electricity Balancing GuideLine*).

Il mercato elettrico nasce come luogo di negoziazione all'ingrosso dei prezzi orari dell'energia, delle quantità scambiate e dei programmi di immissione e prelievo vincolanti nella (e dalla) rete secondo il criterio di merito economico. La sua istituzione ha garantito di fatto un migliore processo di formazione dei prezzi, una gestione più trasparente e flessibile

anche delle situazioni di scarsità di offerta e ha introdotto un elemento di separazione tra l'attività di produzione e di vendita di energia, due attività concorrenziali all'interno della filiera elettrica.
Prima di procedere alla descrizione del mercato elettrico italiano, e dei soggetti che vi operano, di seguito è presentata la struttura della filiera dell'energia, con le quattro fasi di cui si compone.

1.1 La filiera elettrica in Italia

Fino agli anni novanta, in tutto il mondo, il settore elettrico era nazionalizzato: un attore verticalmente integrato, solitamente una società di stato, si occupava di tutta la filiera elettrica.
In Italia, prima della liberalizzazione del mercato, e precisamente dal 1962, la filiera produttiva era gestita totalmente da Enel, in veste di monopolista statale di settore. Le fasi della filiera elettrica che, con la liberalizzazione del mercato, sono state definite separatamente, erano dapprima riunite sotto l'unica gestione di Enel.

La liberalizzazione, avviata in Italia dal d.lgs 79/99[1], realizza la concreta apertura del mercato, intaccando alla base il monopolio di Enel sull'intera filiera, rendendo autonome le varie fasi dal punto di vista economico e operativo, con norme create *ad hoc* per ciascuna, mediante un processo di disaggregazione, detto *unbundling*. Tale processo ha condotto all'identificazione di quattro fasi, dal punto di vista tecnologico, cioè relativo alla gestione fisica dell'energia elettrica, rappresentate in Figura 1.1:

- produzione;
- trasmissione e dispacciamento, funzioni diverse ma riunite nella stessa fase in quanto gestite in modo integrato dallo stesso ente, Terna;
- distribuzione;
- *metering* o utenze.

[1] *Decreto Legislativo 16 marzo 1999, n. 79* (1999)

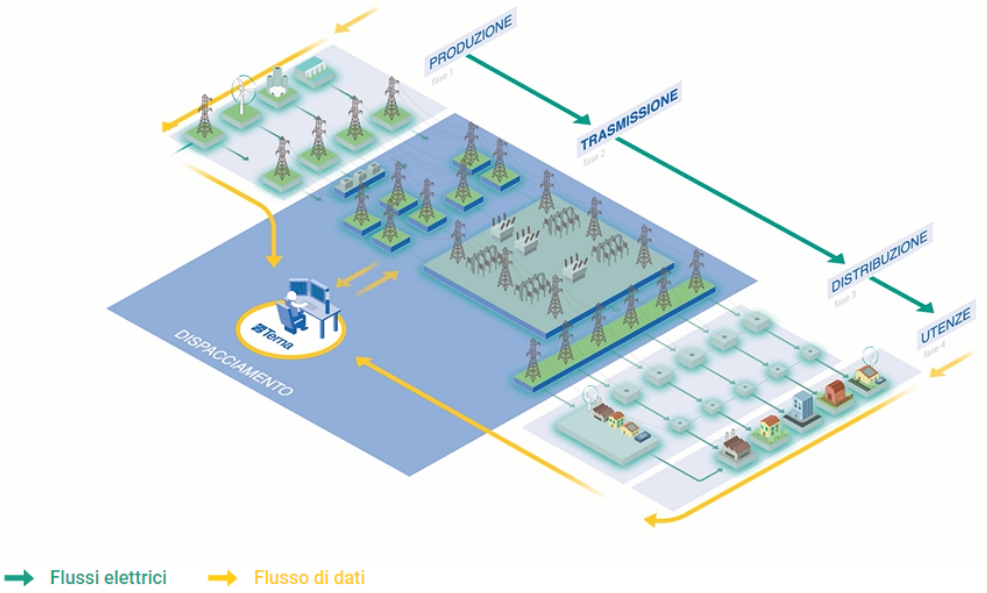


Figura 1.1: Filiera dell'energia elettrica dal punto di vista tecnologico. Fonte[2]

Il d.lgs 79/99 ha rappresentato la prima azione di una serie di provvedimenti attuati per rispondere all'esigenza di promuovere la competizione delle attività di produzione e vendita all'ingrosso e favorire la massima trasparenza ed efficienza dell'attività di dispacciamento. Il decreto avvia la liberalizzazione del settore elettrico, nell'ambito del recepimento della prima direttiva comunitaria (Direttiva 96/92/CE) [3], per la creazione di un mercato interno dell'energia. Nasce così l'*Italian Power Exchange* (IPEX), il mercato elettrico italiano, cuore dell'attività di compravendita all'ingrosso di energia elettrica. È proprio questa la novità strutturale introdotta nella filiera con la liberalizzazione del mercato elettrico: una fase di vendita all'ingrosso e una di vendita al dettaglio, affiancate alle fasi di produzione e distribuzione, ora rese autonome.
La filiera produttiva dell'energia elettrica rimane sostanzialmente uguale, sebbene ora gestita con nuove modalità: ciò che viene modificato è l'aspetto di mercato della stessa. Come illustrato in Figura 1.2, la produzione, la successiva vendita all'ingrosso sul mercato elettrico, e la vendita al dettaglio, effettuata tramite contratti con le società di vendita, sono attività liberalizzate; le fasi intermedie di trasmissione, dispacciamento e distribuzio-

[2] Terna (2021a)

[3] *Direttiva 96/92/CE del Parlamento Europeo e del Consiglio del 19 dicembre 2016* (2016)

ne (affiancata dall'attività di misurazione e controllo), operano in regime di monopolio naturale, detenuto da Terna.

Figura 1.2: Filiera dell'energia elettrica dal punto di vista economico. Fonte[4]

1.1.1 La produzione

La produzione di energia elettrica, che non esiste in natura, avviene mediante la trasformazione in elettricità dell'energia ricavata da fonti primarie. In Italia la produzione di energia elettrica avviene sfruttando principalmente fonti non rinnovabili o fossili (gas naturale, carbone e petrolio), anche se è in continuo aumento lo sviluppo di fonti rinnovabili, come il fotovoltaico e l'eolico, che nel 2021 sono incrementate del 2,1% e del 10,8% rispettivamente, ma anche il geotermico e l'idroelettrico, che in quest'anno hanno subito tuttavia una riduzione del 2,1% e del 5,4% rispettivamente, compensate dall'aumento delle precedenti. Complessivamente, al 2021, la percentuale di copertura del fabbisogno da fonti rinnovabili ha raggiunto il 36%, nella direzione promossa dal Green Deal, accordo finalizzato a trasformare l'Unione Europea in un'economia a zero emissioni entro il 2050[5]. L'introduzione sempre più consistente di produzione da fonti rinnovabili pone la grande

[4] SelectraItalia (2022)

[5] Terna (2021e)

sfida della programmabilità delle stesse. L'energia elettrica non è immagazzinabile, se non in quantità minime rispetto al necessario e poterne disporre al momento del bisogno programmando la sua produzione in vista dell'immediato consumo è di primaria importanza per il corretto funzionamento dell'intero sistema elettrico. Tuttavia questo non è sempre possibile, se solo si pensa a come avviene la generazione dal solare o dall'eolico: le previsioni a volte si discostano dalle condizioni meteorologiche effettive e una giornata meno ventosa del previsto può tradursi in una produzione reale di eolico inferiore a quella prevista. A tale problematica si somma il fatto che vi è una quantità di energia che deve essere necessariamente prodotta per soddisfare il fabbisogno della collettività. Il cosiddetto carico di base, *base load*, è il livello minimo di domanda richiesta in un periodo di 24 ore e pertanto la potenza minima che è necessario fornire continuativamente al sistema elettrico per garantire il funzionamento di tutte le sue componenti. Il carico di picco, *peak load*, riguarda invece i picchi di energia, solitamente di breve durata, figurabile come in Figura 1.3.

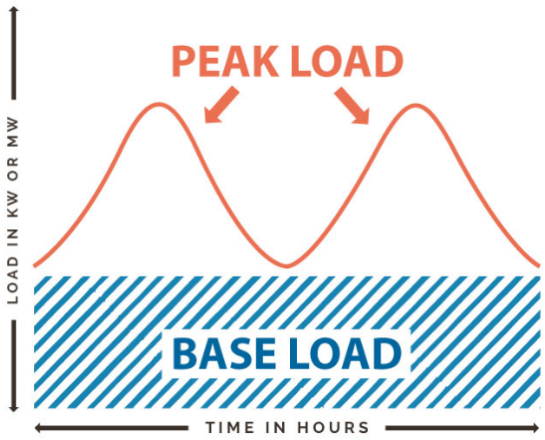


Figura 1.3: Carico di base e di picco della domanda di energia nel corso di 24 ore. Fonte[6]

Rispondere alle due differenti situazioni di carico richiede diverse tecnologie: gli impianti in grado di far fronte a picchi di domanda, accesi e spenti appena prima e subito dopo il picco sono:

- le centrali a gas, centrali termoelettriche cui nucleo motore è un sistema, il turbogas, composto da compressore, camera o sistema di combustione e turbina. Basate sul ci-

[6] De Rooij (2022)

clo termodinamico di Brayton-Joule, sfruttano l'energia chimica prodotta dalla combustione di un mix di gas naturale e combustibile, come il gasolio o il gas metano. Questa viene trasformata in energia termica, la quale aziona una turbina, trasfomandosi in energia meccanica. Mettendo in rotazione l'alternatore, viene infine trasformata in energia elettrica. Le centrali di questo tipo possono essere semplici, combinate o cogenerative, efficienti e in grado di realizzare una grandissima produzione, ma per contro, non sostenibili e inquinanti.

- le centrali eoliche, dette più comunemente parchi eolici, *wind farm* in inglese, sono costituite da un gruppo di turbine: gli aeromotori eolici o aerogeneratori. Localizzati in un territorio delimitato, lontano dai centri abitati, sia su terraferma (*on-shore* e *near-shore*), che in mare aperto (*off-shore*), meno impattanti dal punto di vista acustico e visivo sul paesaggio, e interconnessi tra di loro, producono energia sfruttando la forza del vento. Il movimento delle pale di cui sono composti gli aerogeneratori induce quello di un rotore: la sua energia di rotazione viene trasmessa attraverso l'albero, posto nel palo, verso la base. Un'evoluzione della turbina, l' aerogeneratore magnetoeolico, sfrutta l'attivazione la presenza di magneti permanenti (ad esempio al neodimio) oppure elettromagneti a superconduttore del tipo Maglev. La creazione di un campo elettromagnetico viene impiegata per ridurre l'attrito sperimentato dal rotore e dall'asse del pignone principale del rotore aeronautico.
  È una delle fonti rinnovabili tradizionali, diffusa soprattutto in Paesi con ventosità costante e direzionale con ridotta turbolenza, come la Danimarca e i Paesi Bassi. In Italia, con impianti presenti quasi esclusivamente al Sud e sulle isole, copre, insieme al solare, il 14,4% dell'intera produzione di energia, una quota ancora abbastanza ridotta. Per questo tra i risultati principali al 2030 predisposti da Terna vi è l'incremento ulteriore della capacità installata di eolico e fotovoltaico (+26 GW dal 2025 al 2030) [7].

- le centrali solari, in cui l'energia associata alla radiazione solare viene convogliata per produrre elettricità. Vi sono due tipologie di centrale solare: le fotovoltaiche e le termodinamiche.
  Le prime utilizzano l'effetto fotovoltaico per produrre elettricità, ovvero la capacità di alcuni materiali semiconduttori di generare elettricità se esposti alla radiazione luminosa. Il principale problema di questi impianti è dovuto al fatto che funzionano esclusivamente durante le ore di luce, ma spesso sono proprio questi i momenti in cui

[7] Terna (2021d)

è maggiore la richiesta.
Le seconde, dette anche "a concentrazione", utilizzano non pannelli fotovoltaici, ma specchi (parabolici lineari o circolari, lineari a riflettore e a torre): convogliando i raggi solari verso un ricevitore, riscaldano il fluido termovettore contenuto in esso, che si trasforma in vapore e viene da qui indirizzato da un sistema di tubazioni fino ad azionare una turbina, da cui l'energia meccanica prodotta viene trasmessa all'alternatore, che la trasforma in elettricità [8].

- le centrali idroelettriche, sistemi ingegneristici finalizzati alla conversione dell'energia cinetica generata da masse d'acqua in movimento in energia elettrica. Ve ne sono di diverse tipologie: ad acqua fluente, che sfruttano la velocità di una corrente d'acqua, a bacino, artificiale, se creato da una diga, naturale, se già esistente, come un lago, o ad accumulazione. Queste ultime utilizzano un bacino di raccolta anche a valle: quando la domanda di energia è minore, l'acqua del bacino a valle viene riportata nel bacino a monte mediante un sistema di pompaggi, recuperando massa idrica, che viene impiegata per una maggiore produzione in caso di bisogno. Diffusi soprattutto nel nord Italia, grazie ad abbondanza delle acque di scioglimento delle nevi, questi impianti hanno un notevole impatto visivo, in quei luoghi adatti ad ospitarli per la loro conformazione ambientale, e questo è uno dei principali punti negativi dell'idroelettrico. Tuttavia, sfrutta una risorsa rinnovabile e non inquinante, fornendo anche un buon quantitativo di energia. Grazie alla possibilità di attivarle e disattivarle in pochi minuti con l'immediata apertura delle saracinesche idrauliche, sono impiegate nella risposta a picchi di fabbisogno.

Gli impianti chiamati a garantire il soddisfacimento del fabbisogno minimo, provvedendo a generare energia in modo continuo durante il giorno, sono invece:

- le centrali nucleari, che funzionano riscaldando un fluido termovettore attraverso l'energia liberata dalle reazioni nucleari, trasformando l'energia nucleare in termica e poi elettrica. Il cuore del funzionamento di una centrale nucleare, come in quelle termoelettriche, risiede in un ciclo termodinamico che origina calore non dalla combustione di un gas o di olio combustibile (carbone, nafta, orimulsion o metano), bensì dalla fissione nucleare dell'uranio-235 e del plutonio-239 contenuti nel combustibile composto principalmente da uranio-238. A differenziare le tipologie di centrale è il refrigerante del reattore e del moderatore impiegati: il più diffuso è l'acqua (pressurizzata,

[8] EnelGreenPower (2021)

bollente o pesante pressurizzata), ma può essere impiegata anche grafite, acqua leggera o gas. Attorno all'impiego di energia nucleare vi è un forte dibattito che vede scontrarsi pro e contro: questi impianti generano emissioni particolarmente basse di CO_2, ma non sfruttano una fonte rinnovabile, risultando poco sostenibili nel lungo periodo, e dall'altra parte pongono il problema dello smaltimento di scorie nucleari. Consentono di ridurre la dipendenza da petrolio e gas, garantendo ingenti produzioni di energia e una maggiore stabilità politica, ma portano con sè un problema di sicurezza e potenziale pericolo in caso di incidenti, come la storia ci ha insegnato con i disastri di Chernobyl e Fukushima.

- le centrali a carbone, centrali termoelettriche in cui la produzione di energia elettrica avviene a partire dalla combustione del carbon fossile: il calore viene impiegato per riscaldare l'acqua generando vapore, il quale, altamente pressurizzato, fa girare le pale di una turbina collegate ad un alternatore. Come il petrolio e il gas naturale, il carbone è un combustibile fossile presente in natura, ma allo stato solido. La sua formazione risale a 345 milioni di anni fa: il clima caldo e umido di allora, e un'alta concentrazione di CO_2 favorirono la crescita di alberi imponenti, abbattuti dopo secoli da potenti inondazioni che crearono uno strato di legname non degradabile dai batteri esistenti al tempo. La forte pressione dei sedimenti venutisi a formare e l'assenza di ossigeno ha portato alla formazione di questa roccia sedimentaria, che dalla sua scoperta è stata estratta e bruciata incessantemente. Ancora oggi gli impianti a carbone rispondono al 4,9% della produzione netta italiana[9] e nel 2021, dopo un costante calo negli anni, l'energia prodotta bruciando la fonte fossile in assoluto più inquinante è aumentata del 9% [10].

- le centrali geotermiche, che sfruttano l'energia termica proveniente dalle profondità terrestri per produrre energia elettrica. Come nelle centrali termoelettriche, il vapore viene utilizzato per produrre energia termica, ma se in queste veniva prodotto artificialmente da un generatore, con l'utilizzo di un combustibile, nelle centrali geotermiche si utilizza il vapore originato dall' evaporazione dell'acqua piovana per contatto con le temperature elevate del sottosuolo. Infatti la centrale geotermica dispone di strumenti per la depurazione del vapore attraverso l'estrazione di gas incondensabili di scarto, come la CO_2. Vi sono tre tipologie principali di questo impianto: a vapore dominante

[9] Ferraino (2022)
[10] Bongioanni (2022)

(*dry steam*), ad acqua dominante (*flash*) e a ciclo binario.
L'energia geotermica è un'alternativa rinnovabile, utilizza energia non intermittente e inesauribile, a zero emissioni di CO_2, e grazie al riciclo del vapore prodotto, permette il riutilizzo anche degli scarti di produzione. Rispetto alle altre fonti rinnovabili consente una produzione molto maggiore a parità di potenza elettrica installata, ma soprattutto in modo continuativo. Gli svantaggi principali di questo tipo di impianti risiedono nell'impatto visivo e nell'odore prodotto dalle emissioni di idrogeno solforato, ma per queste ultime sono stati studiati sistemi di abbattimento, mentre per l'aspetto estetico, sono sempre più diffusi progetti di bio-architettura [11].

- le centrali a biomasse, che sfruttano l'energia ricavabile dalle biomasse con processi termochimici: combustione diretta, pirolisi, o gassificazione, con l'estrazione di gas di sintesi. In alternativa, vengono impiegati processi biochimici, di digestione anaerobica o aerobica, sfruttando la degradazione della sostanza da parte di enzimi e batteri specifici. Biomassa è il termine con il quale viene indicata dal d.lgs 397/03[12] (art. 2, comma 1) "la parte biodegradabile dei prodotti residui provenienti dall'agricoltura (comprendente sostanze animali e vegetali) e dalla silvicoltura e da industrie connesse, nonchè la parte biodegradabile dei rifiuti industriali e urbani". Il facile reperimento della materia prima rende l'energia prodotta da biomassa un ottima risorsa, nonostante il trasporto possa incidere sulla sua sostenibilità ambientale.
- le centrali biogas, che utilizzano la digestione anaerobica per la produzione del biogas a partire dalle biomasse: in un ambiente umido e privo di ossigeno (il digestore), la biomassa fermenta grazie a particolari enzimi e batteri producendo biogas, costituito principalmente da metano e anidride carbonica. Si distingue in digestione "a secco", la biomassa con un contenuto solido minimo del 30%, mentre se la percentuale si aggira attorno ad un 10-15%, si parla di digestione "a umido". I vantaggi principali di questa tecnologia riguardano il risparmio economico dovuto alla produzione del biogas da biomasse, sfruttando un materiale altrimenti inutile e difficile da depositare e smaltire, per contro la lavorazione di questo materiale causa cattivo odore, il che rende questi impianti di difficile collocazione vicino a centri abitati.

Per soddisfare il fabbisogno energetico nazionale, oltre a produrre energia da fonti rinnovabili in modo sempre più consistente, l'Italia acquista energia elettrica anche da altri

[11] E.ONEnergia (2020)
[12] *Decreto Legislativo 29 dicembre 2003, n. 387* (2003)

Paesi: nel 2021 l'86,5% del fabbisogno nazionale è stato assicurato dalla produzione nazionale, pari alla produzione netta del parco di generazione decurtata dell'energia destinata ai pompaggi, mentre il restante 13,5% riguarda il saldo dell'energia scambiata con l'estero, in particolare con Francia e Svizzera, sulle cui frontiere la capacità di interconnessione è maggiore[13].

L'importazione è legata infatti a due fattori principali: il differenziale di prezzo con i Paesi confinanti e la capacità delle interconnessioni transfrontaliere. L'Italia è storicamente un paese importatore di energia elettrica: da anni presenta uno spread positivo di prezzo dell'energia elettrica con i Paesi della frontiera Nord. Ad esempio, nel 2020, come mostrato in Figura 1.4 lo spread medio del costo di 1 KWh rispetto a questi Paesi è stato di 4 euro, in seguito ad un calo del 59% rispetto al valore dell'anno precedente (9 €/MWh nel 2019) dovuto alle condizioni di emergenza globali. Il differenziale minimo di 0,2 €/MWh sulla frontiera slovena, e massimo di circa 6 €/MWh su quella francese, sono dovuti alla presenza di tecnologie dai costi marginali inferiori rispetto all'Italia: in Francia più del 70% della produzione di energia proviene dal nucleare, l'Austria copre il 60% con idroelettrico, in Slovenia, invece, i 2/3 dell'energia prodotta provengono da idroelettrico e nucleare. Anche la Svizzera presenta uno spread negativo con l'Italia, essendo di fatto un ponte che collega l'Italia con Francia e Germania, paese con elevati livelli di produzione da lignite ed eolico[14].

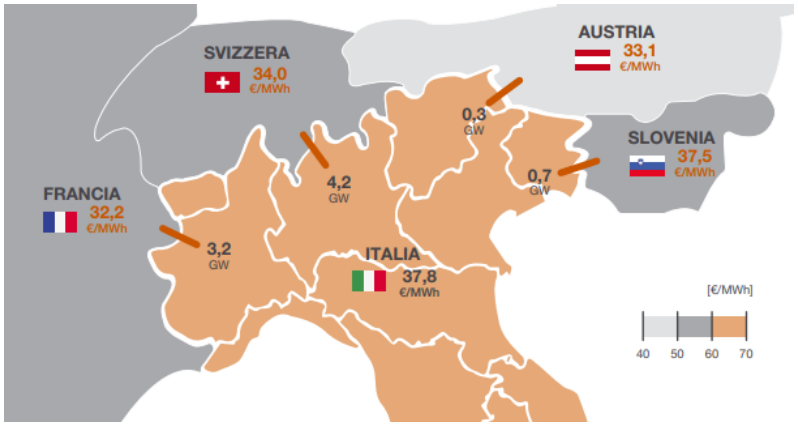


Figura 1.4: Prezzo medio annuo 2020 dell'energia elettrica [€/MWh] e capacità di interconnessione 2020 [GW] tra Italia e Paesi della frontiera Nord. Fonte[14]

[13] Terna (2022d)

[14] Terna (2021b)

Nonostante il costo inferiore dell'energia nei Paesi sopra citati, da cui la convenienza ad importare, la generazione sul territorio nazionale rappresenta la quota principale di provenienza di energia in Italia. Comprende diverse attività, dall'approvvigionamento delle materie prime, la trasformazione dell'energia primaria in energia elettrica, l'immissione dell'energia prodotta nella rete, alla costruzione e manutenzione degli impianti. Con la liberalizzazione del mercato, la produzione di energia, prima monopolizzata da Enel, è stata portata ad un regime di mercato concorrenziale e dal 1° Gennaio 2003 ha visto l'introduzione di un'importante limitazione per le società produttrici di energia: ciascuna non può controllare (direttamente o indirettamente) più del 50% del mercato (energia elettrica generata e importata). Laddove si rileva capacità eccedente di produzione la società è portata alla vendita, mossa che di fatto ha contribuito alla nascita di nuovi operatori nel settore. Accanto alla produzione assicurata da Enel, gestita tramite la società Enel Produzione, sono quindi nati nuovi operatori del mercato elettrico, in grado di produrre autonomamente energia grazie a nuovi centrali elettriche di loro proprietà e di venderla. Tra i più importanti: Eni, Edison, A2A, EPH, Iren, Engie, Sorgenia, ERG, Alperia, Axpo Group, Saras.

1.1.2 La trasmissione e il dispacciamento

L'energia, importata o prodotta e venduta all'ingrosso, viene trasportata, dalle società produttrici ai distributori locali, sulla rete di trasmissione ad alta e altissima tensione, gestita da Terna S.p.A., che opera in monopolio per garantire la sicurezza e l'efficienza del sistema. L'energia viaggia dalle unità di produzione attraverso la rete ad altissima tensione a livelli di 380 kV (kiloVolt) e 220 kV percorrendo grandi distanze. Circola poi a livelli attorno ai 150 kV, 132 kV e 60 kV nella rete ad alta tensione, che si distribuisce più capillarmente nel territorio, fino a raggiungere le cabine primarie. La gestione di questi flussi di energia avviene in tempo reale attraverso un sistema che fa capo al Centro Nazionale di Controllo (CNC), rappresentato in Figura 1.5, il cuore del sistema elettrico italiano, che attraverso oltre 100 schermi di controllo e un wallscreen di 40 metri quadrati, monitora 293 linee, tra cui 9 interconnessioni con l'estero, 3 cavi sottomarini e 281 linee nazionali a 380 kV.

Vincolo sostanziale per il corretto funzionamento della rete è l'equilibrio tra domanda e offerta sul sistema elettrico nazionale, ovvero tra energia immessa e prelevata, che Terna

[15] Terna (2022b)

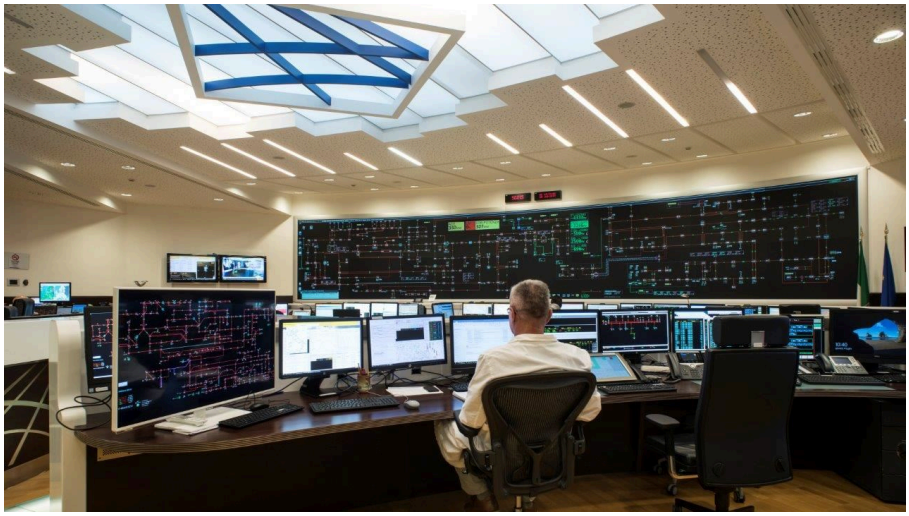


Figura 1.5: Il Centro Nazionale di Controllo di Terna, a Roma. Fonte[15]

provvede a mantenere costantemente mediante l'attività di dispacciamento. A tal fine, questa attività consiste ne:

- la verifica dei transiti di potenza lungo tutti i 74.723 km di linee elettriche gestite da Terna e le 431 stazioni di trasformazione e smistamento;
- la programmazione delle indisponibilità di rete e degli impianti di produzione con diversi orizzonti temporali;
- la previsione del fabbisogno elettrico nazionale;
- il confronto di coerenza tra fabbisogno e programma delle produzioni, determinato come esito del mercato libero dell'energia (Borsa Elettrica e contratti fuori Borsa);
- l'acquisizione di risorse per il dispacciamento.

Quest'ultima attività è quella rispetto cui Terna è chiamata a intervenire su un orizzonte temporale più lungo. Si tratta infatti di disporre gli investimenti sulla rete elettrica, necessari per garantire la sicurezza e la continuità degli approvvigionamenti e migliorare la qualità del servizio, riducendo le congestioni. L'infrastruttura di trasmissione deve, infatti, evolvere continuamente per integrare lo sviluppo delle fonti rinnovabili, ridurre le perdite di energia in rete e permettere un avanzamento nei processi di decarbonizzazione accordati a livello europeo.

Ogni anno, Terna definisce il Piano di Sviluppo della rete, relativo a tutti i progetti di sviluppo dei 10 anni successivi, selezionati sulla base di un attento percorso di analisi e valutazione, includendo informazioni sullo stato di avanzamento dei progetti già avviati. In riferimento al Piano di Sviluppo 2021[16], i nuovi interventi programmati sono rappresentati in sintesi in Figura 1.6, tra cui compare la realizzazione del "Tyrrhenian Link", rappresentato più nello specifico in Figura 1.7.

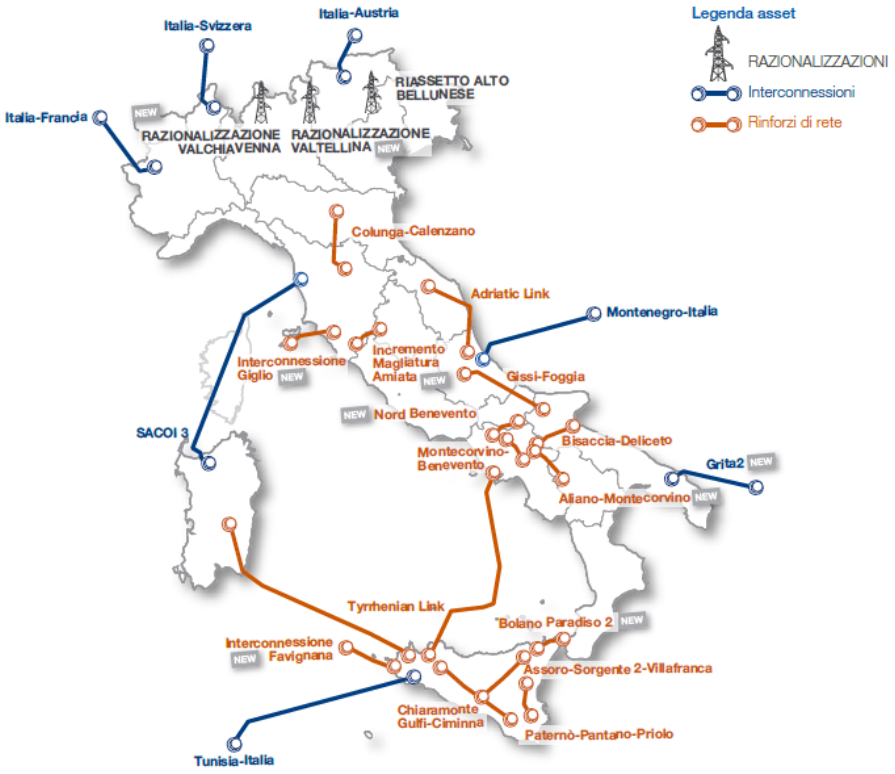


Figura 1.6: Interventi sulla RTN programmati da Terna e annunciati nel Piano di Sviluppo 2021. Fonte[17]

Il progetto è volto alla creazione di due collegamenti sottomarini (Est dalla Sicilia alla penisola e Ovest dalla Sicilia alla Sardegna), da 500 MW ciascuno in corrente continua per una lunghezza totale di 950 km: l'opera, per cui Terna investirà circa 3,7 miliardi, consentirà l'integrazione delle fonti rinnovabili e favorirà il processo di decarbonizzazione del

[16] Terna (2021b)
[17] Terna (2021b)

sistema elettrico sardo. In tal senso, era stato progettata anche la più grande infrastruttura ad oggi realizzata e funzionante in Italia, "SA.PE.I" (il collegamento Sardegna - Penisola Italiana). Attualmente quest'opera ha reso la rete elettrica della Sardegna più stabile dal punto di vista della regolazione di frequenza in condizioni di normale esercizio. SA.PE.I, raffigurato schematicamente in Figura 1.8, è un'infrastruttura da record: collocato a 1640 metri sotto il livello del mare, è il cavo elettrico sottomarino più profondo al mondo, e con i suoi 435 km di lunghezza, il più lungo cavo sottomarino da 1000 MW del pianeta.

Figura 1.7: Le due tratte, Est e Ovest, del Tyrrhenian Link. Fonte[18]

Figura 1.8: La tratta del collegamento SA.PE.I. Fonte[19]

[18] Terna (2022f)
[19] Terna (2021f)

Terna, in qualità di gestore della Rete di Trasmissione Nazionale e responsabile delle suddette funzioni, ha pertanto un importante ruolo nel controllo di un numero crescente di attori, sia dal lato della produzione che della domanda, in seguito all'apertura del mercato elettrico su entrambi i fronti.

Al contrario, le attività di trasmissione e dispacciamento hanno subito un percorso di accentramento: dopo il processo di *unbundling*, è stato subito ritenuto più idoneo concedere a un solo soggetto un ruolo attivo in questi settori, riconoscendo il monopolio naturale che li governa e dando la possibilità, a più fornitori, di poter intervenire solo nella fase successiva di trasporto dell'energia elettrica.

Sulla base di ciò nel 1999 nasce il Gestore della Rete di Trasmissione Nazionale (GRTN) per la gestione dell'operatività delle infrastrutture di rete, le attività di trasmissione e dispacciamento, e Terna, come società proprietaria della concessione di tali infrastrutture. Proprietà e gestione della Rete di Trasmissione Nazionale sono separate, passando dal monopolio di Enel ad un modello *Independent System Operator*, che permette a tutti i produttori di avere lo stesso trattamento nell'accesso alla rete.

Nel 2004 Terna diviene completamente autonoma da Enel e scompare la necessità di una gestione terza della rete; per garantire l'autonomia operativa di Terna come gestore della Rete di Trasmissione Nazionale, il Ministero dell'Economia e delle Finanze acquista tramite la CDP (Cassa Depositi e Prestiti) il 29,99% del capitale di Terna. L'anno successivo il GRTN perciò cambia le proprie funzioni diventando Gestore dei Servizi Elettrici (GSE). Parallelamente si assiste alla nascita di ulteriori attori quali il GME (Gestore del Mercato Elettrico), AU (Acquirente Unico) e RSE (Ricerca sul Sistema Energetico), descritti al meglio in seguito.
La remunerazione dei servizi di trasmissione e dispacciamento assicurati da Terna si basa su un sistema tariffario stabilito dall'Autorità di Regolazione per Energia, Reti e Ambiente (ARERA) attraverso specifiche delibere e non attraverso un meccanismo di formazione del prezzo sul mercato. La remunerazione di Terna (5,6%) è la più bassa nel settore energetico italiano ed è ben al di sotto della media europea (6,4%).

Con la separazione della filiera è stato necessario stabilire una pianificazione integrata per l'adeguamento della rete di trasmissione, in precedenza gestita e coordinata, insieme alla capacità di produzione, da un unico soggetto (Enel). Per quanto riguarda l'attività di trasporto e distribuzione dell'energia elettrica, come discusso nel punto seguente, è tuttora concentrata nelle mani di pochi operatori per motivi sia economici sia di sicurezza del sistema elettrico nazionale.

1.1.3 La distribuzione

Nell'ambito della distribuzione dell'energia, ovvero del trasporto all'utente finale dell'elettricità, operano società in regime di concessione. L'attività consiste nel trasporto dell'energia dalle cabine primarie, che trasformano l'alta in media tensione, alle cabine secondarie, che portano l'elettricità a una bassa tensione, e nella manutenzione della sezione locale di rete. La distribuzione di energia ha subito una suddivisione su base geografica, con delle ripartizioni su scala regionale e, talvolta, provinciale e viene gestita sulla base di monopoli naturali locali. Ciò ha fatto sì che molte regioni abbiano visto la nascita di più soggetti sul territorio per la distribuzione e che grandi municipalizzate locali abbiano dovuto svolgere delle divisioni delle proprie attività.

Per quanto concerne la vendita si ha avuto, invece, il maggiore cambio. La liberalizzazione del mercato ha infatti concesso a numerosi operatori di potersi presentare sul mercato, con un progressivo passaggio che ha portato alla divisione tra il Servizio di Maggior Tutela e il Mercato Libero.

1.1.4 Il *metering*

Il compito principale del *metering* consiste nell'acquisire, elaborare e convalidare le misure di energia scambiata sulla Rete Rilevante, che comprende la Rete di Trasmissione Nazionale (RTN), l'interconnessione con l'estero e le reti di distribuzione in alta tensione direttamente connesse alla RTN, e su altre reti riguardanti le produzioni di interesse di Terna. L'obiettivo è capire quanta energia è stata immessa o prelevata da ciascun operatore, per le finalità di previsione del fabbisogno da parte di Terna.
L'acquisizione delle misure avviene:

- in maniera diretta, per quanto riguarda l'energia immessa e prelevata dagli impianti di produzione e di prelievo connessi alla Rete Rilevante, mediante telelettura dei contatori; le misure rilevate vengono poi convalidate ed elaborate dal sistema di misura,
- in maniera indiretta, per quanto riguarda le misure relative ai Sistemi di Distribuzione Chiusa (SDC), acquisite e convalidate dai Sistemi di Acquisizione Secondari (SAS) e solo successivamente elaborate dal Sistema di Misura[20].

[20] Terna (2022c)

1.2 Soggetti del mercato italiano

Prima di procedere nella descrizione delle strutture e delle dinamiche del mercato, occorre presentare i principali soggetti che operano nella borsa e nella filiera elettrica italiana, distinguendo tra attori istituzionali e non.

1.2.1 Soggetti non istituzionali

- Il produttore: "persona fisica o giuridica che produce energia elettrica indipendentemente dalla proprietà dell'impianto" (art. 2 comma 18 d.Lgs 79/99)[21]. Il produttore può generare energia per autoconsumo, cioè per coprire le proprie necessità in misura non inferiore al 70% annuo, venendo definito in tal caso "autoproduttore" (art. 2 comma 2 d.Lgs 79/99) o per venderla sul mercato;
- gli importatori: i primi 10 in Italia, al 2020, sono Eni, Edison, Enel Global Trading, Shell Energy Europe Limited, DXT Commodities Sa, Gunvor International BV, A2A, Gazprom Italia, Enet Energy Sa, Hera Trading [22];
- i clienti: "sono le imprese o società di distribuzione, gli acquirenti grossisti e gli acquirenti finali di energia elettrica" (art. 2 comma 2 D. Lgs 79/99). Tra questi pertanto rientrano i distributori, i fornitori e i consumatori finali.
- i distributori: si occupano del trasporto e della consegna al cliente finale attraverso le reti di distribuzione dell'energia elettrica a media e bassa tensione. Occupandosi della rete in regime di monopolio naturale, sono responsabili dell'allacciamento degli utenti, sia idonei, sia vincolati, e del servizio di misura.
- i fornitori, o società di vendita: si occupano della vendita al dettaglio al cliente finale. Controparte del contratto di fornitura con il cliente finale, si occupano di garantire la somministrazione alle condizioni stabilite dal contratto, della fatturazione e dell'assistenza al cliente. Possono non essere produttori diretti e dunque acquistare l'energia o il gas all'ingrosso per poi rivenderla al cliente finale.

[21] *Decreto Legislativo 16 marzo 1999, n. 79* (1999)
[22] Arera (2022)

- i clienti idonei: indicati come "persona fisica o giuridica che ha la capacità [...] di stipulare contratti di fornitura con qualsiasi produttore, distributore o grossista, sia in Italia che all'estero" (art. 2 comma 6 d.Lgs 79/99). In altre parole, i grandi consumatori, cui consumo annuo supera i 40 GWh (soglia progressivamente ridotta già a partire dal d.Lgs 79/99), che possono scegliere con chi stipulare contratti di fornitura.

- i clienti vincolati, non rientrando nella categoria dei clienti idonei, possono "stipulare contratti di fornitura esclusivamente con il distributore che esercita il servizio nell'area territoriale dove è localizzata l'utenza" (art. 2 comma 7 d.Lgs 79/99). Sono principalmente le utenze domestiche. La distinzione tra clienti idonei e vincolati, introdotta originariamente dalla Direttiva Europea 96/92/CE, viene annullata solo nel 2007: "a decorrere dal 1° luglio 2007 è cliente idoneo ogni cliente finale di energia elettrica", accelerando così la liberalizzazione dei mercati [23].

- i traders, o operatori del mercato elettrico, sono soggetti privati o giuridici operanti nel Mercato dell'Energia, comprando, vendendo o scambiando energia elettrica. All'interno di questo mercato sono presenti ad oggi fornitori del mercato libero (come Eni, Enel Energia, Sorgenia, Alpiq, il Gruppo Hera, ecc.), gruppi di acquisto, grossisti e cooperative. I primi 10 trader italiani per vendita di energia elettrica sono: Enel, con 85.723 GWh venduti, Edison (14.165 GWh) e A2A (13.211 GWh). Seguono Hera, Axpo Group, Eni, Acea, E.ON, Duferco, Alperia[24].

1.2.2 Soggetti istituzionali

Il Gestore dei Servizi Energetici (GSE), interamente partecipato dal Ministero dell'Economia e delle Finanze, in qualità di capo gruppo, coordina le società Acquirente Unico (AU), Gestore dei Mercati Energetici (GME) e Ricerca sul Sistema Energetico (RSE). Il GSE rappresenta l'evoluzione del Gestore della Rete di Trasmissione Nazionale (GRTN), in seguito alla cessione dell'attività di gestione della rete di trasmissione a Terna S.p.A nel 2005 e l'acquisizione di RSE, ex CESI Ricerca, nel 2010. Tra le sue principali funzioni

[23] *Legge del 23 agosto 2004, n.239* (2004)
[24] Arera (2022)

vi è l'incentivazione economica della produzione di energia da fonti rinnovabili e la gestione dei flussi finanziari ad essa associati. Il GSE, come promotore dell'efficienza e della sostenibilità energetica, è anche responsabile dell'emissione dei "certificati verdi", titoli negoziabili corrispondenti ad una certa quantità di emissioni di CO_2, ottenibili su richiesta dei produttori di energia da fonti rinnovabili. Tali certificati vengono rivenduti a prezzo di mercato alle imprese che producono energia da fonti fossili, che li acquistano per poter raggiungere una soglia, imposta per legge e sempre crescente dal 2004, della propria produzione.

Tra le società controllate dal GSE, il Gestore dei Mercati Energetici (GME) gestisce il mercato dell'energia elettrica (IPEX, *Italian Power Exchange*), ambientale e del gas naturale, secondo criteri di trasparenza ed obiettività, al fine di promuovere la concorrenza tra i produttori assicurando la disponibilità di un adeguato livello di riserva. Gestisce il Mercato a Termine (MTE) e nel ruolo di "gestore del mercato elettrico designato" (NEMO, *Nominated Electricity Market Operator*) coordina il Mercato del Giorno Prima (MGP) e il Mercato Infragiornaliero (MI) per l'Italia, ma anche il Mercato per la negoziazione di Prodotti Giornalieri (MPEG) e il Mercato dei Servizi di Dispacciamento (MSD), in cui agisce per conto di Terna S.p.A. come controparte nelle transazioni. Sempre nel ruolo riconosciutogli di gestore designato, il GME è membro fondatore del PCR, il progetto *Price Coupling of Regions* promosso dai maggiori PXs (*Power EXchanges*) europei per definire una soluzione tecnica per l'integrazione dei mercati europei del giorno prima.
Il GME inoltre organizza i Mercati per l'Ambiente (con il mercato dei certificati verdi, dei Titoli di Efficienza Energetica (TEE), e delle Unità di Emissione (EU-ETS, *Emission Trading System*)) ed ha assunto la gestione della piattaforma P-GAS, per la gestione dell'obbligo di offerta per legge di una quota delle importazioni di gas prodotto in paesi non europei. Il GME opera in base alle previsioni regolatorie delle Autorità di Regolazione per l'Energia, le Reti e l'Ambiente (ARERA).

La seconda società partecipata al 100% dal GSE è l'Acquirente Unico, a cui è affidato il ruolo di garante, per legge, della fornitura di energia elettrica e gas, provvedendo alla disponibilità necessaria per far fronte alla domanda di tutti i Clienti del "mercato di salvaguardia" e del "mercato di maggior tutela", cioè per i clienti che ancora non hanno scelto un fornitore sul libero mercato. Il compito dell'Acquirente Unico è quello di acquistare energia elettrica alle condizioni più favorevoli sul mercato e di cederla alle imprese di vendita al dettaglio per rifornire gli utenti domestici e le PMI (con meno di 50 dipendenti

e fatturato non superiore ai 10 milioni di euro) che non acquistano sul mercato libero. L'Acquirente Unico può acquistare energia sulla borsa elettrica o attraverso contratti bilaterali. I prezzi di vendita del "mercato tutelato" sono stabiliti dall'Autorità di Regolazione per l'Energia e l'Ambiente.

Terza e ultima partecipata al 100% dal GSE è la società di Ricerca sul Sistema Energetico (RSE), che si rivolge principalmente all'innovazione e al miglioramento delle prestazioni del sistema elettrico, dal punto di vista dell' economicità, della sicurezza e della sostenibilità.

L'Autorità di Regolazione per l'Energia, le Reti e l'Ambiente (ARERA) è un organismo indipendente istituito con la legge numero 481 del 1995 di liberalizzazione del mercato dell'energia e del gas. Svolge la fondamentale attività di regolazione e di controllo negli ambiti dell'energia elettrica e del gas naturale, oltre che nei settori dei servizi idrici, del ciclo dei rifiuti e del telecalore.
La regolazione delle tariffe nel mercato tutelato, l'accesso alle reti, il funzionamento del mercato e la protezione dei consumatori sintetizzano le principali attività di cui si occupa l'Autorità. In particolare, l'ARERA aggiorna con cadenza trimestrale le condizioni economiche di riferimento per i clienti che non hanno ancora scelto il mercato libero nei settori energetici e promuove l'uso razionale dell'energia, incentivando la diffusione dell'efficienza energetica e l'adozione di misure per uno sviluppo sostenibile. Inoltre, stabilisce, per i settori energetici, le tariffe per l'utilizzo delle infrastrutture e predispone ed aggiorna il metodo tariffario per la determinazione dei corrispettivi per il servizio idrico integrato e per il servizio integrato dei rifiuti. Spetta a questo organismo definire i livelli minimi di qualità dei servizi per gli aspetti tecnici, contrattuali e per gli standard di servizio ed accrescere i livelli di tutela, di consapevolezza e l'informazione ai consumatori.

Terna - Rete Elettrica Nazionale S.p.A, infine, è la società proprietaria della Rete di Trasmissione Nazionale ad alta ed altissima tensione. È il primo operatore di rete indipendente d'Europa e tra i principali al mondo per chilometri di linee, con 74.855 km (al 31 dicembre 2021, in aumento dello 0.2% rispetto al 2020) di linee elettriche ad alta e altissima tensione gestite attraverso la controllata diretta Terna Rete Italia S.p.A. Le infrastrutture di cui Terna detiene proprietà comprendono anche 896 stazioni di trasformazione e smistamento (7 in più attivate rispetto al 2020), 1 Centro Nazionale di Controllo (CNC), 3 centri di teleconduzione, 26 linee di interconnessione con i paesi confinanti (Francia,

Svizzera, Austria, Slovenia, Montenegro, Grecia), e sistemi di HVDC (*High Voltage Direct Current*) per l'interconnessione della rete peninsulare italiana con altre reti – nazionali e internazionali. L'opera con tecnologia HVDC più recente è SA.PE.I, che raggiunge la profondità record di 1.640 metri sotto il livello del mare, collega Sardegna e la Penisola Italiana, con una lunghezza di 435 km e una potenza di 1.000 MW.

Dal 2005 Terna unifica in sè i due ruoli di *Trasmission* e *System Operator*. In qualità di *Trasmission Operator* (TO), si occupa del trasporto dell'energia elettrica in alta tensione dai punti di produzione a quelli di distribuzione, coordinando gli interventi di sviluppo della rete elettrica e la manutenzione. In qualità di *System Operator* (SO), Terna è responsabile del servizio di dispacciamento nel sistema elettrico nazionale, garante del bilanciamento continuo tra la domanda di energia, da parte del consumatore, e l'offerta, ovvero l'energia prodotta. Con il fine ultimo di mantenere l'equilibrio di rete, Terna si pone come coordinatore centrale con potere di controllo su entrambi i fronti, della produzione e della domanda.
É con tale ruolo che Terna opera nel Mercato per il Servizio di Dispacciamento (MSD), dove si approvvigiona dei necessari servizi per garantire la sicurezza e l'adeguatezza del sistema, i cosiddetti *servizi ancillari* o *ausiliari*, con il fine ultimo del bilanciamento in tempo reale. Essi riguardano:

- la risoluzione di congestioni intrazonali,
- la regolazione di frequenza, relativa all'attivazione della riserva per l'equilibrio della rete e della tensione,
- l'approvvigionamento di riserva statica, relativa alla creazione di una certa banda di capacità, che garantisca l'operabilità entro range di sicurezza.

Per comprendere la differenza tra le due funzioni relative alla riserva, basti pensare ad un generatore di 500 MW, portato alla produzione di 400 MW: ne restano 100 MW liberi, che costituiscono la riserva. L'attivazione di riserva per 50 MW, che porta la generazione a 450 MW, costituisce la cosiddetta regolazione.

Il Mercato per il Servizio di Dispacciamento in cui opera Terna, rientra nel Mercato a Pronti, che comprende, oltre all'MSD, il Mercato dell'Energia, ovvero l'insieme di Mercato del Giorno Prima, Mercato Infragiornaliero e Mercato dei Prodotti Giornalieri, gestiti

dal NEMO (*Nominated Electricity Market Operators*) italiano, il GME.
A valle, il mercato elettrico si distingue, in base alla tipologia di contrattazione che intercorre tra le parti, in:

- Mercato Elettrico "a Pronti", o mercato "spot", essenzialmente riconducibile alla Borsa elettrica, in cui ogni operatore ha come controparte il mercato stesso, e ricerca l'incontro tra domanda e offerta;
- Mercato Elettrico "a Termine", o mercato "forward", con obbligo di consegna e ritiro, in cui sono stipulati accordi grazie all'incontro di due parti e i prodotti scambiati sono tendenzialmente standardizzati.

La struttura generale del Mercato Elettrico, è rappresentata graficamente in Figura 1.9.

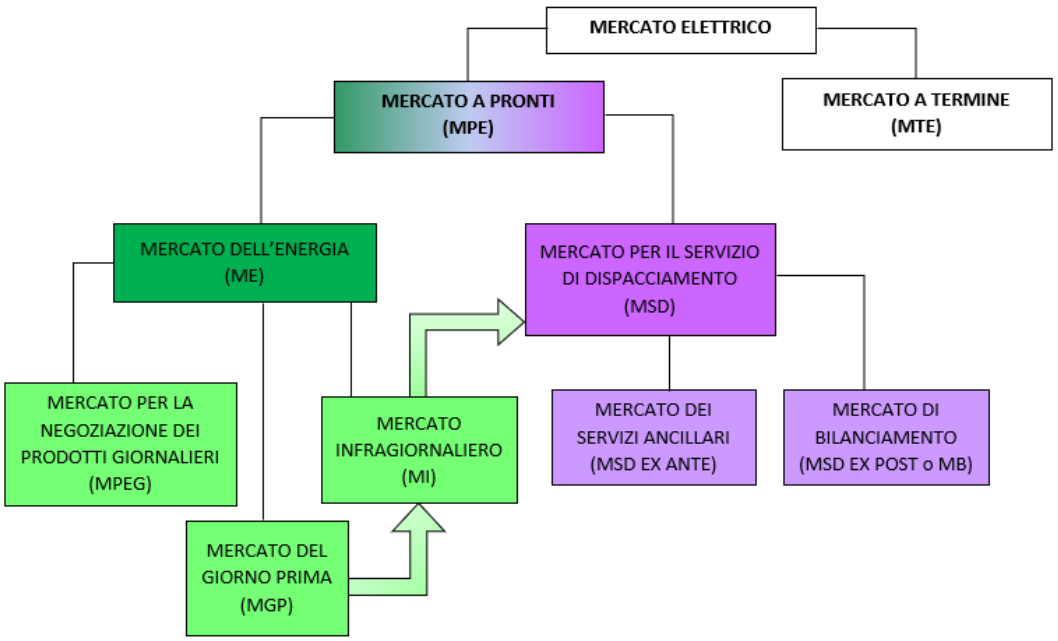


Figura 1.9: Struttura del mercato elettrico italiano.

1.3 Mercato elettrico a pronti

Sebbene in tutti i Paesi occidentali i Mercati a Pronti, o "spot", siano liberalizzati, in realtà il disegno del modello di mercato non è così uniforme a livello internazionale, e in particolare sono due i modelli adottati: il modello integrato e il modello a borsa elettrica.

1.3.1 Principali modelli di MPE

Modello integrato

Il modello integrato è il modello adottato nei mercati liberalizzati degli Stati Uniti. Rappresenta sostanzialmente un'evoluzione del modello a borsa elettrica, adottato in Europa, e in principio anche negli Stati Uniti. Il gestore del sistema elettrico (*System Operator*, SO) ottimizza lo *unit commitment* e il dispacciamento a livello centralizzato, tenendo conto di tutti i vincoli del sistema elettrico in ogni fase. Energia e servizi ancillari sono procurati simultaneamente e co-ottimizzati, in un modello che integra la risoluzione del mercato con la verifica della realizzabilità dei programmi accordati. Per fare ciò, la rete di trasmissione viene rappresentata in modo dettagliato, in ogni singolo nodo e linea, in tutti gli algoritmi di *clearing*, il processo di calcolo delle obbligazioni reciproche degli operatori contraenti, e già nelle fasi iniziali di contrattazione. I mercati costruiti su questo modello sono detti *mercati nodali*, in quanto facenti riferimento al valore dell'energia elettrica nei singoli nodi di rete. I prezzi adottati sono detti "prezzi nodali" (*Locational Marginal Prices*, LMP).

Modello a borsa elettrica

Nel modello a borsa elettrica, adottato nel mercato europeo, solo i principali vincoli ("strutturali") sulla capacità di trasporto della rete vengono riflessi, e sulla base di questi la rete viene suddivisa in zone di mercato ("*bidding zones*"), tra le quali lo scambio è condizionato ai limiti fisici di transito con le zone confinanti. A differenza del modello integrato infatti, la rappresentazione della rete elettrica implementata negli algoritmi è piuttosto

semplificata e limitata alla suddivisione in zone, consentendo di evidenziare solo le principali congestioni. Eventuali disequilibri vanno risolti pertanto in mercati successivi.
La determinazione dei prezzi, detti "prezzi zonali", avviene attraverso un meccanismo di aste in ogni zona di mercato. In questi *mercati zonali* energia e servizi ancillari sono procurati in modo sequenziale: l'approvvigionamento di questi ultimi avviene sulla base di aste a termine dedicate ai singoli servizi, o in appositi mercati "spot", come il Mercato per il Servizio di Dispacciamento, in Italia.

1.3.2 Suddivisione in zone di mercato

Dal 1° gennaio 2021 la configurazione delle zone di mercato italiane è cambiata come esito di un processo pluriennale di revisione, iniziato nel 2015. In ottemperanza ai nuovi criteri individuati dal regolamento EU CACM, riassumibili in sicurezza, efficienza del mercato e robustezza, l'Italia è il primo Paese europeo a vedere l'entrata in vigore del nuovo paradigma. Fino al 2018, le zone di mercato in Italia erano dieci: 6 zone "geografiche" e 4 "virtuali". L'ampliamento della capacità di trasporto della rete e la concomitante dismissione della capacità di generazione, che ha visto, solo nel 2020, la dismissione di circa 1,2 GW di capacità termoelettrica, ha portato alla chiusura dei quattro poli di produzione, le quattro "zone virtuali", di Priolo, Foggia, Brindisi e Rossano (quest'ultimo nel 2021). Per riflettere, invece, gli effetti dell'aumento concomitante della generazione rinnovabile sui flussi di potenza, è stata introdotta una nuova zona (Calabria). Ad oggi le zone di mercato in Italia sono pertanto sette, tutte classificate come zone "geografiche", per via della corrispondenza territoriale con le regioni italiane e loro raggruppamenti (Fig. 1.10):

- Nord (NORD), comprendente Val D'Aosta, Piemonte, Liguria, Lombardia, Trentino, Veneto, Friuli Venezia Giulia, Emilia Romagna;
- Centro Nord (CNORD), comprendente Toscana, Marche;
- Centro Sud (CSUD), comprendente Lazio, Abruzzo, Campania, Umbria;
- Sud (SUD), comprendente Molise, Puglia, Basilicata;
- Sicilia (SICI);
- Sardegna (SARD);

- Calabria (CALA).

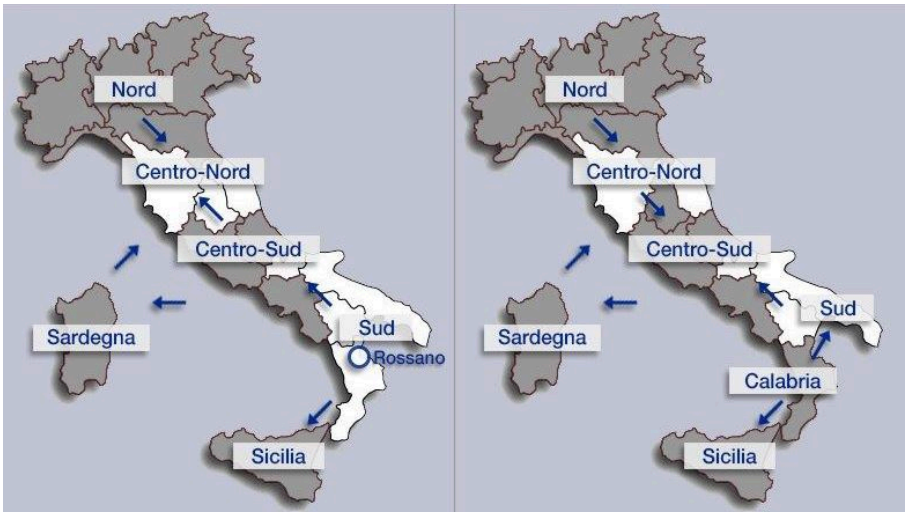


Figura 1.10: A sinistra le zone del mercato elettrico in vigore fino al 31 dicembre 2020 (con il polo di produzione di Rossano in provincia di Catanzaro), a destra la nuova configurazione zonale. Fonte[25]

Accanto alle zone di mercato in cui è suddiviso il territorio nazionale, sono definite delle zone di mercato per i Paesi confinanti, dette zone "virtuali estere": a differenza della suddivisione in zone del mercato italiano, presente anche in Norvegia e Svezia, per tutti i Paesi europei, i confini di mercato coincidono con i confini nazionali, all'interno dei quali vige il libero scambio, limitato solo con l'estero. Vengono pertanto integrati in questo modo negli algoritmi di risoluzione di mercato, in cui vanno considerati tutti i flussi di energia, per poter definire correttamente l'equilibrio. Tra queste l'Austria (AUST), la Corsica (CORS), la Grecia (GREC), la Francia (FRAN) e le varie altre nazioni europee collegate all'Italia, oltre alle zone rappresentative dell'interconnessione dedicata al*market coupling*, come Francia Coupling (XFRA), Austria Coupling (XAUS), Slovenia Coupling (BSP) e la più recente Grecia Coupling (XGRE).
Il *market coupling* è un meccanismo di integrazione dei mercati che permette di evitare di separare l'acquisto della capacità di trasporto dalla compravendita di energia elettrica. Grazie ad un algoritmo che accoppia tutti i sistemi, determina il valore dell'energia elettrica

[25] Terna (2022a)

nelle varie zone europee (il "*clearing*" delle offerte di acquisto e di vendita di energia), allocando contestualmente la capacità giornaliera di transito sulla frontiera, con l'obiettivo di massimizzare il surplus economico complessivo dei partecipanti al mercato e incrementare il benessere sociale.

È aperto il dibattito su quale sia la struttura migliore per il mercato dell'energia. La suddivisione in zone in Italia è stata sin dal principio adottata, sia per conformazione geografica, che per differenziare i prezzi di acquisto a seconda del bilancio tra capacità di generazione di energia elettrica e domanda, che varia da zona a zona (fornendo opportuni "segnali di prezzo"). Tale scelta ha proprio per questo delle conseguenze sul prezzo, e potrebbe potenzialmente riversarsi sul consumatore finale. Per livellare le differenze di prezzo tra zone, il prezzo per l'energia adottato come costo finale per il consumatore è il cosiddetto PUN (Prezzo Unico Nazionale), pari alla media pesata dei prezzi zonali (sul lato offerta) ponderata per le quantità acquistate in tali zone. In questo modo, un costo maggiore per l'energia in alcune zone viene distribuito a livello nazionale: al Sud, dove c'è più offerta, soprattutto grazie alla grande produzione da fonti rinnovabili, i prezzi sarebbero inferiori alla Sicilia, un sistema isolato in cui gran parte della domanda viene soddisfatta grazie allo scambio con la Calabria. Un prezzo unico a livello nazionale fa sì che laddove c'è maggiore offerta e i prezzi sarebbero naturalmente inferiori, si paghi invece un po' di più, contribuendo in parte al maggiore prezzo da pagare al produttore in Sicilia, in modo che il differenziale non venga scaricato solo sulle spalle del consumatore dell'isola. Inoltre, al produttore che volesse dar vita un nuovo impianto conviene farlo dove c'è meno offerta, perché ci guadagna di più; ma così facendo migliora anche la quantità di elettricità offerta nella zona e quindi l'efficienza generale del sistema elettrico.

La suddivisione in zone si riflette nella suddivisione del mercato dell'energia, a partire dal Mercato del Giorno Prima, che conserva la struttura zonale, fino al Mercato dei Servizi di Dispacciamento, in cui nell'ottica dell'unitarietà del mercato elettrico nazionale, si implementano i limiti di scambio tra zone. Per questo il processo di definizione delle nuove zone di mercato ha richiesto un attento lavoro di revisione metodologica: è ciò che sta alla base di un'impostazione efficiente del mercato. L'obiettivo è, infatti, quello di cogliere correttamente l'andamento dei principali flussi di potenza, a seconda delle condizioni di domanda e offerta, identificando sezioni critiche per la trasmissione e fissandone i limiti di transito. L'eliminazione di questi "colli di bottiglia" nella capacità di trasporto della rete è compito che Terna porta avanti grazie allo sviluppo della rete.

1.3.3 Il Mercato del Giorno Prima

Il Mercato del Giorno Prima (MGP) è la sessione di mercato in cui in Italia vengono scambiati la maggior parte dei volumi di energia venduti e acquistati: è da questo mercato che, pertanto, si colgono segnali di variazioni dei prezzi. Si basa su un meccanismo ad asta, dove si contrattano blocchi orari di energia, per ciascuna ora del giorno successivo, già dalle ore 8 del nono giorno antecedente la consegna e fino alle 12:00 del giorno precedente. La comunicazione degli esiti del MGP avviene entro le ore 12.58 del giorno precedente il giorno di consegna[26].

Le 24 ore rappresentano il *market time unit* italiano, ovvero il periodo di tempo per il quale vengono costruite le curve di domanda e di offerta e quindi definito il prezzo di mercato, o il più breve se, nel confronto tra diverse zone di mercato, è differente[27]. Questo lasso temporale assume, inoltre, una certa rilevanza dal momento che nel mercato italiano non è possibile effettuare offerte "complesse", ma solo "semplici", con base oraria. Ciò significa che per ogni offerta sulle 24 ore è come se si effettuassero 24 offerte, una per ora, ognuna indipendente dall'altra, che possono essere accettate in un'ora e non in quella successiva. Si tratta chiaramente di una semplificazione, che non tiene conto della reale attuabilità del programma, perché, ad esempio, un impianto a carbone non può essere funzionante in un'ora, spento in quella successiva e poi riacceso ancora, come se il suo avvio fosse istantaneo. Sarà compito delle sessioni di mercato successive correggere ciò. Nel resto d'Europa vale invece la possibilità di fare offerte complesse, come ad esempio le *bloc bids*, per cui se vengono accettate, devono garantire il servizio per almeno un certo numero di ore, o le *minimum income*, per cui se vengono accettate, devono garantire che il ricavo in questo periodo di tempo sia per almeno un certo valore.

IL MECCANISMO DI FORMAZIONE DEI PREZZI

Le contrattazioni, nelle quali il GME agisce come controparte centrale, avvengono tra operatori in vendita, rappresentanti dei produttori di energia, che presentano offerte circa la quantità e il prezzo minimo al quale sono disposti ad vendere, e operatori in acquisto, rappresentanti delle società di vendita al consumatore finale, che presentano offerte circa la

[26] GME (2022)

[27] *Commission Regulation (EU) No 543/2013 of 14 June 2013* (2013)

quantità e il prezzo massimo al quale sono disposti ad acquistare. Il punto di incontro tra domanda e offerta è selezionato da EUPHEMIA (NEMOCommittee, 2019), acronimo di *Pan-European Hybrid Electricity Market Integration Algorithm*, l'algoritmo centrale di accoppiamento dei prezzi a livello europeo sul mercato del giorno prima.

Questo meccanismo si inserisce nell'iniziativa PCR (*Price Coupling of Regions*)[28], avviata da otto borse elettriche europee per lo sviluppo di un'unica soluzione di accoppiamento dei prezzi, o *price coupling*, armonizzata su tutta l'area UE, con la creazione di un mercato unico del giorno prima a livello europeo, chiamato *Single Day Ahead Coupling* (SDAC) e di un mercato unico infragiornaliero, il *Single Intraday Coupling* (SIDC). I TSO delle frontiere italiane, in cooperazione con le rispettive borse di energia, hanno implementato il *market coupling* sui confini italiani, in base a quanto previsto dal target model europeo per l'allocazione giornaliera della capacità di interconnessione. Nel febbraio 2015 l'Italia, attraverso la Francia, Austria e Slovenia, e dal 2020 anche attraverso la Grecia, entra nel mercato elettrico unificato dal progetto *Multi-Regional Coupling* (MRC), e il GME diventa membro del PCR. Un mercato elettrico integrato su scala europea accrescerà la liquidità, l'efficienza e il benessere sociale.

Ad oggi, l'obiettivo può essere considerato raggiunto, con l'inclusione nel SDAC della quasi totalità dei Paesi europei, anche in seguito all'uscita dall'Inghilterra nel Gennaio 2021 (Fig. 1.11). Il passo finale, l'inclusione del confine tra Croazia e Ungheria nel SDAC *coupling*, in giallo nella Figura 1.11 è avvenuto il 20 Aprile 2022 con il go-live del Core Flow-Based Market Coupling (Core FB MC)[29].

[28] EPEXSPOT (2021)

[29] HUPX (2022)

Figura 1.11: Stato attuale dell'implementazione del PCR. Fonte[30]

L'algoritmo EUPHEMIA, di cui una descrizione dettagliata è riportata in NEMOCommittee (2019), calcola simultaneamente le quantità accettate per ogni offerta sottomessa, i prezzi marginali zonali e le posizione nette, ottimizzando il social welfare, cioè massimizzando la felicità collettiva. Dal punto di vista del consumatore, ciò si ottiene minimizzando il prezzo che paga, dal punto di vista del produttore, massimizzando il prezzo che gli viene riconosciuto, massimizzando complessivamente le quantità accettate nel mercato; con benessere sociale si fa, infatti, riferimento alla somma del surplus del consumatore e del surplus del produttore.
Le offerte sono accettate dopo la chiusura della seduta di mercato per ogni *market time unit*, cioè ogni 24 ore, e ordinate sulla base del merito economico e nel rispetto dei limiti di transito tra le zone. La curva aggregata delle offerte di vendita, che raccoglie le offerte dal lato produzione, ha una pendenza positiva, dall'offerta più economica (la più a sinistra), a crescere, come mostrato in Figura 1.12.

[30] NEMOCommittee (2022)

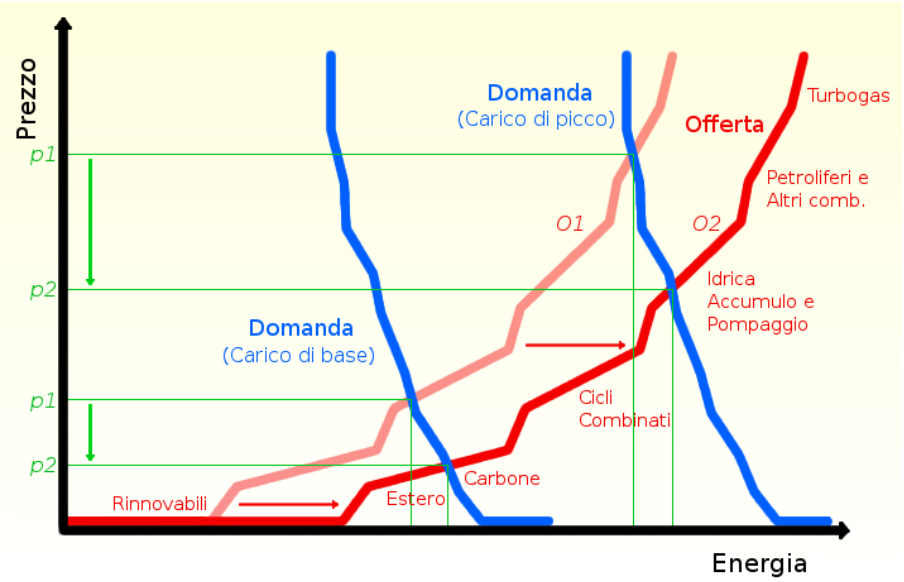


Figura 1.12: Formazione dei prezzi nel Mercato del Giorno Prima, con effetto sul prezzo di una maggiore produzione rinnovabile. Fonte[31]

L'"ordine di merito" fa riferimento al fatto che la disposizione delle offerte rispecchia i costi di generazione associati alle diverse tecnologie. Gli impianti da fonte rinnovabile sono caratterizzati da costi marginali molto bassi, quasi nulli, e tendono ad essere le fonti selezionate per prime, con alta priorità, sottraendo spazio ed escludendo dal mercato le fonti a più alti costi marginali. Gli impianti nucleari hanno un costo marginale basso e andrebbero a collocarsi immediatamente dopo le fonti rinnovabili, dove presenti; in Italia, l'energia prodotta da questi impianti è importata e risulta ancora abbastanza conveniente. A salire di prezzo, ci sono gli impianti a carbone (convenienti qualora non siano introdotte significative "carbon tax"), gli impianti a ciclo combinato, molto flessibili rispetto agli altri impianti per far fronte al carico di base, e con efficienza molto elevata, attorno al 50-55%. Procedendo verso destra si sale nei costi arrivando ai pompaggi, attivati al bisogno dagli impianti idroelettrici ad accumulo, per giungere, infine, agli impianti con costi marginali più elevati, come i turbogas semplici o con motori diesel/olio combustibile, che, seppure con un'efficienza del 25-30%, sono spesso utili per la copertura dei carichi di picco.
La rappresentazione schematizzata della curva di offerta di fig, 1.12, mostra anche come l'aumento di energia da fonti rinnovabili comporti una maggiore riduzione del prezzo dell'energia se tale maggiore disponibilità si ha nei momenti di picco della domanda, piut-

[31] Dataenergia (2014)

tosto che durante il carico di base (graficamente ciò è riscontrabile nella differenza $p1 - p2$, in verde). Questo è il motivo per cui il fotovoltaico, che produce energia durante le ore con carico medio e di picco, ha un effetto riduttivo sul prezzo di mercato maggiore rispetto a tecnologie come l'eolico, che invece tende a distribuire la loro produzione lungo tutta la giornata.
Qualunque sia la fonte rinnovabile prevalente, l'ordinamento delle offerte di vendita per merito economico ha incentivato la produzione di energia da impianti più efficienti e da fonti rinnovabili, proprio in virtù della priorità di cui godono, che "spinge" le unità più costose fuori dal mercato.
Analogamente alla formazione della curva dell'offerta, il GME raccoglie le offerte di acquisto di energia, ovvero tutte le quantità richieste dagli operatori in acquisto con il massimo prezzo che sono disposti a pagare per quella quantità, formando la curva aggregata della domanda, con pendenza negativa. Come mostrato in Figura 1.12, la curva di domanda è particolarmente anelastica: l'energia elettrica è infatti uno di quei beni la cui richiesta è scarsamente influenzata dal prezzo, in virtù della sua essenzialità.

Tutte le offerte di vendita e le offerte di acquisto, riferite sia alle unità di pompaggio che alle unità di consumo, appartenenti alle zone virtuali estere che sono accettate sul MGP, vengono valorizzate al prezzo marginale di equilibrio della zona a cui appartengono, il prezzo marginale zonale. Tale prezzo è determinato, per ogni ora, dall'intersezione della curva di domanda e di offerta e si differenzia da zona a zona, in presenza di limiti di transito saturati. In tal senso il MGP adotta una struttura zonale, e si distinguono zone importatrici ed esportatrici (con prezzo zonale più basso).
Il prezzo marginale zonale è il costo di produzione di 1 MWh in più in quella zona. Se nella zona adiacente il costo marginale di produzione è inferiore, sarà preferibile importarlo, a meno che il limite di scambio non sia stato saturato. In tal caso, dovrà essere prodotto internamente, per non incidere sulla congestione, o potrebbe non essere prodotto, se quel MWh in più ha un costo che i consumatori della zona non possono sostenere. Le offerte di acquisto accettate e riferite alle unità di consumo appartenenti alle zone geografiche italiane (ad eccezione quindi del pompaggio e dell'export) sono invece valorizzate al Prezzo Unico Nazionale (PUN). La differenza tra quanto viene incassato dai produttori e l'esborso dei consumatori è definito "rendita da congestione", ed è esattamente pari al prodotto tra il flusso interzonale ed il delta prezzo. La rendita da congestione rappresenta il valore della capacità di trasporto. In quanto tale, viene incamerato da Terna, responsabile del funzionamento della rete, per sovvenzionare nuovi investimenti sull'interconnessione in

modo da aumentare la capacità di scambio. Nella realtà viene restituita subito ai consumatori, perché in Italia gli investimenti sono approvati e pagati dalle Autorità: il settore in questo modo risulta, seppur meno trasparente, più regolamentato, perché le Autorità, facenti capo al Ministero della Transizione Ecologica, riescono a tener conto dei tempi per la realizzazione dei progetti.

Il Mercato del Giorno Prima è a partecipazione volontaria. La scelta di entrarvi viene fatta sulla base di una valutazione di opportunità e capacità operative. Essendo questo il mercato con maggiore volume di scambio, e quindi con maggiore liquidità e competitività, alcuni produttori potrebbero ritenere più conveniente non parteciparvi. I produttori che si avvalgono di questa possibilità sono però una ristretta cerchia, quella dei produttori abilitati alla partecipazione al Mercato per il Servizio di Dispacciamento. Come vedremo in seguito, tale mercato è a partecipazione obbligatoria per le unità produttive "rilevanti", le quali potrebbero quindi valutare la possibilità di offrire solo in MSD. Il soddisfacimento di vincoli di rete con carattere locale, permette di avere meno competizione, al massimo con produttori vicini a livello geografico, quindi meno liquidità e di conseguenza un maggiore potere da parte dei produttori sul prezzo dell'energia. Questo vale soprattutto in quelle ore in cui Terna potrebbe avere bisogno di comprare energia: il produttore abilitato può decidere di non offrire o farlo a prezzi altissimi sul mercato dell'energia, per poi offrire in MSD, dove ha maggiore potere contrattuale.
In generale, non tutte le unità produttive possono optare per questa strategia, perché, come detto, non tutte sono abilitate a partecipare al MSD. Per le UP non abilitate, il principale motivo per cui potrebbero non offrire, o offrire a prezzi molto elevati, tutta la capacità, riguarda il fatto che alcuni operatori, in certe condizioni di stress, possono pilotare il prezzo ed evitare che si abbassi. Nel mercato italiano non c'è un meccanismo di controllo che previene queste situazioni in casi di forte stress, a differenza di quello che succede nel mercato con struttura nodale, dove all'interno degli algoritmi sono implementati criteri oggettivi per definire se qualche operatore sta esercitando potere di mercato e le offerte vengono automaticamente corrette. Nel mercato statunitense l'assunzione è che ogni produttore offre al costo variabile di produzione. Se questo criterio non viene rispettato, l'offerta viene riportata automaticamente al costo marginale con l'aggiunta di una penalità. Nei mercati a zone, come quello italiano, questo automatismo non esiste e vi sono situazioni che riflettono il limite sottile tra scarsità e potere di mercato. Quando c'è scarsità, c'è poca liquidità e i prezzi salgono; quindi, è giusto dare questo segnale perché incentiva nuovi ingressi nel mercato, ma sconfina nell'esercizio di potere di mercato. In

Sicilia sfruttando il potere di mercato che il principale operatore ha sull'isola, il prezzo risulta più alto: non vi sono altri motivi se non questo ai prezzi storicamente più elevati perché la tecnologia siciliana è la stessa di quella della penisola. L'operatore rialza l'offerta perché satura la capacità di importazione.

1.3.4 Il Mercato Infragiornaliero

Il Mercato Infragiornaliero, o *Intraday*, (MI) opera nello schema di mercato italiano inglobato unitamente al Mercato del Giorno Prima e al Mercato per il Servizio di Dispacciamento, a raccordo tra il Mercato dell'Energia e il Mercato dei Servizi. Come il MGP, anche il MI è a partecipazione volontaria.
In esito al Mercato dell'Energia i programmi delle singole unità di produzione sono, in generale, tecnicamente fattibili, ma non generalmente compatibili con la sicurezza e l'adeguatezza del Sistema Elettrico Nazionale. Per assicurare la realizzabilità di questi programmi, garantendo il continuo bilanciamento tra immissioni e prelievi, e tenendo conto dei vincoli tecnici delle unità produttive, in MI gli operatori possono apportare modifiche ai programmi definiti in MGP attraverso ulteriori offerte di acquisto o vendita. Le negoziazioni riguardano quantità di energia che verranno consumate nel giorno stesso; se le correzioni apportate non dovessero risolvere il mercato con prontezza, in caso di eventi imprevisti o previsioni aggiornate, Terna compensa nell'MSD eventuali squilibri, regolando l'energia immessa e prelevata in rete.

La nuova struttura a doppio meccanismo

Fino al mese di Settembre 2021, il Mercato Infragiornaliero era strutturato in sette sessioni (MI1, MI2, ... , MI7), come visibile in Figura 1.13 avviate sequenzialmente a partire dalle 12.55 del giorno precedente la consegna, *d-1* subito dopo la chiusura del MGP, fino alle 17.00 del giorno *d*.

Oggigiorno, le negoziazioni sul MI avvengono con un sistema a doppio meccanismo, basato sullo svolgimento di tre sessioni d'asta complementari regionali CRIDA (CRIDA1, CRIDA2, CRIDA3) e una sessione di negoziazione continua XBID, aperte 24 ore al

giorno, 365 giorno all'anno, agli operatori di mercato di tutta Europa, come riportato in Figura 1.14.

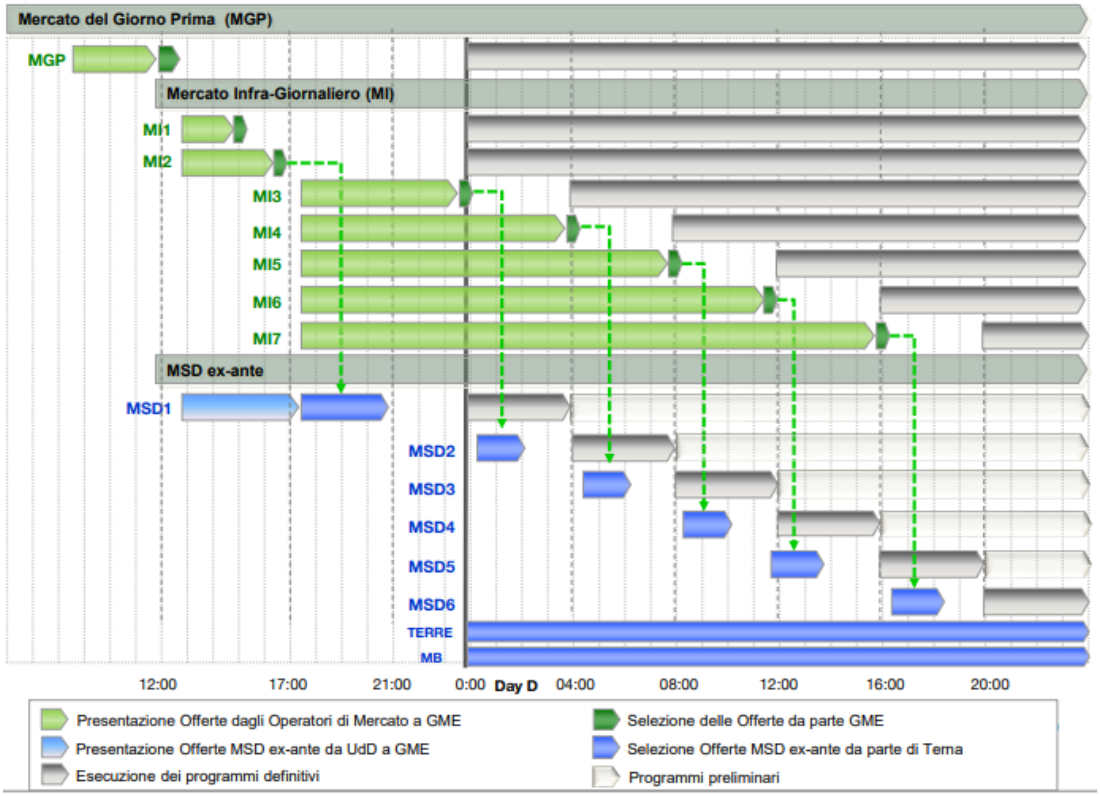


Figura 1.13: Integrazione tra il Mercato Infragiornaliero e il Mercato per il Servizio di Dispacciamento: scenario precedente. Fonte[32]

Tale ristrutturazione si pone nell'ottica della creazione di mercati infragiornalieri più efficienti, capaci di sostenere la crescita di produzione di energia da fonti rinnovabili e intermittenti, rendendo possibile una correzione rapida di eventuali cali o eccessi energetici, grazie all'interconnessione a livello europeo di risorse e fabbisogni. L'avviamento di tale soluzione di mercato rientra nel progetto di *Single Intra-Day Coupling* (SIDC), per lo sviluppo e realizzazione di un meccanismo di allocazione implicita della capacità di scambio contestuale all'accoppiamento di domanda e offerta in zone d'Europa diverse.

Tale meccanismo si basa su piattaforma centralizzata a livello europeo, chiamata XBID ("*Cross-Border Intraday*"), che funge da collegamento dei sistemi di negoziazione loca-

[32] Terna (2020)

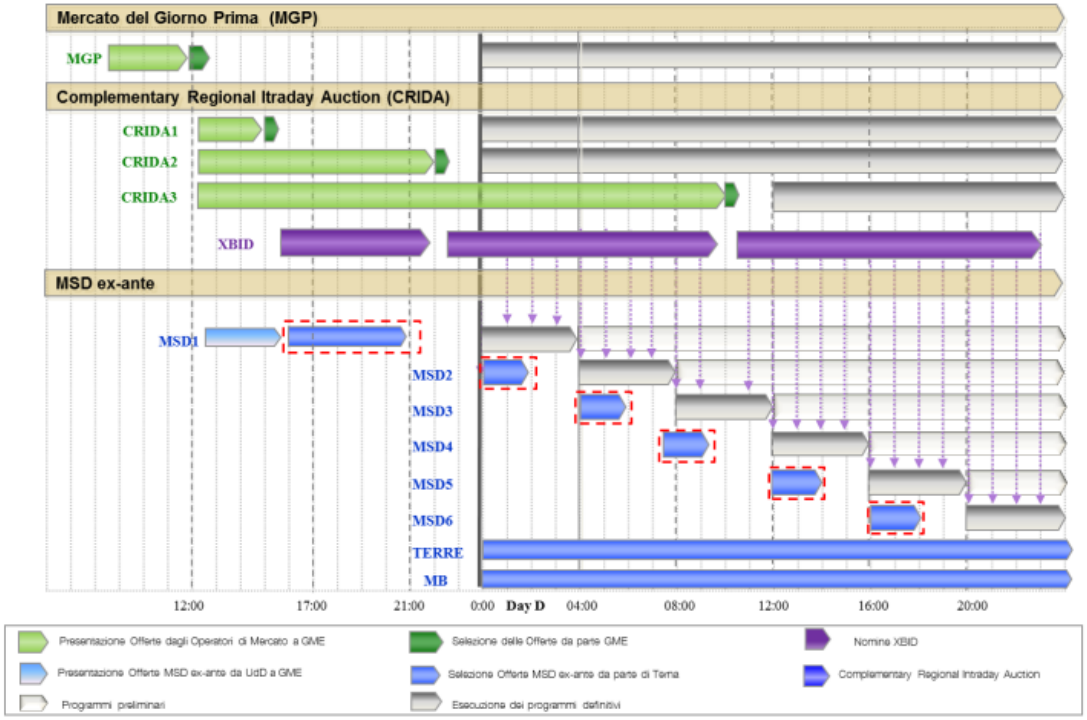


Figura 1.14: Integrazione tra il Mercato Infragiornaliero e il Mercato per il Servizio di Dispacciamento: scenario attuale. Fonte[33]

li (*Local Trading System*), gestiti dai singoli NEMO. La piattaforma riceve in entrata le offerte di acquisto e di vendita in ciascuna zona di mercato, caricate in tempo reale dai NEMO delle diverse zone del mercato europeo e le capacità di trasporto disponibili tra le diverse zone, calcolate e comunicate dai TSO locali, come Terna. La contrattazione continua, che si svolge grazie a questa piattaforma, si rivolge per l'appunto all'integrazione crescente di fonti rinnovabili, che portano con sé una notevole aleatorietà. Capita spesso, infatti, che un operatore abbia bisogno di correggere velocemente la propria posizione, per evitare di incorrere in prezzi di sbilanciamento, nel caso in cui, in prossimità dell'ora h, si renda conto, ad esempio, che il vento è stato meno forte di quanto si attendesse.

Tuttavia, questo meccanismo, oltre ad essere poco efficiente per grandi contrattazioni, non consente di derivare il prezzo alla capacità di trasporto, perché si dispone di singole accoppiate di domanda e offerta, che non consentono di valorizzare la capacità di interconnessione.

[33] Terna (2020)

Per ovviare a questo svantaggio, i regolamenti europei prevedono l'introduzione di un meccanismo complementare a quello della contrattazione continua: le aste implicite intraday ("*Intraday Auctions for Pricing* o IDA"), con un principio d'azione molto simile a quello del MGP.
CRIDA ("*Complementary Regional Intraday Auctions*") è costituito da un sistema a tre aste, due delle quali svolte nel giorno antecedente la consegna e una nel corso del giorno *d*, con i seguenti orari:

- la prima asta, dalle 15:00 *d-1* (ore negoziabili 0:00-24:00 *d*) con la capacità residua del mercato del giorno prima;
- la seconda asta dalle 22:00 *d-1* (ore negoziabili 0:00-24:00 *d*) con la capacità disponibile in esito al processo di ricalcolo IDCC1[34];
- la terza asta dalle 10:00 *d* (ore negoziabili 12:00-24:00 *d*) con la capacità disponibile in esito al processo di ricalcolo IDCC2[35].

Le CRIDA permettono di fornire segnali di prezzo della capacità interzonale nel timeframe infragiornaliero e di estrarre la rendita da congestione derivante dall'allocazione della capacità, risolvendo il principale limite della negoziazione continua di XBID.

Nella nuova struttura del MI, a valle di ogni asta vi sono programmi che possono cambiare. Gli operatori prima di ogni sessione MSD e in *h-1*, devono dichiarare il proprio programma tramite una nomina a Terna. In questo risiede la maggiore differenza rispetto al precedente meccanismo e la più grande sfida per il TSO italiano. Infatti, in precedenza, le negoziazioni erano "unit based", cioè gli operatori partecipavano al Mercato Infragiornaliero con ogni singola unità, cosa che permetteva a Terna di conoscere, a valle della chiusura del mercato, il programma esatto di ciascuna unità. La seduta di offerta per ciascuna sessione terminava alcune ore prima del periodo orario di consegna, e il Mercato dei Servizi di Dispacciamento ex -ante, ovvero la fase di programmazione di MSD, operava semplicemente a correzione degli esiti del Mercato dell'Energia, senza imporre a preventivo vincoli di sicurezza su tali esiti.
L'attuale struttura del Mercato Infragiornaliero prevede che gli operatori partecipino non

[34]IDCC1 (Intraday Capacity Calculation 1) previsto in esercizio per la fine del 2022
[35]IDCC2 (Intraday Capacity Calculation 2) in esercizio all'avvio del nuovo MI

più con ogni singola unità produttiva, bensì a portafoglio. Ciò significa che un operatore per una certa zona partecipa con la sua posizione complessiva; a valle del mercato avrà probabilmente una nuova posizione complessiva, data dagli aggiornamenti effettuati nel MI, ma non sappiamo come rifletterà quei cambiamenti sulle singole unità produttive. Per questo si richiede di nominarli a Terna. Questa situazione diventa più complicata da gestire per Terna, in quanto ora i programmi possono cambiare fino ad *h-1*, per effetto delle contrattazioni del mercato infragiornaliero, ed eseguendo la prima sessione di MSD in anticipo rispetto ad *h-1*, Terna deve tenere conto del fatto che le nomine fatte sono solo provvisorie e potenzialmente modificabili. Il cambiamento della struttura di MI ha reso necessario un cambio di paradigma anche per la fase di programmazione di MSD, MSD ex-ante. La sua esecuzione non può più essere esclusivamente a correzione degli esiti dei mercati dell'energia, ma si svolgerà in parallelo ad essi, andando in parte a correggere le negoziazioni già avvenute, in parte imponendo vincoli preliminari a quelle successive, fornendo degli estremi, a definizione degli "intervalli di fattibilità", in cui quegli operatori abilitati a MSD potranno nominare[36].

1.3.5 Il Mercato dei Prodotti Giornalieri

Nel Mercato dei Prodotti Giornalieri (MPEG) ha luogo la negoziazione continua dei prodotti giornalieri con obbligo di consegna dell'energia. Le contrattazioni che si concludono in MPEG sono infatti considerate definitive e vincolanti. In particolare, essi sono scambiabili con:

- "differenziale unitario di prezzo": il prezzo determinato dalla negoziazione rappresenta il differenziale rispetto al PUN, al quale gli operatori sono disposti a negoziare tali prodotti;
- "prezzo unitario pieno": il prezzo determinato dalla negoziazione coincide con il valore unitario di scambio dell'energia elettrica oggetto dei contratti negoziati.

Le due tipologie di prodotto sono associabili a diversi profili di consegna, rispondendo ad un fabbisogno di carico di base, quotato per tutti i giorni di calendario, quando

[36] Terna (2022e)

il sottostante è l'energia elettrica da consegnare in tutti i periodi rilevanti appartenenti al giorno oggetto di negoziazione, ma anche ad un fabbisogno di carico di picco, quotato dal lunedì al venerdì, quando il sottostante è l'energia elettrica da consegnare nei periodi orari rilevanti dal nono al ventesimo del giorno oggetto di negoziazione. Il GME individua nelle DTF ("Disposizioni tecniche di funzionamento") i prodotti giornalieri e i corrispondenti profili di consegna, nel rispetto della quantità di energia sottostante ciascun prodotto giornaliero fissata a 1 MW moltiplicato per il numero di ore sottostanti il prodotto medesimo. Ad oggi, i prodotti giornalieri vengono negoziati a "differenziale unitario di prezzo", con profili di consegna Baseload e Peak Load[37].
Al MPEG possono partecipare tutti gli operatori del mercato elettrico che siano anche operatori della PCE ("Piattaforma Conti Energia"), ovvero che siano abilitati a registrare transazioni sui conti energia, nella propria disponibilità, e in queste negoziazioni il GME si pone come controparte centrale. In qualità di operatore di mercato qualificato, il GME registra sulla PCE la posizione netta in consegna di energia risultante dalla negoziazione dei prodotti giornalieri, secondo le modalità previste nella Disciplina del mercato elettrico. Per ulteriori approfondimenti sul MPEG, si rimanda al *Testo Integrato della Disciplina del Mercato Elettrico - Capo 1 BIS* (2021).

1.3.6 Il Mercato per il Servizio di Dispacciamento

Nel Mercato dell'Energia gli unici vincoli allo scambio tra offerte riflettono la fisicità del sistema elettrico, e sono i limiti di scambio tra zone, ma non le limitazioni tecniche di ogni singola linea del sistema, né altre necessità come quella di avere una riserva di energia.
Per questo esiste un operatore di sistema, Terna in Italia, e una piattaforma, il Mercato per il Servizio di Dispacciamento, strumento di garanzia della sicurezza del Sistema Elettrico Nazionale. Detto anche Mercato dei Servizi, o *Ancillary Service Market*, questo è il luogo in cui Terna, il TSO italiano, si approvvigiona delle risorse necessarie alla gestione e al controllo del sistema, ovvero dei cosiddetti *servizi ancillari*, provvedendo:

[37] GME (2022)

- alla creazione della riserva di energia per la rialimentazione del sistema, a garanzia del bilanciamento in tempo reale e del ripristino delle condizioni di sicurezza, a fronte della disconnessione dalla rete o in caso di squilibri tra immissioni e prelievi;
- alla risoluzione delle congestioni intrazonali generate in esito al Mercato dell'Energia e la regolazione della tensione di gruppi di generazione, relativa all'attivazione della riserva creata, con il fine ultimo di mantenere gli elementi di rete entro i corretti limiti di funzionamento.

Il Mercato per il Servizio di Dispacciamento è un mercato "a variazione". L'approvvigionamento dei servizi viene, infatti, operato mediante modifica dei programmi delle UP: Terna accetta offerte in incremento o decremento della produzione rispetto al programma ottenuto in esito ai Mercati dell'Energia, se questi non consentono di assicurare i servizi. Ad esempio, se un programma con un certo assetto di produzione e di carico, comporta, per la corrente su questa linea, il superamento dei limiti di trasmissione, l'intervento di Terna consiste nella diminuzione della produzione a monte e nell'aumento della produzione a valle, così da contrastare il flusso sulla linea, evitando la congestione.
La partecipazione a MSD è obbligatoria, in ogni giorno e ora, per tutte le Unità Abilitate, ovvero tutti gli impianti di generazione, disponibili all'esercizio, con:

- potenza nominale superiore ai 10 MVA (megavoltampere),
- capacità di rispondere a certi requisiti tecnici, in termini di gradiente di erogazione, tempistiche di risposta e durata di erogazione del servizio, ad eccezione di quelle alimentate da fonti rinnovabili non programmabili: di fatto, vengono considerate tutte le unità programmabili, come gli impianti termoelettrici, idroelettrici a bacino, ad eccezione di alcune unità soggette alla legge Seveso, perché connesse a industrie petrolchimiche, quindi che devono rimanere libere per rispondere a leggi di sicurezza dell'impianto a cui son connesse.

Il meccanismo di formazione dei prezzi

Tali impianti sono obbligati a sottomettere in MSD un'offerta di esercizio, al prezzo che vogliono: in MSD possono partecipare solo tali Unità Abilitate e le offerte possono essere presentate solo dai relativi utenti del dispacciamento. Le offerte sono poi accettate

sulla base del merito economico, compatibilmente con la necessità di assicurare il corretto funzionamento del sistema e sono valorizzate al prezzo offerto. Il meccanismo di pricing adottato è "*pay-as-bid*", e non il prezzo marginale, come nei Mercati dell'Energia (incrocio domanda e offerta in cui definisco quantità e prezzi).

Le motivazioni della scelta di tale tipologia di prezzaggio sono diverse: la più semplice è che il prezzo marginale potrebbe portare ad un esborso eccessivo per il consumatore finale. La controparte in MSD è Terna, non i consumatori, e riverserebbe quell'esborso maggiore in bolletta. Questa tesi non è esente da critiche: essendo il mercato dell'energia un "gioco ripetuto", giacché si ripete ogni giorno, il prezzo *pay-as-bid* converge al prezzo marginale: l'operatore tenderà a offrire non il suo valore reale (vero costo variabile), bensì il prezzo massimo a cui lui si attende che possa essere accettato. I mercati elettrici sono mercati liberi non perfettamente competitivi, in cui gli operatori di mercato tentano di massimizzare i loro profitti, adeguando le loro strategie di offerta sulla base delle condizioni di esercizio attese e tenendo conto dell'interazione tra i diversi mercati (*INC/DEC game*[38]). Queste strategie impattano sul costo di approvvigionamento dei servizi, soprattutto in presenza di ridotta liquidità (Graf, Quaglia and Wolak, 2020). È quindi cruciale rilassare i vincoli che impongono approvvigionamenti localizzati (e contrarre i fabbisogni complessivi), ampliare la platea di risorse in competizione e massimizzare l'efficienza del disegno di mercato.

Il vero motivo, che determina l'utilizzo del prezzaggio *pay-as-bid*, risiede nella natura matematica del problema della formazione dei prezzi in MSD. In MGP è un problema lineare, in cui non vi sono variabili intere, ma solo prezzi moltiplicanti quantità, e non si tiene conto dei vincoli tecnici. In MSD è un *problema a misto interi*: vi sono variabili intere, e vengono considerati i vincoli tecnici, riguardanti la natura fisica degli impianti. Ad esempio, un impianto termoelettrico, può essere spento, e quindi può avere potenza zero, ma se viene acceso deve stare ad un minimo tecnico, cioè almeno ad una certa potenza. Una volta raggiunto tale livello, può variare più o meno linearmente nel range tra il minimo e il massimo: è una variabile intera, e la teoria del prezzo marginale decade.
La differenza tra i due approcci di calcolo dei prezzi, può essere analizzata supponendo, ad esempio, di disporre di due unità produttive per far fronte a una domanda di 100 MW,

[38] Presenza di opportunità per gli operatori di mercato di ottenere profitto dalla differenza tra il modello di mercato elettrico e il meccanismo in cui effettivamente operano

entrambe con capacità massima 80 MW e potenza minima 50 MW, ma la prima offre di produrre a 10 €/MWh, la seconda a 100 €/MWh.

Nel MGP si risolve il problema senza tenere conto dei vincoli tecnici, ovvero della potenza minima per cui si pensa di poter portare la potenza di UP da zero al massimo. La soluzione ottima si ha accettando 80 MW dalla prima UP e 20 MW dalla seconda. Il prezzo marginale del mercato è 100 €/MWh, perché se si dovesse consumare 1 MW in più rispetto a questa soluzione, non si potrebbe comprarlo dalla prima UP, in quanto è già al massimo e si andrebbe ad accettarlo sulla seconda.

Nell'MSD si risolve il problema tenendo conto del minimo tecnico, per cui la soluzione ottima si calcola dopo aver considerato che entrambe le UP devono stare almeno a 50 MW per poter essere disponibili a produrre. Viene soddisfatto in questo modo il fabbisogno di 100 MW. Il prezzo marginale del mercato è 10 €/MWh, perché se si dovesse consumare 1 MW in più rispetto a questa soluzione, si andrebbe ad accettarlo sulla prima UP. Questo però significa che la seconda UP è stata forzata a produrre per coprire la domanda di 100 MW, e sta venendo remunerata con il prezzo marginale di 10 €/MWh, anziché con 100 €/MWh (la sua offerta di produzione); quindi, tecnicamente la seconda UP è costretta a rimetterci. Nessun meccanismo funziona se costringe qualche sua parte a perderci: a valle del mercato se qualcuno è stato costretto in questa situazione, viene reintegrato della quota, e questo è il vero motivo dell'introduzione del meccanismo di prezzaggio *pay-as-bid*.

Negli USA è definito il prezzo nodale, cioè un prezzo marginale ad ogni nodo, poi applicato a tutte le risorse attaccate a quel nodo, ma in realtà per riequilibrare questo prezzo vengono pagate delle quote ("*mequal payments*") fuori dal mercato, non riflesse opportunamente dal prezzo marginale.

L'approvvigionamento di riserva

Il Mercato per il Servizio di Dispacciamento assicura la reperibilità continua di energia elettrica, e quindi la stabilità del Sistema, attraverso la predisposizione di riserve di potenza. Il gestore della rete di trasmissione, Terna, sfrutta tali risorse per bilanciare eventuali cali di energia nella rete, che non è possibile coprire neanche a seguito delle negoziazioni a breve termine sul Mercato Infragiornaliero.

La riserva è impiegata per rispondere alla regolazione di frequenza e alla regolazione di tensione. In base al tempo di attivazione, in cui è chiamata a fare regolazione, la riserva di energia si suddivide in tre tipologie:

- riserva primaria, impiegata per la regolazione primaria,
- riserva secondaria, impiegata per la regolazione secondaria,
- riserva terziaria, suddivisa in sostituzione ("a salire" e "a scendere"), rotante ("a salire" e "a scendere") e pronta (solo "a salire"), per la regolazione terziaria.

La riserva primaria non è un servizio acquistato in MSD: tutti i generatori italiani sono obbligati a fornire regolazione primaria, riservando una quota parte della loro capacità nominale. Per questa fascia di bilanciamento un impianto sul continente deve predisporre almeno l'1,5% della propria capacità massima come riserva primaria. La Sardegna, essendo asincrona, necessita di più margine, perché dispone di meno generatori che reagiscono velocemente. La quota riservata per la primaria sale al 10%.

La riserva secondaria e terziaria sono servizi acquistati in MSD. Rispetto ai programmi acquistati in MGP per il giorno seguente ad una certa ora, Terna deve assicurare di avere, ad esempio, almeno 2000 MW di terziaria, per via dell'incertezza sulle previsioni, e 1000 MW di secondaria, perché il gruppo più grande che scatta è di 1000 MW. Dunque, tra l'esito dei programmi in MGP relativo a tutte le UP e la loro potenza massima, tenendo conto della loro rapidità d'azione, Terna deve assicurarsi di disporre di questo spazio. A questo scopo, se tutti i gruppi sono usciti al massimo, occorre abbassare la loro produzione per creare spazio e accenderne altri, visto che è stata imposta la riduzione della loro produzione. Quando in tempo reale verrà attivata con un ordine, tale richiesta verrà pagata perché è come accettare una quantità offerta sul mercato.

Sono pertanto tre, le tipologie di riserva negoziate in MSD: secondaria e terziaria, che si distingue a sua volta in terziaria pronta, rotante e di sostituzione.

- Secondaria, detta *aFRR* ("*automatic Frequency Restoration Reserve*"): rapida, erogata mediamente in 180 secondi;
- Terziaria pronta, rotante e di sostituzione: differenziate a seconda del tempo entro cui viene erogata e della durata di erogazione. Alcune centrali infatti sono veloci e possono garantire riserva rotante, altre no e Terna deve sapere a quali UP rivolgersi per garantire le quote di terziaria.

– Terziaria pronta: erogata entro 15 minuti con gradiente di 50 MW/min e sostenuta per almeno 120 minuti;

– Terziaria rotante: più rapida, erogata entro 15 minuti e a sostegno della modulazione di frequenza per almeno 120 minuti. Nella terminologia europea è indicata come *mFRR* ("*manual Frequency Restoration Reserve*"), ma sta ad indicare il complementare di quello che in Italia si intende come terziaria rotante:

fabbisogno di mFRR = fabbisogno di rotante – fabbisogno di secondaria

– Terziaria di sostituzione: erogata entro 120 minuti con gradiente di 0,67 MW/min e senza limiti di durata. Nella terminologia europea è indicata come *RR* ("*Replacement Reserve*"), ma sta anche qui ad indicare il complementare di quello che indichiamo come terziaria di sostituzione:

fabbisogno di RR = fabbisogno di sostituzione – fabbisogno di rotante

La mappatura tra i prodotti scambiati nel Mercato Italiano dei Servizi, e i prodotti europei, è utile per il confronto con gli altri Paesi. In Italia viene adottato un approccio "a matrioska" alla dimensionalità delle riserve, rappresentato in Figura 1.15: il fabbisogno di secondaria è incluso nel fabbisogno di terziaria rotante, a sua volta incluso nel fabbisogno di terziaria di sostituzione.

secondaria $\subset$ *terziaria rotante* $\subset$ *terziaria di sostituzione*

Questo significa che, ad esempio, un fabbisogno di secondaria di 500 MW e di rotante di 1000 MW, per Terna si traduce nella creazione di uno spazio, non di 1500 MW, bensì di 1000 MW, di cui 500 devono essere abbastanza veloci da garantire la secondaria. Così come, se per la riserva di sostituzione è richiesto un fabbisogno di 3000 MW, per la rotante di 1000 MW e per la secondaria 500, Terna deve assicurare uno spazio di 3000 MW attivabile in 120 minuti, di cui 1000 MW attivabili in massimo 15 minuti e 500 in massimo 180 secondi, per poter supportare la regolazione di frequenza nelle corrette tempistiche.

L'approccio europeo definisce le quantità al netto del sottoinsieme a cui appartengono. Infatti, se riserva secondaria e aFRR coincidono, la mFRR è la riserva rotante al netto della secondaria, la RR è la riserva di sostituzione al netto della rotante (e della secondaria).

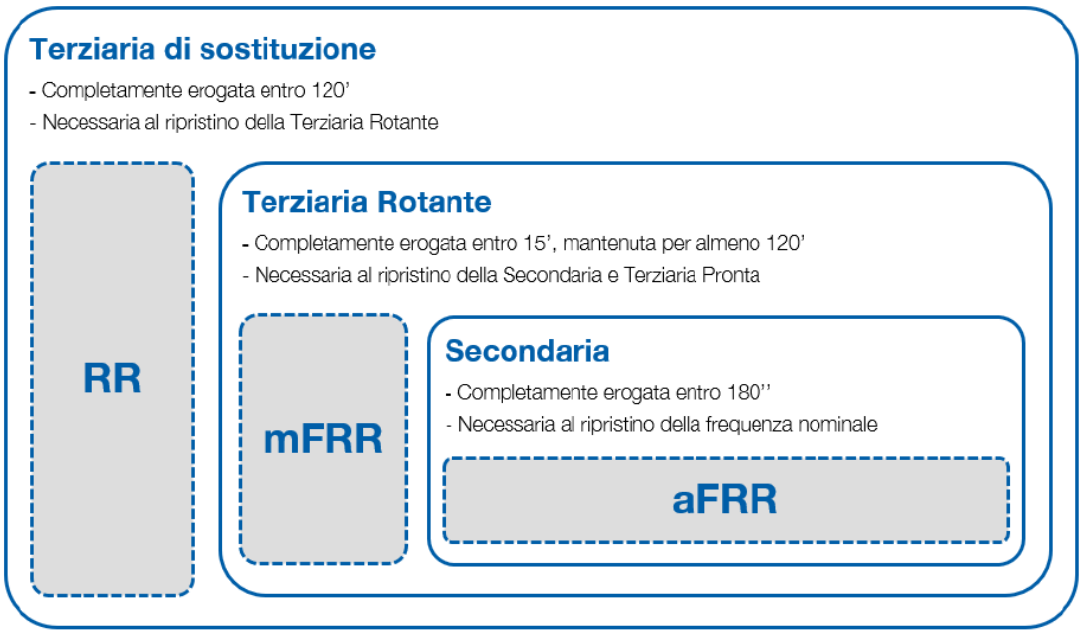


Figura 1.15: Mappatura tra i prodotti del Mercato italiano per il Servizio di Dispacciamento e i prodotti europei. Fonte[39]

La regolazione di frequenza

A seguito di un evento perturbativo, cioè di perdita di una certa quantità di potenza prodotta, la frequenza non torna naturalmente al suo valore nominale, ma si devono mettere in atto azioni, automatiche e non, che permettano di ripristinare i normali valori di esercizio. Per fare ciò, Terna si approvvigiona della capacità di riserva necessaria per mettere in atto azioni per la regolazione della frequenza.

Il meccanismo di attivazione delle tre riserve per la regolazione di frequenza è ben riassunto in Figura 1.16.

Quando si verifica uno squilibrio di rete a causa, ad esempio, del blocco di una UP, la frequenza si abbassa, ed entro 30 secondi scatta la *regolazione primaria*, a sostegno della variazione di potenza per almeno 15 minuti. La regolazione primaria è la risposta automatica di ogni singolo generatore, attivata per mezzo di regolatori automatici disposti negli impianti e non centralmente coordinata, se non dal fatto che esiste una frequenza comune a tutto il sistema e in base a quella i generatori rispondono. La reazione si verifica da parte

[39] Terna (2020)

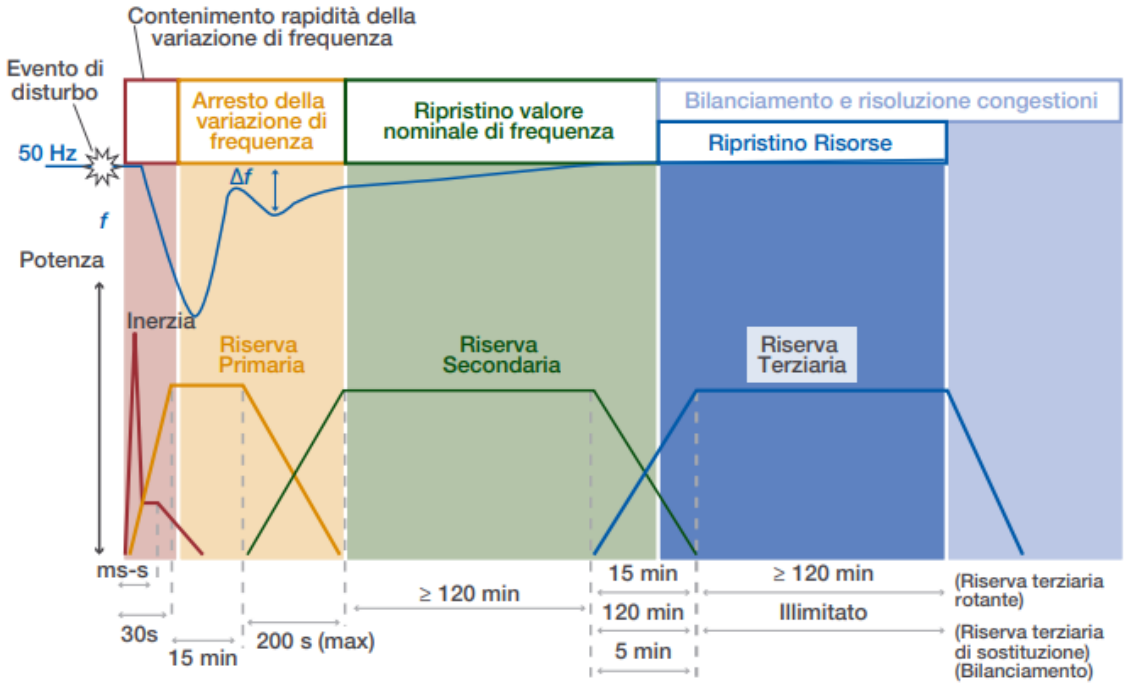


Figura 1.16: Schema della regolazione di frequenza in funzione del tempo di intervento. Fonte[40]

di quelli interconnessi alla stessa area sincrona (con lo stesso segnale di frequenza), motivo per cui se in Italia scatta un generatore, reagiscono anche quelli in Germania, perché l'Italia è interconnessa al resto d'Europa. Eccezione in Italia è rappresentata dalla Sardegna, connessa all'Italia e quindi all'Europa con collegamenti in corrente continua che rendono asincroni i due sistemi: se succede qualcosa in Italia, non per forza reagirà anche la Sardegna.
La regolazione primaria viene attivata dai singoli generatori, che immettono potenza per frenare la derivata di frequenza, senza ripristinarla al suo valore nominale di 50 Hz, ma semplicemente arrestando la variazione, evitando che la frequenza scenda troppo, causando blackout. Per la tipologia di regolazione a cui è predisposta, finalizzata al contenimento dello squilibrio di rete, la riserva primaria è detta "*Frequency Containment Reserve*" (FCR).

In 100-200 secondi intervengono i servizi di "*Frequency Restoration Reserve*" (FRR), che come da definizione, hanno l'obiettivo di ripristinare l'equilibrio di rete: inizialmente

[40] Terna (2021b)

scatta la *regolazione secondaria*, con capacità di sostenere la differenza di potenza per almeno 120 minuti. È coordinata centralmente a livello nazionale, mediante un ordine, da parte di Terna, ai vari gruppi a servire la regolazione secondaria, di quanto aumentare la frequenza per ripristinare il bilancio tra produzione e carichi e far ritornare la frequenza al livello nominale di 50 Hz. A questo punto potrebbe essersi esaurita la riserva predisposta a regolazione secondaria. Se le risorse pronte a reagire a un nuovo evento sono state già attivate e utilizzate tutte, occorre ripristinarle e ricreare le bande di secondaria. A questo scopo interviene la regolazione terziaria.

In 15-120 minuti scatta la *regolazione terziaria*, per il ripristino delle riserve, con capacità di sostenere il differenziale di potenza per 120 minuti (riserva rotante) o illimitatamente (riserva di sostituzione). A tale scopo, Terna impartisce disposizioni di esercizio come, ad esempio, l'entrata in servizio di centrali elettriche di riserva o la variazione della potenza prodotta da quelle già in servizio.

La regolazione di tensione

Il secondo parametro fondamentale per cui deve essere garantita una adeguata regolazione è la tensione. Il controllo della tensione è strettamente correlato alla gestione della potenza reattiva circolante in rete.

Per il corretto funzionamento del sistema occorre garantire che la tensione ai nodi sia entro un certo range. La regolazione di tensione è un meccanismo complesso che vede la coordinazione di molteplici fattori della rete: ad esempio, le centrali di generazione alzano e abbassano il loro *set point* per assorbire o immettere energia cambiando la tensione ai nodi.
Se un'unità abilitata risulta accesa nel momento del bisogno, la fornitura di questo servizio è obbligatoria, senza alcuna remunerazione. I costi per la regolazione di tensione, in Italia, si materializzano quando in una porzione di rete nessuna UP è accesa, nel momento del bisogno. Terna deve provvedere ad accenderne, per poter quindi provvedere al servizio di regolazione. Pertanto, per le Unità Abilitate, cambiare il *set point* è obbligatorio, se necessario, ma non è obbligatorio farsi trovare accese al bisogno: se occorre, va sostenuto il costo dell'accensione.

Il numero minimo di unità produttive che devono risultare in servizio per garantire la regolazione di tensione è definito *VRI*, "Vincolo a Rete Integra". Tali vincoli sono rappresentati da cluster che rappresentano porzioni di rete (es. Campania, Foggia...), mo-

strando la presenza di vincoli anche a livello intrazonale, cioè all'interno di ogni singola zona. Ad esempio, per una certa ora della giornata di domani, per mantenere le tensioni nel range di sicurezza, occorre almeno una unità accesa in Campania: se dai programmi definiti dai Mercati dell'Energia, almeno una UP risulta accesa, Terna non dovrà sostenere costi, mentre se non vi sono UP accese, Terna in MSD provvede a soddisfare tale bisogno, sostenendo a questo punto il costo di accensione.

Come per la regolazione della frequenza, anche la regolazione della tensione prevede una serie di azioni su più livelli gerarchici:

- Regolazione primaria di tensione: consiste nella modulazione della potenza reattiva in uscita dal gruppo di generazione sulla base dello scostamento della tensione ai morsetti del medesimo gruppo. La regolazione primaria di tensione ha, quindi, carattere prettamente locale e viene fornita da tutti i gruppi di generazione rotanti mediante il Regolatore Automatico di Tensione (RAT) per la regolazione a livello di gruppo e mediante il Sistema Autonomo per la Regolazione della Potenza Reattiva e della Tensione (SART) a livello di centrale;

- Regolazione secondaria di tensione: consiste nella modulazione della potenza reattiva in uscita dall'unità che effettua la regolazione sulla base dello scostamento della tensione su alcuni nodi predefiniti, chiamati "nodi pilota". Tale regolazione ha carattere regionale e viene effettuata mediante il Regolatore Regionale di Tensione (RRT), che impartisce segnali di livello a centrali elettriche e stazioni afferenti alla stessa Area.

Il servizio di regolazione della tensione ad oggi è unicamente effettuato da impianti di produzione rotanti e da strumenti di regolazione gestiti direttamente da Terna. A tal proposito, Terna sta valutando l'opportunità di avviare progetti pilota per testare la fornitura di potenza reattiva da parte di impianti ad oggi non abilitati. Per ulteriori approfondimenti Terna (2021b).

La logica "energy only"

Nel Mercato per il Servizio di Dispacciamento non vengono effettuate offerte separate per i servizi. Sono effettuate offerte in aumento o decremento della potenza per una certa quantità ad un certo prezzo, ma senza essere differenziate in base all'utilizzo finale. L'unica eccezione riguarda la regolazione di secondaria: essendo un servizio particolarmente veloce, può essere fatta per questo un'offerta a parte.

Quello che viene pagato al produttore da Terna sono le quantità accettate per spostare il suo programma.

Se un produttore, come esito del Mercato dell'Energia, ha accordato di generare una certa quantità di energia, ricevendo una remunerazione pari all'energia venduta su MGP moltiplicata per il *clearing price* zonale, ma Terna, sul Mercato dei Servizi, ritiene necessario alzare la produzione, l'UdD (Utente di Dispacciamento) di tale produttore vende energia a Terna, quella in più richiesta, e riceve il proprio prezzo d'offerta. Se Terna ritiene necessario ridurre la sua produzione, l'UdD di tale produttore acquista energia da Terna e la paga al proprio prezzo d'offerta. In pratica, il produttore dovrà restituire il prezzo di quella quantità, per cui era stato remunerato precedentemente: non c'è un pagamento esplicito da parte di Terna per questo cambiamento di programmi.

Se invece l'unità produttiva, che in MSD si ritiene necessario essere funzionante, è, come esito del Mercato dell'Energia, spenta, Terna paga una specifica offerta di accensione (€/avviamento) per ogni avviamento in più fatto su MSD rispetto al MI, senza però offrire una remunerazione all'UP per la capacità "riservata". La riserva di energia è in questo senso un servizio gratuito, che risulta in costi solo se l'impianto è spento.

Nel caso esemplificativo, in cui nel sistema vi siano due unità produttive, UP1 e UP2, possono presentarsi diversi scenari:

1. entrambe le unità produttive sono accese come esito del Mercato dell'Energia, ma i vincoli non sono soddisfatti: l'equilibrio di sistema (con consumi pari alla somma dell'energia prodotta da UP1 e UP2) causa una congestione di rete nella linea tra UP1 (a monte) e UP2 (a valle). Terna interviene abbassando la produzione di UP1 (UP1 restituisce soldi sulla base del prezzo in acquisto che aveva offerto), e alzando quella di UP2 (Terna paga UP2 sulla base del prezzo di vendita che aveva offerto): l'accettazione della quantità in acquisto su UP1 per la risoluzione della congestione ha richiesto indirettamente l'accettazione di una medesima quantità in vendita su UP2.

2. UP1 esce a potenza massima e UP2 è spenta e non può essere accesa velocemente: l'equilibrio di sistema (con consumi pari alla somma dell'energia prodotta da UP1 e UP2) comporta l'impossibilità, per UP1 perché è al massimo, UP2 perché spenta, di fornire riserva a salire (a scendere sì perché se ho un eccesso di produzione è sufficiente abbassare UP1). Terna interviene accendendo UP2, per creare la riserva necessaria, (Terna paga UP2 perché la obbliga ad accendersi e a portarsi al minimo tecnico) e abbassando UP1 (UP1 restituisce soldi perché Terna ha decrementato la sua produzione). L'energia di riserva che verrà poi prodotta da UP2 verrà pagata se sarà necessaria, ma inizialmente Terna paga solo per la garanzia di averla.

Dal punto di vista economico, unendo gli esiti di Mercato dell'Energia e Mercato dei Servizi, UP1 e UP2 hanno partecipato al Mercato dell'Energia, quindi ricevono il prezzo zonale derivante dal Mercato dell'Energia. Nel MSD, Terna modifica i loro programmi: chi vede la propria produzione ridursi, restituisce un prezzo, quello in acquisto, chi aumenta la produzione vedrà pagata la quota aggiuntiva. Nei due casi riportati ad esempio, UP1 scende in entrambi. Il suo ricavo in quell'ora è rappresentato da:

$$ricavo = (quantità\ MGP \times prezzo\ MGP) - (energia \times prezzo\ a\ scendere)$$

Il ricavo è decrementato, ma non significa che lo sia anche il profitto, perché UP1 non sosterrà il costo variabile di produzione; quindi, quello che offrirà a scendere è qualcosa che fa sì che UP1 non ci vada a perdere. Ad esempio, immaginando che sul Mercato dell'Energia gli siano stati accettati 100 €/MWh per produrre, e il costo variabile di produzione di UP1 è 80 €/MWh. Il massimo prezzo che ragionevolmente offre a scendere sono, come minimo, 80 €/MWh. Se Terna lo accetta a scendere, UP1 restituisce gli 80 €/MWh, che sta comunque risparmiando, perché non sta producendo e si tiene comunque 100 €/MWh, che è guadagnato. Quindi sia che lo faccia scendere o no per UP1 è indifferente. Nella realtà molto spesso i produttori offrono molto meno del costo variabile di produzione: se un'unità di produzione ha costo variabile pari a 40 €/MWh, ma offre 80 €/MWh, l'UP ci sta guadagnando dal fatto che Terna lo sta obbligando a scendere: oltre ai 100 €/MWh che sono guadagnati, ci sta risparmiando 80 €/MWh non producendo, quando in realtà ne spenderebbe 40 (il suo vero costo variabile di produzione), quindi un ulteriore guadagno.

L'evoluzione del MSD:
le unità virtuali abilitate miste (UVAM)

Il Mercato dei Servizi è un mercato obbligatorio per tutte e sole le unità abilitate a parteciparvi. Si tratta di grandi e medi impianti, unità programmabili e con potenza installata superiore a 10 MVA, e ad oggi sono circa pari a 250 unità di produzione e pompaggio[41].

Nell'ottica della transizione energetica, la crescente necessità di flessibilità del sistema elettrico, associata alla riduzione delle ore di produzione degli impianti termoelettrici tradizionali, rende essenziale l'approvvigionamento di servizi di rete anche da nuove risorse, meglio ancora se rinnovabili. Inoltre, la diversificazione delle risorse che partecipano a MSD può contribuire a minimizzare i costi complessivi per il sistema elettrico, quindi quelli dell'utente finale, dal momento che comporta un aumento dei partecipanti e quindi della competizione.

Per questo Terna, in accordo con l'ARERA, ha avviato un processo di progressiva apertura del mercato dei servizi alle risorse dapprima non abilitate, attraverso la definizione di progetti pilota, per avviare il processo di revisione del mercato, e testare le caratteristiche di nuove risorse. Tali progetti mirano all'introduzione in MSD di impianti di generazione di tipo non rilevante, rinnovabili e non (<10 MVA) ed degli accumuli, incrementando la quantità di risorse disponibili a garanzia della sicurezza del sistema e, allo stesso tempo, diversificando la tipologia di risorse abilitate ai servizi. Essi sono:

- Progetto UVAM (Unità Virtuali Abilitate Miste), che, come i loro predecessori UVAP e UVAC, insieme al Progetto Pilota per la Regolazione Secondaria, e al Progetto UPR (Unità di Produzione Rilevanti), mirano all'aggregazione in unità virtuali di risorse distribuite, e inizialmente non abilitate, per generazione, accumulo e consumo;
- FRU (Progetto Pilota Riserva Ultra-Rapida), nell'ambito della definizione di nuovi servizi di rete necessari per garantire l'integrazione delle rinnovabili;
- UPI (Unità di Produzione Integrate con sistemi di accumulo), che insieme al Progetto Pilota per la Regolazione di Tensione, mira a ottimizzare la fornitura di servizi già esistenti da parte di nuove tecnologie, inizialmente non abilitate.

I progetti pilota delle UVAC e delle UVAP sono stati avviati per l'abilitazione al MSD rispettivamente della domanda (da giugno 2017) e della produzione distribuita (da dicem-

[41] Terna (2021b)

bre 2017). Essi sostanzialmente, miravano all'aggregazione di punti di prelievo e punti di immissione rispettivamente a formare un'unica unità virtuale. Tali progetti si sono conclusi a novembre 2018 per dare avvio al progetto pilota delle UVAM, che abilita negli stessi aggregati unità di consumo, di produzione e sistemi di accumulo. Ciò che conta per Terna è il programma equivalente, quindi di quanto l'UVAM, come aggregato di unità, può aumentare o diminuire la produzione su ordine di Terna. In particolare, requisito da soddisfare è la flessibilità "a scendere" o "a salire", di almeno 1 MWh entro 15 minuti dalla richiesta di Terna.
Parallelamente all'abilitazione delle unità virtuali (UVA), Terna ha avviato il progetti pilota delle UPR, per la partecipazione volontaria a MSD di Unità di Produzione Rilevanti non oggetto di abilitazione obbligatoria. Con questa definizione si intendono, ad esempio, i parchi eolici o solari di grossa taglia (>10 MVA), impianti di generazione di tipo rilevante, ma non oggetto di abilitazione obbligatoria a MSD.

Punto di forza dei progetti UVA è il loro approccio decentralizzato, che rende possibile utilizzare la potenza di riserva nelle prossimità delle aree con instabilità e squilibri di rete. Per questo motivo è importante una diffusione regionale capillare e a tal proposito Terna ha istituito 18 zone di aggregazione[42] all'interno delle quali i partecipanti al progetto UVAM devono offrire almeno 1 MW di potenza di riserva "a scendere" e "a salire", per i servizi di risoluzione delle congestioni, riserva terziaria rotante, di sostituzione e bilanciamento[43].

Anche rispetto alla regolazione economica cambiano alcuni aspetti. La remunerazione prevede due meccanismi, il primo legato all'energia attivata (€/MWh) e l'altro alla disponibilità (corrispettivo fisso, €/MW). La scelta di remunerare la disponibilità, a differenza di quanto si verifica per le grandi centrali, è motivata dal fatto che le risorse partecipanti, dal lato consumatori, sono principalmente stabilimenti produttivi industriali, disponibili a ridurre i propri prelievi di energia. Tali soggetti, per fornire flessibilità nel mercato dei servizi, devono sostenere costi fissi di investimento per installare e mettere a punto le apparecchiature necessarie a sviluppare il servizio e costi annuali di gestione dell'operatività.

Un altra importante novità introdotta con i progetti UVA è la figura del *Balancing Ser-*

[42] Terna (2020)
[43] Marchisio et al. (2022)

vice Provider (BSP). È propriamente il soggetto titolare della UVA e responsabile della prestazione dei servizi negoziati sul MSD, e non deve necessariamente coincidere con l'Utente del Dispacciamento (*Balancing Responsible Party*, BRP). Il BSP non ha legame contrattuale con il BRP e fornisce direttamente i servizi al gestore di rete, mentre il BRP è il responsabile del pagamento dei corrispettivi di sbilanciamento. Per ulteriori approfondimenti, *L'apertura delle risorse distribuite al mercato dei servizi: quale bilancio? (2022)*.

Il Mercato per il Servizio di Dispacciamento è suddiviso in due fasi: una fase di programmazione (MSD ex ante) ed una fase di tempo reale (Mercato di Bilanciamento, MB) ciascuna delle quali è composta da 6 sessioni, e ciascuna delle quali si conclude un esito. La sua struttura è rappresentata in Figura 1.14, assieme all'integrazione del progetto TERRE, implementata a partire da gennaio 2021.

La fase di programmazione (MSD ex-ante)

La prima fase, suddivisa in 6 sessioni di programmazione distinte (MSD1, MSD2, MSD3, MSD4, MSD5 e MSD6), secondo lo schema di Figura 1.14 produce come esito le offerte di acquisto e vendita di energia accettate da Terna, ai fini della risoluzione delle congestioni residue e della costituzione di un adeguato margine di riserva.
La programmazione è necessaria per motivi tecnici: sarebbe l'ideale poter fare tutto in tempo reale, per evitare incertezze, ma gli impianti non hanno flessibilità infinita: in un istante non posso portare la produzione da 0 a 100 o spegnerlo in un minuto, bensì devo predisporlo in anticipo. L'algoritmo di risoluzione del MSD ex-ante è un problema di ottimo con una funzione obiettivo nelle variabili e nei vincoli di mercato, allo scopo di minimizzare i costi nelle ore sottofase. L'algoritmo risolve simultaneamente tre questioni, in ordine di funzionalità:

1. *Unit commitment* (UC), nel quale l'obiettivo è minimizzare il costo totale di generazione in un un dato periodo, definendo un'adeguata programmazione delle UP, ovvero quali unità produttive tenere accese e quali spente per ogni istante di tempo.
   È un problema a variabili intere, giacché considera nella sua risoluzione i vincoli tecnici degli impianti: infatti, se un impianto è avviato, deve stare ad un potenza minima, e l'unico momento in cui si trova nel range tra potenza zero e potenza minima è proprio mentre sta rampando verso la minima. Se non ci fossero vincoli tecnici di funzionamento, non ci sarebbe il problema di *unit commitment*, perché non occorrerebbe stabilire

se una certa UP deve essere accesa o spenta, ma solo quanto produrre e, secondo tale logica, potrebbe farlo anche nel range di potenza tra zero e la minima.

2. *Optimal power flow (economic dispatch)* (OPF - ED), nel quale l'obiettivo è definire quanto le unità produttive attive devono produrre. Questo valore va definito, perché se un'UP è accesa e funzionante, può produrre qualsiasi valore compreso tra la potenza minima e massima dell'impianto.

3. *Security constrained* (SC) nel quale l'obiettivo è riflettere adeguatamente nel problema i vincoli tecnici di sicurezza, cioè la realtà fisica della rete, pertanto: come si distribuiscono i flussi in base ai carichi, quali sono le unità che produrranno, dove si creeranno congestioni.

L'algoritmo risulta pertanto definito come:

$$\min \left\{ \begin{array}{c} \text{costo quantità accettate in vendita - costo quantità accettate in acquisto +} \\ \text{costo avviamento + costo cambio assetto + costo atteso di utilizzo della} \\ \text{riserva secondaria in tempo reale + costo atteso azioni di bilanciamento in} \\ \text{tempo reale} \end{array} \right\}$$

sotto le condizioni relative ai:

- vincoli di sistema (bilancio di energia, fabbisogno di riserva, limiti di transito tra zone, vincoli di rete (come tensione e sovraccarico),
- vincoli di offerta (coerenza tra le quantità accettate e prenotate con le quantità, offerte),
- vincoli tecnici (tempo di avviamento, tempo di permanenza in/fuori servizio).

Il problema di ottimo in MSD ex-ante agisce sui programmi esito del Mercato dell'Energia, che definiscono alcuni impianti come spenti, altri come accesi e per questi un programma iniziale di produzione. Terna va a sostituire i programmi iniziali di fabbisogno con le sue previsioni, aggiunge tutti i vincoli, per trovare la soluzione a minimo scostamento (minimo costo per Terna) dalla soluzione di partenza che rispetta tutti i vincoli che prima non erano considerati.

- L'ALGORITMO MSD EX-ANTE: LA FUNZIONE OBIETTIVO
  Più precisamente, si tratta di minimizzare l'esborso sul mercato, dato dal costo dell'energia per le quantità accettate in vendita, e quindi acquistate da Terna nel momento in

cui ordina un aumento della produzione rispetto agli esiti del Mercato dell'Energia, al netto di quelle accettate in acquisto, cioè vendute da Terna nel momento in cui ordina una riduzione della produzione alle UP rispetto ai programmi. In tal caso, Terna incassa il valore corrispondente alla quantità in decremento, restituito dagli operatori che lo avevano incassato come vendita in MGP. Terna cerca di ridurre i costi sul Δenergia cercando quell'impianto che fa pagare meno la quantità richiesta in più.
I costi di avviamento riguardano l'accensione delle UP che risultano spente nel momento del bisogno ai fini della regolazione. Il gettone di avviamento ha un costo crescente al crescere delle quantità, fatto che conferisce convessità al problema di ottimo e garantisce la convergenza. Ad esempio, un impianto con potenza massima di 400 MW accetta offerte su due gradini: 200 MW (il minimo) e 200 MW in più sul minimo. Il prezzo dovrà essere più basso per il primo gradino e più alto per il secondo. Di conseguenza, l'andamento dei prezzi non è coerente con quelli reali per gli operatori, perché la centrale consuma di più in fase di accensione, e meno quando è a regime. Per questo è stato introdotto il gettone di avviamento, che copre i costi in più in fase di accensione. Come descritto in precedenza, Terna sostiene il costo di avviamento solo se l'UP esce spenta nelle 24 ore sul Mercato dell'Energia e ne viene ordinata l'accensione, di fatto aumentando il numero di avviamenti per quell'impianto.
Il costo per il cambio di assetto riguarda il passaggio tra configurazioni, quindi quando si tratta di accendere qualcosa nell'impianto.

Quelli appena enunciati sono costi reali, i successivi costi enunciati nella funzione obiettivo sono costi fittizi, che servono per trovare un trade-off ottimale sulle scelte, e infatti sono indicati come costi "attesi". In fase di creazione dello spazio per la riserva, Terna sostiene i costi per l'energia che movimenta per crearsi questo spazio, e non la riserva che vi sta sopra, né l'utilizzo che ne sarà fatto; quando verrà usata, sarà adeguatamente remunerata. Pertanto, disponendo di unità produttive con diversi costi per essere accese e per essere movimentate dalla potenza minima alla massima, Terna sceglie in base alla probabilità con cui si aspetta di attivare la riserva: nell'algoritmo vengono inseriti scenari con diversa probabilità di utilizzare una certa quantità di riserva, per avere un trade-off che tiene conto del rischio futuro di avere costi più elevati. Se la probabilità di attivare riserva è bassa, Terna sceglie quelle UP con costo di attivazione più basso, perché quasi sicuramente il costo per la movimentazione non sarà coinvolto. Se, invece, la probabilità di attivare riserva è elevata, la scelta verte su costi di movimentazione più bassi, a discapito di un maggiore costo di attivazione: questa penalità permette, infatti,

di far convergere meglio l'algoritmo.

- L'ALGORITMO MSD EX-ANTE: I VINCOLI

1. VINCOLI DI SISTEMA:
   - Vincoli di bilancio di energia: la quantità in acquisto deve essere uguale alla quantità in vendita, affinché il consumo sia pari alla produzione e vi sia bilanciamento nella rete;
   - Vincoli di fabbisogno di riserva: la riserva deve essere distribuita in un certo modo tra le aree geografiche nelle varie tipologie (secondaria, terziaria);
   - Limiti di transito tra le zone: di questo vincolo fisico se ne tiene conto anche nel Mercato dell'Energia. In MSD, in cui si risolve un problema nodale, si considera comunque la struttura zonale del mercato, seppur sia una rappresentazione semplificata della rete, perché la riserva viene approvvigionata a livello nazionale/aggregati di zone/zone stesse e non a livello nodale. Il limite di transito serve perché se Terna ritiene di dover spostare riserva tra zone, le serve conoscere i limiti di transito tra le zone;
   - Vincoli di rete (tensione, sovraccarico elettrodi): riguardano la struttura particolareggiata del mercato, quindi a livello nodale.

2. VINCOLI DI OFFERTA: che mirano al mantenimento della coerenza tra le quantità accettate e prenotate con le quantità offerte (indivisibilità delle offerte di minimo e spegnimento), e della congruenza sulla base dei criteri di convessità;

3. VINCOLI TECNICI DEGLI IMPIANTI
   In MGP gli operatori non possono inserire i loro vincoli tecnici (come la potenza minima), motivo per cui l'esito di MGP potrebbe essere insostenibile a livello pratico. Nel MI una prima correzione tenta di rendere fattibili i programmi. Ma è in MSD che i vincoli vengono implementati dettagliatamente, per garantire che l'esito sia tecnicamente eseguibile dagli impianti. In particolare, devono essere considerati il tempo e il profilo della rampa di avviamento , il gradiente di presa/rilascio di carico, il tempo di permanenza in servizio/fuori servizio, l'energia UP idroelettriche (per gli accumuli e impianti di pompaggio in generale): a differenza di un impianto

termoelettrico che può stare alla sua potenza minima tranquillamente, l'impianto di accumulo esaurisce l'energia, quindi si carica e scarica

- L'ALGORITMO MSD EX-ANTE: VARIABILI
  Le variabili coinvolte nella funzione obiettivo sono le quantità accettate in acquisto e vendita, se l'unità di produzione è accesa o spenta, e la quantità in produzione per ogni UP.

Il Mercato di Bilanciamento

Il Mercato di Bilanciamento (MB), ovvero la fase in tempo reale, è anch'essa suddivisa in 6 sessioni, ma solo per la presentazione dell'offerta; in realtà è come se fosse un mercato continuo, perché ogni 15 minuti risolve il mercato per i 15 minuti successivi, come rappresentato in Figura 1.14.
Il Mercato di Bilanciamento è un mercato di "ultima istanza", in cui è possibile effettuare la ripresentazione di offerte per l'uso in tempo reale. Alcune decisioni non possono essere più riviste, come tipicamente l'accensione e lo spegnimento e, fatta eccezione per quegli impianti che sono veloci da accendere, in tempo reale è possibile solo modulare all'interno delle bande, accettando in aumento/decremento quantità fattibili in base ai tempi. È per questo che in MSD ex-ante, Terna crea la riserva che si aspetta di dover utilizzare, a titolo gratuito. Per le offerte riservate da Terna in MSD ex-ante, i prezzi possono essere modificati solo in termini migliorativi, mentre per quelle non riservate possono essere modificati liberamente.
In definitiva, non c'è più il problema di *unit commitment*, giacché non è possibile decidere se accendere o spegnere UP, e l'algoritmo si riduce a un problema di *AC optimal power flow* (*security constrained* ed *economic dispatch*). Il problema di ottimizzazione ha quindi funzione obiettivo nelle variabili definite in precedenza, con vincoli uguali all'MSD ex-ante, al netto di quelli che riguardano l'accensione/spegnimento.

Integrazione a livello europeo: le piattaforme di bilanciamento

Con l'entrata in vigore, il 18 dicembre 2017, del "Regolamento Balancing", Regolamento (UE) 2017/2195 della Commissione del 23 novembre 2017, l'integrazione a livello europeo rivolge la propria attenzione ai mercati di bilanciamento dei Paesi UE, con nuovi orientamenti e discipline allo scopo di garantire un approvvigionamento economicamente efficiente dei servizi di bilanciamento.

Se il Mercato dell'Energia è stato soggetto di una completa integrazione a livello europeo, con la creazione di un unico algoritmo sia a livello Single-Day (SDAC) che Infra-Day (SIDC), il Mercato dei Servizi, per la sua natura di supporto e intervento, rimane un mercato parzialmente nazionale. Nel MSD ex-ante Terna predispone la riserva a livello nazionale, che attiva in caso di necessità nel MB. È proprio in questa fase di MSD, quando Terna decide dove attivare la riserva, che agisce il Regolamento Balancing: per poterla attivare nel modo più efficiente possibile, permette di andare a reperire riserva anche al di fuori dei confini nazionali, se conveniente.

Una delle novità è rappresentata dalle piattaforme di bilanciamento, progettate secondo un modello multilaterale TSO-TSO con attivazione delle offerte per ordine di merito economico a livello europeo. Su ogni piattaforma (eccetto quella per il collegamento europeo a fine di compensazione, *netting*, degli sbilanciamenti) viene scambiato un prodotto standard caratterizzato da specifici tempi di attivazione:

1. Piattaforma RR (progetto TERRE - *Trans European Replacement Reserve Exchange*), avviata il 13 gennaio 2021, per lo scambio energia di bilanciamento da *Replacement Reserve*, assimilabile a riserva terziaria di sostituzione;
2. Piattaforma mFRR (progetto MARI - *Manually Activated Reserves Initiative*), in attesa di go-live il 24 luglio 2022, per lo scambio energia di bilanciamento da *manual Frequency Restoration Reserve*, assimilabile a riserva terziaria rotante;
3. Piattaforma aFRR (progetto PICASSO - *Platform for the International Coordination of Automated Frequency Restoration and Stable System Operation*), go-live in programma il 24 luglio 2022, per lo scambio energia di bilanciamento da *automatic Frequency Restoration Reserve*, assimilabile alla riserva recondaria;
4. Piattaforma IN (progetto IGCC - *International Grid Control Cooperation*), avviata il 27 gennaio 2020 per l'ottimizzazione dei volumi di riserva secondaria attivati per il

controllo della frequenza di rete, tramite compensazione esplicita degli sbilanciamenti in tempo reale di segno opposto delle aree dei TSO.
L'IGCC è stato individuato da ENTSO-E, la rete europea dei gestori dei sistemi di trasmissione, come piattaforma di riferimento per l'implementazione del processo di compensazione dello sbilanciamento, definito, in ambito europeo, come *Imbalance Netting*.

Il primo progetto introdotto per l'integrazione dei mercati di bilanciamento è stato proprio quest'ultimo, IGCC, che rappresenta sostanzialmente il criterio di razionalità del *modus operandi* dei sistemi elettrici europei. Una volta scattata la regolazione primaria, viene attivata la regolazione secondaria: tutti i Paesi europei intervengono a regolare l'ammanco di energia, per riportare il sistema alla frequenza costante. Occorre a questo punto ripristinare la frequenza nominale del sistema: viene misurato l'ammanco di energia, dovuto per esempio alla perdita di un generatore in Italia, guardando allo scarto con l'estero. Se ad esempio in Germania si sta producendo più eolico di quello che serve al sistema, invece che produrre di più in Italia per il bilanciamento della rete e abbassare la produzione tedesca, l'integrazione basata su IGCC permette di compensare le due posizioni, aggiornando i programmi, con la conseguente riduzione dei costi per il cittadino europeo.

Fino a fine 2020, il mercato funzionale alla gestione da parte di Terna del sistema in tempo reale poteva essere identificato esclusivamente con le attività del mercato di bilanciamento (MB). A partire da gennaio 2021 è stata avviata la partecipazione alla piattaforma di bilanciamento europea per lo scambio di *Replacement Reserve* attraverso la piattaforma TERRE[44]. Il progetto europeo TERRE introduce un'ulteriore fase per eseguire il bilanciamento del sistema quanto più possibile in tempo reale. Le principali differenze introdotte dalle linee guida per la RR rispetto al mercato di bilanciamento italiano sono:

1. Processo orario schedulato: offerte sottomesse dagli operatori valide al massimo per l'ora successiva, e processo orario di selezione di tali offerte;

2. Remunerazione a prezzo marginale: gli scambi tra TSO verranno regolati al prezzo marginale, sia per quanto riguarda le offerte che i fabbisogni accettati;

[44] Entso-E (2022)

3. Prodotto standard per la RR convertito e scambiato: l'Italia (operativa con un modello di *central dispatch*) potrà convertire i prodotti prima di sottometterli, in modo tale che la sottomissione non comporti rischi per la sicurezza del sistema.

Entro un'ora dalla consegna (entro le 15, per la consegna dalle 16 alle 17), gli operatori devono nominare il punto di funzionamento dei loro impianti in base a tutte le soluzioni avvenute nei mercati precedenti, incluse le contrattazioni bilaterali. Sulla base di questi esiti, sottomettono offerte sulla piattaforma TERRE per aumentare o diminuire la propria produzione. Ciascun TSO europeo sottomette il proprio fabbisogno di attivazione all'algoritmo che seleziona le risorse più economiche per sopperire a tali fabbisogni.
TERRE è quindi un processo di bilanciamento aggiuntivo a MB, anticipato di 1 ora: MB in tempo reale entra in gioco un'ora dopo, di 15 in 15 minuti per correggere eventuali errori. Quindi, ogni ora, dopo la chiusura di TERRE, si attiva MARI e poi PICASSO e infine MB in tempo reale.

Pertanto, se prima l'identificazione dell'aggregato di zone poteva avvenire tenendo come riferimento la sola sessione di MB, ad oggi occorre considerare che il mercato per il bilanciamento è articolato in più segmenti indipendenti, dove il TSO scambia prodotti di bilanciamento con caratteristiche differenti e valorizzati a prezzi potenzialmente differenti.

A proposito della caratterizzazione del modello di dispacciamento italiano come *central dispatch*, il Regolamento (UE) 2017/2195, l' *Electricity Balancing Guideline* (EB-GL), ha richiesto ai TSO europei di definire tramite una serie di termini, condizioni e metodologie, la propria strategia operativa, nell'ambito di un modello centralizzato di dispacciamento, *central dispatching* o di uno maggiormente decentrato, il *self dispatching*. L'idea dietro l'EB-GL è quella di favorire la migrazione verso il *self dispatch*, definito come modello di riferimento europeo. Ad oggi, infatti, questo è il modello impiegato dalla maggior parte dei Paesi europei, come si vede in Figura 1.17. L'Italia, continua ad adottare il modello centralizzato, che è storicamente il modello che va per la maggiore oltreoceano.

Il Regolamento (EU) 2017/2195 ha consentito ai TSO come Terna che in precedenza adottavano un modello centralizzato di mantenerlo, notificandolo all'Autorità di regolazione competente. Tale parziale apertura consegue dal riconoscimento della principale potenzialità del modello: nell'assetto *central dispatch*, il TSO determina i programmi di produzione e consumo nell'ambito dell'*Integrated Scheduling Process* (ISP). Dispone di tutti gli elementi per movimentare le risorse di dispacciamento più convenienti disponi-

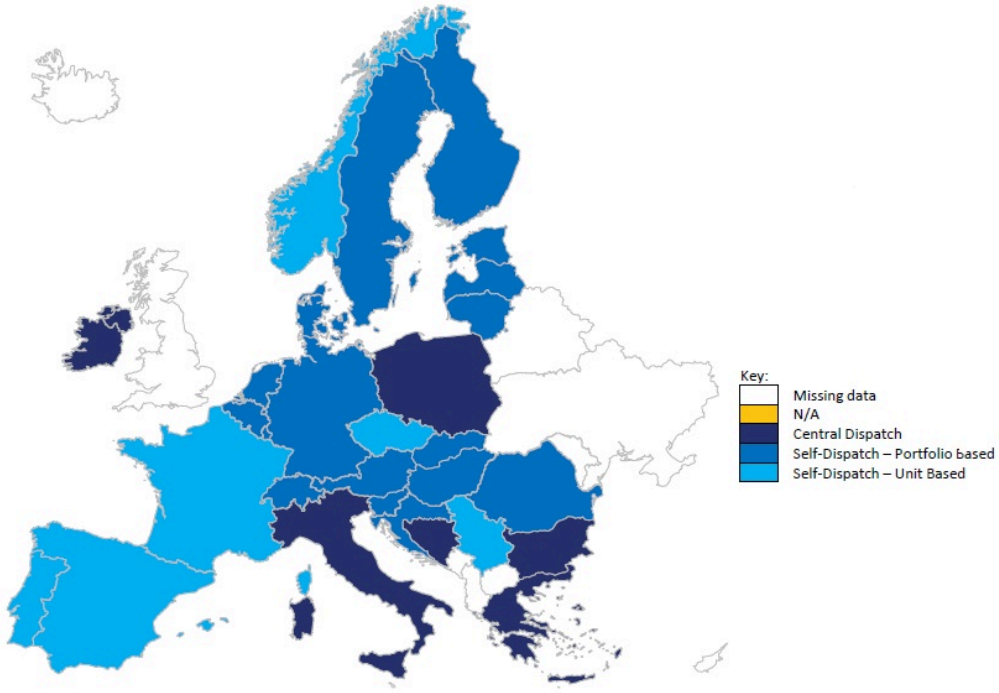


Figura 1.17: Modello di bilanciamento adottato nei Paesi Membri UE. Fonte[45]

bili ed acquisire in sicurezza i servizi richiesti per realizzare le regolazioni centralizzate, di tensione e di frequenza. Il maggior lavoro di regia centrale affidato al TSO si traduce quindi in uno schema in grado di conseguire, seppur con i dovuti aggiustamenti, addirittura una migliore efficienza di esercizio del sistema rispetto al modello *self dispatch*.

Nel modello di *self dispatch*, agli utenti del dispacciamento è invece lasciata la facoltà di determinare piuttosto liberamente i programmi di produzione e consumo degli impianti nel rispettivo portfolio. Questi soggetti (cosiddetti *Balance Responsible Party*, BRP) sono poi tenuti a garantire l'equilibrio tra domanda e offerta all'interno del proprio gruppo di bilanciamento (generazione interna al portfolio e potenza acquistata dall'esterno devono uguagliare consumi interni e cessioni) e sono i responsabili di eventuali sbilanciamenti.

[45] WGAS (2021)

1.4 Mercato a Termine

Produttori, società di vendita, traders, consumatori, abilitati a registrare transazioni sulla PCE, possono stipulare contratti a lungo termine per coprirsi dal rischio di variabilità del prezzo. Sono contrattazioni bilaterali, nelle quali gli operatori indicano la tipologia, il periodo di consegna, il numero dei contratti e prezzo al quale sono disposti ad acquistare/vendere. Questo viene indicato come "differenziale unitario di prezzo", ovvero come differenza rispetto al prezzo che si forma nei mercati spot: non si fissa un vero e proprio prezzo, ma si fa uno "sconto" su quello definito in MPE. Sono negoziabili contratti in risposta al fabbisogno del carico di base o del carico di picco, con periodi di consegna pari al mese, al trimestre o all'anno. Per questi ultimi due orizzonti temporali, è previsto il meccanismo "a cascata": al termine della sessione dell'ultimo giorno di negoziazione, le posizioni sul contratto annuale vengono divise in equivalenti posizioni sui contratti con scadenza inferiore (mensile e trimestrale). Allo stesso modo, una posizione su un contratto trimestrale viene trasformata in equivalenti posizioni sui corrispondenti contratti mensili. Per ulteriori approfondimenti, si rimanda al testo della *Disposizione tecnica di funzionamento n. 01 rev. 05 MTE* (2016). In parole semplici, se nel Mercato a Pronti sono svolte contrattazioni relative alla soddisfazione immediata (o quasi) del fabbisogno energetico, sul Mercato a Termine, vengono contrattate quantità e prezzi per il futuro.

2

Analisi preliminari

In questo capitolo vengono descritti i dati utilizzati per le analisi, a livello di effetto marginale sulla variabile risposta, i costi di approvvigionamento di riserva di energia.

Le analisi preliminari sono svolte con l'obiettivo di valutare quali variabili impattano e come nell'esposizione al rischio di Terna. Il rischio che si vuole definire è quello per il TSO italiano di incorrere in costi per l'approvvigionamento di riserva di energia non programmati e viene analizzato come funzione di particolari variabili, fornite da Terna e di seguito presentate.

I dati forniti hanno originariamente frequenza oraria, la frequenza con cui sono naturalmente prodotti nel mercato elettrico italiano. Per valutare il rischio in un orizzonte temporale di 1 giorno e di 30 giorni, tutte le serie storiche orarie sono state aggregate sulle 24 ore mediante somma algebrica dei valori orari, ottenendo serie storiche giornaliere. Per ogni variabile si dispone di 1734 osservazioni giornaliere, che si estendono nel periodo temporale compreso tra il 1° gennaio 2017 e il 30 settembre 2021.

I modelli elaborati in seguito, sui risultati di queste analisi, utilizzano i dati giornalieri così ottenuti e permettono di calcolare il rischio associato ai costi, su base giornaliera. Questo sia per quanto riguarda il rischio a 1 giorno, che per quello a 30 giorni. Per quest'ultimo come vedremo, verrà impiegato un metodo simulativo per estendere la modellazione, su base giornaliera, ad un orizzonte temporale di 30 giorni.

Il capitolo si apre con la descrizione del fenomeno d'interesse, i costi giornalieri di approvvigionamento di riserva di energia, per i quali vengono mostrate a livello qualitativo le caratteristiche della serie storica a disposizione. Successivamente sono descritte le variabili di cui si compone il dataset, rappresentandone la relazione con la variabile d'interesse e l'evoluzione anno per anno.

2.1 Il dataset

Come evidenziato nel Documento per la Consultazione 325/2021/R/eel di Arera[46], contenente una serie di orientamenti per la definizione di un sistema di incentivazione ai fini della riduzione dei costi di dispacciamento, negli ultimi anni, si è assistito ad un forte aumento dei costi in MSD. Tra le cause, soprattutto, viene riconosciuto "l'aumento delle movimentazioni per vincoli locali di tensione (VRI), in condizioni di basso fabbisogno ed elevata produzione di energia elettrica da fonti non programmabili, in presenza di situazioni potenzialmente vulnerabili a comportamenti non competitivi da parte dei produttori". Inoltre, le esigenze relative alla regolazione di tensione hanno portato a dover estendere il perimetro degli Impianti Essenziali per la Sicurezza del Sistema (IESS), con conseguente aumento dei relativi costi.

Nel documento sopra citato, Arera annuncia un aumento dei costi di dispacciamento anche nel prossimo futuro. Questo infatti risulta associato all'atteso aumento della diffusione delle fonti rinnovabili non programmabili, diretto alla riduzione del numero di impianti programmabili in servizio in esito al mercato dell'energia. Di qui, la necessità di movimentare sul MSD risorse in grado di erogare servizi ancillari (in particolare di regolazione di tensione), nonché di garantire adeguati margini di riserva per far fronte ad una maggiore aleatorietà della produzione.

Nelle analisi che seguono porteremo alla luce nei nostri dati le evidenze espresse da Arera circa l'andamento presente e futuro dei costi e la loro relazione con le variabili sopracitate e fornite da Terna. I vincoli locali di tensione (VRI), la produzione da rinnovabile,

[46] *Documento per la Consultazione n. 325/2021/R/eel* (2021)

la quantità di riserva movimentata, sono tutte variabili rappresentate e analizzate nel par. 2.1.2.

2.1.1 I costi di approvvigionamento

La variabile di interesse c_t, oggetto delle analisi e della modellazione, rappresenta i costi sostenuti da Terna per l'acquisto di riserva di energia in MSD ex-ante, la fase di programmazione, e nel Mercato di Bilanciamento, la fase in tempo reale. Vengono analizzati i costi totali, ottenuti come somma dei costi sostenuti nelle due fasi del mercato.

Il nostro interesse è quello di comprendere la struttura dei costi complessivi, a fini predittivi, ma le due fasi di contrattazione si distinguono per il differente sottostante e il differente fine con cui viene scambiata energia.
Nel MSD ex-ante, la fase di programmazione, Terna acquista i servizi relativi all'approvvigionamento di riserva di energia, costruendosi uno spazio di manovra per far fronte alla risoluzione delle congestioni, alla costituzione dei margini di riserva secondaria e terziaria di potenza in anticipo rispetto al tempo reale. Per questo motivo, in MSD si verifica molto più frequentemente l'accensione, piuttosto che lo spegnimento di poli produttivi, per non ridurre risorse in vista del bilanciamento. I costi si formano nel momento in cui risulta necessario accendere un'unità produttiva, tra quelle abilitate, che, da esito del Mercato dell'Energia, risulta spenta.
Nel Mercato di Bilanciamento (MB), la fase in tempo reale, Terna acquista i servizi relativi alla regolazione di frequenza e tensione, pertanto legati all'attivazione della riserva. Quest'ultima fase si svolge, infatti, nel giorno stesso a quello cui si riferiscono le offerte, ovvero in tempo reale, e i costi si formano nel momento in cui Terna è costretta ad uscire dallo spazio di manovra creato in MSD ex-ante.

La serie storica dei costi di approvvigionamento giornalieri totali, ottenuti sommando algebricamente i costi in MSD ex-ante e in MB, fornita da Terna come variabile di interesse della nostra analisi, è mostrata in Figura 2.1. Il periodo temporale a cui si riferisce, va dal 1° gennaio 2017 al 30 settembre 2021, per un totale di 1734 osservazioni.

Nel grafico la linea rossa verticale corrisponde al 30 settembre 2020. Per poter stimare e poi validare il modello, si considera la ripartizione in insieme di stima e insieme di verifica e tale linea segna la ripartizione nei due sottoinsiemi: la stima dei modelli viene effettuata sulla finestra 1° settembre 2017 - 30 settembre 2020, per un totale di 1369 osservazioni,

mentre la verifica dei modelli stimati è svolta sui 365 giorni a venire, fino al 30 settembre 2021. Tutte le analisi preliminari presentate in seguito mostrano l'andamento delle variabili nel periodo di stima, fino al 30 settembre 2020.

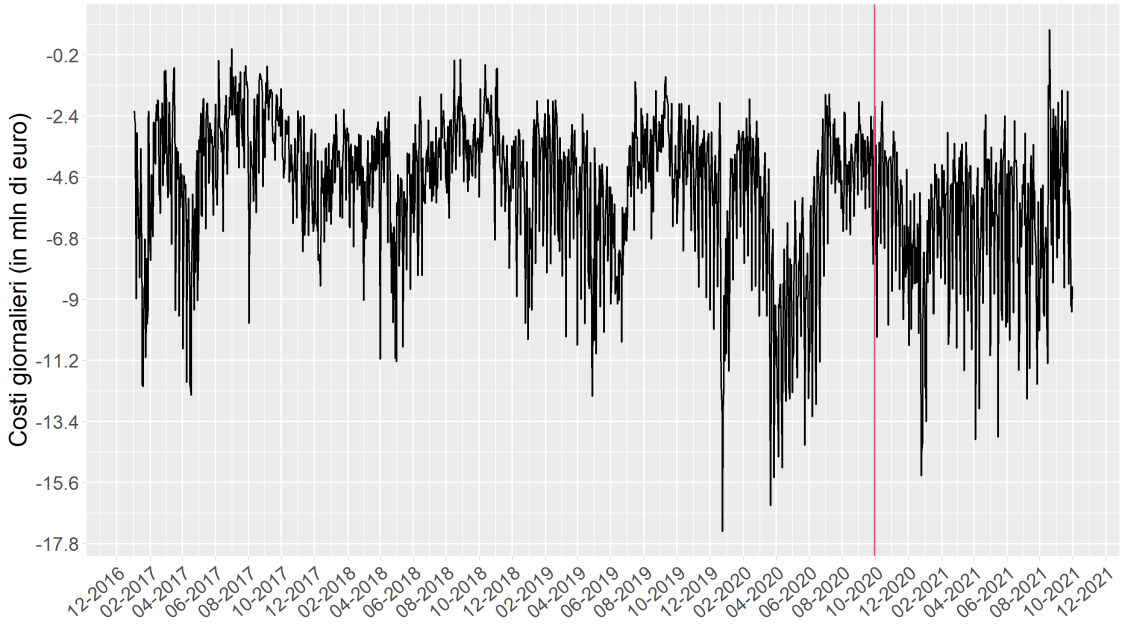


Figura 2.1: Serie storica dei costi giornalieri sostenuti da Terna in MSD (in milioni di euro) dal 1/1/2017 al 30/9/2021.

Ciò che a prima vista si coglie osservando la serie dei costi, è il loro graduale aumento, in valore assoluto, arrivando a picchi di 17.349.663 €/MWh giornalieri a fine 2019, e 16.422.518 €/MWh giornalieri all'inizio del secondo trimestre del 2020. Inoltre, si nota un aumento della variabilità dei costi stessi nel 2021.

Secondo il Documento per la Consultazione 325/2021/R/eel di Arera[47], già dal 2016 i costi delle movimentazioni per esigenze di regolazione di tensione sono aumentati, soprattutto per effetto della riduzione degli impianti programmabili accesi in esito al mercato dell'energia.
Nel corso del 2016 e del 2017 si sono registrati aumenti dei costi di dispacciamento causati da arbitraggi sugli sbilanciamenti, soprattutto da parte degli utenti del dispacciamento

[47] *Documento per la Consultazione n. 325/2021/R/eel* (2021)

di unità di consumo, poi superati con le nuove modalità di calcolo del segno di sbilanciamento zonale introdotte con la Delibera 419/2017/R/eel[48] della Direzione Mercati Energia all'Ingrosso e Sostenibilità Ambientale (DMEA) e adottate da Terna a partire dall'1 settembre 2017. Sempre nel 2017 si è verificato un aumento dei costi per l'essenzialità, in seguito all'aumento degli impianti assoggettati al regime di reintegro dei costi.
Negli ultimi anni è stato registrata una graduale riduzione della produzione eolica, fondamentale per mantenere in condizioni di sicurezza il sistema elettrico nazionale, o porzioni del medesimo, oltre che per la gestione di congestioni locali verificate nelle principali direttrici maggiormente soggette a tali fenomeni. Infine, nel 2020, a causa dell'emergenza Covid e della correlata riduzione della domanda, si sono registrati ulteriori aumenti dei costi di dispacciamento in seguito all'esigenza di movimentare gli impianti sul MSD, assicurando la disponibilità di adeguati margini di riserva e - soprattutto - per esigenze di regolazione di tensione.

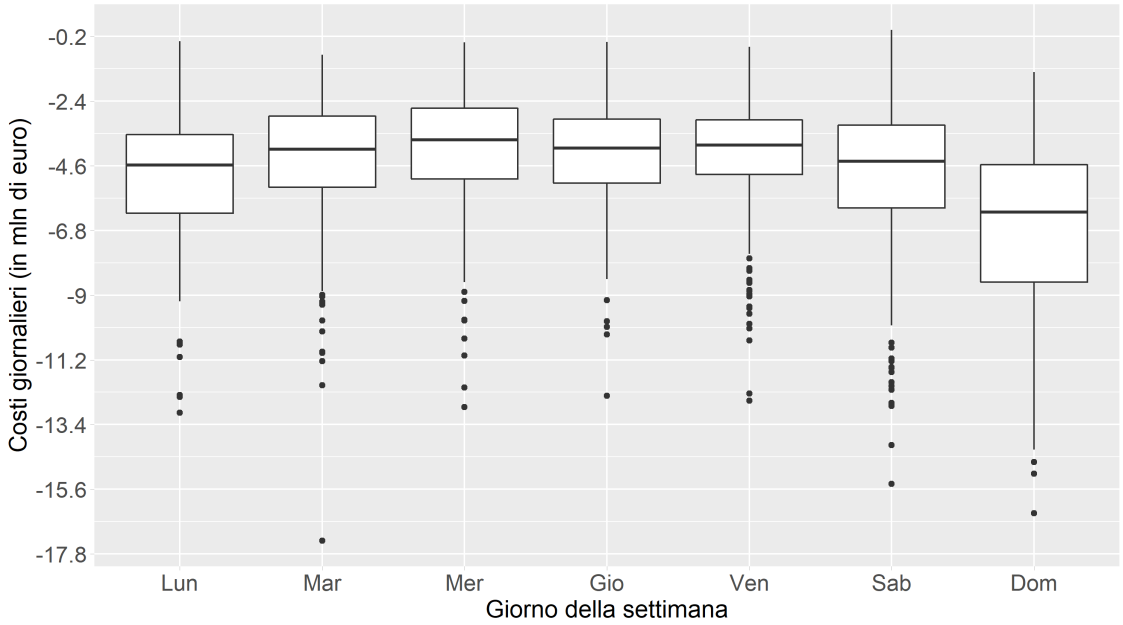


Figura 2.2: Boxplot dei costi giornalieri di approvvigionamento per giorno della settimana.

Un'altra caratteristica strutturale dei costi giornalieri è data dalla loro stagionalità, presente sia a livello annuale che settimanale. La componente settimanale è osservabile in

[48] *Delibera 08 giugno 2017 n. 419/2017/R/eel* (2017)

Figura 2.2: i costi sono più elevati durante i giorni festivi, oltre che maggiormente variabili. Anche i prefestivi e i postfestivi, corrispondenti al sabato e al lunedì, registrano costi leggermente più elevati e variabili, mentre i restanti giorni feriali, dal martedì al venerdì, si attestano all'incirca sullo stesso livello.

Queste evidenze motivano l'inclusione di variabili di calendario nella modellazione delle componenti strutturali. Si noti che i costi sono entità dal valore negativo negativi, per cui valori piccoli (negativi) della variabile risposta rappresentano costi maggiori.

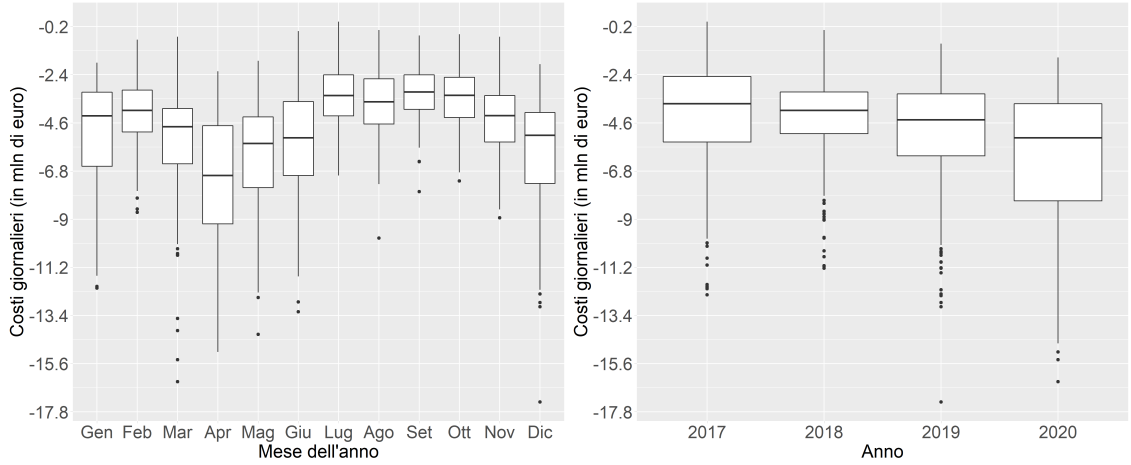


Figura 2.3: Boxplot dei costi giornalieri di approvvigionamento per mese dell'anno (a sinistra) e per anno (a destra).

La struttura temporale dei costi si caratterizza anche per una componente annuale, che emerge osservando i costi giornalieri aggregati per mese dell'anno, in Figura 2.3 o considerando la serie dei costi mensili, in Figura 2.4, ovvero la somma sui 30 giorni dei costi giornalieri.

Nel grafico a sinistra di Figura 2.3, si evidenzia quanto visto anche dal grafico della serie temporale, ovvero costi più elevati e variabili in primavera, in particolare nel mese di aprile e marzo, che come dicembre presenta *outliers* anche molto elevati. I mesi a costi più contenuti e meno variabili sono quelli estivi, per le temperature stabili e la riduzione di fabbisogno per le chiusure estive. Nel grafico a destra si trova il boxplot dei costi giornalieri, da inizio 2017 a settembre 2020, raggruppati per anno. È evidente quanto detto a inizio del paragrafo sull'andamento dei costi: nel corso degli anni sono aumentati in media, in mediana (linea centrale nella scatola), ma anche nei quantili, valori minimo, massimo e valori estremi. Nel grafico in Figura 2.4 è riportata l'intera serie da gennaio 2017 a settembre

2021, con la ripartizione in insieme di stima e verifica rappresentata dalla linea rossa, nel mese di settembre 2020. I costi di approvvigionamento sono particolarmente elevati nei mesi primaverili e autunnali, in cui a seconda dello scostamento inatteso dalle temperature medie, si effettuano movimentazioni non programmate sul MSD più frequentemente. Un picco molto elevato (corrispondente al costo mensile più sostenuto) riguarda il mese di aprile 2020, mese in cui il fenomeno appena descritto si è unito all'effetto della pandemia da Covid-19. La ripresa si è verificata durante l'estate e l'autunno, per poi precipitare nuovamente a inizio 2021 (secondo picco massimo di costi). La dipendenza lineare tra costi mensili, appena significativa, è tra m e $m - 12$, quindi di anno in anno i costi risultano in relazione tra di loro, com'è facile pensare, per la ciclicità delle stagioni e degli eventi anno dopo anno.

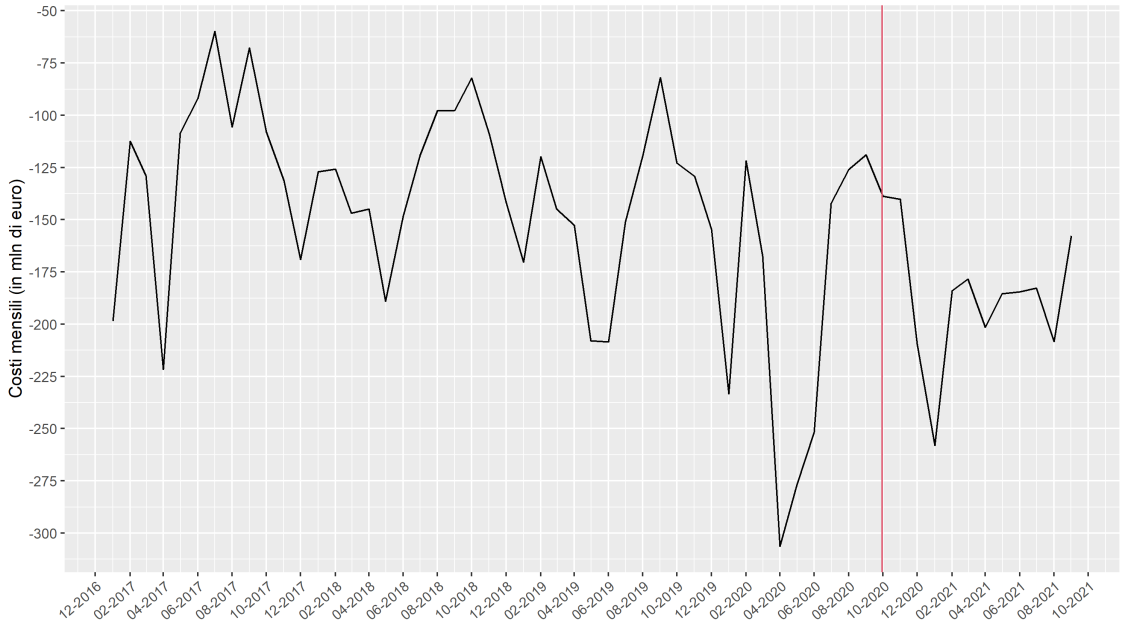


Figura 2.4: Serie storica dei costi mensili sostenuti da Terna in MSD (in milioni di euro) dal 1/1/2017 al 30/9/2021.

2.1.2 Le variabili

Il dataset per le analisi si compone di variabili fornite da Terna, che come la serie dei costi, si estendono dal 1° gennaio 2017 al 30 settembre 2021. Nelle nostre analisi, c_t, la serie dei costi è la variabile risposta da modellare, mentre le variabili elencate di seguito, costituiscono l'insieme $\mathbf{x}_t$ (o equivalentemente $\mathcal{I}_t = \{x_s, s \leq t\}$), ovvero l'insieme dell'informazione disponibile fino al tempo t, rispetto al quale sono condizionate le previsioni.

Anche per queste variabili si dispone dei dati a frequenza giornaliera per t, $t = (1, \ldots, n)$, n=1734 osservazioni per ogni variabile. Per la stima dei modelli presentati in seguito verranno utilizzate le prime n=1369 osservazioni, rappresentate nei grafici a venire di:

- D_t: serie storica giornaliera del fabbisogno di energia nazionale, ovvero la domanda di energia (espressa in MWh) formulata in esito ai mercati dell'energia;
- $Wind_t$, PV_t, $Hydro_t$: serie storiche giornaliere rispettivamente della produzione di energia eolica, energia fotovoltaica, energia idroelettrica da centrali a bacino (in MWh);
- $aFRR_t$: serie storica giornaliera del fabbisogno di riserva secondaria nazionale (in MWh) (si veda il paragrafo 1.3.6, sezione "L'approvvigionamento di riserva");
- RR_t: serie storica giornaliera del fabbisogno di riserva terziaria nazionale (in MWh) (si veda il paragrafo 1.3.6, sezione "L'approvvigionamento di riserva");
- VRI_t: serie storica giornaliera dei vincoli a rete integra, ovvero del numero di unità produttive in esercizio sul territorio nazionale a supporto della messa in sicurezza del sistema elettrico (si veda il paragrafo 1.3.6, sezione "La regolazione di tensione");
- TTF_t: serie storica giornaliera del prezzo del gas nel TTF (*Title Transfer Facility*), il mercato virtuale del gas naturale con sede in Olanda. I valori dell'indice, media aritmetica delle quotazioni giornaliere riferite al mese di fornitura, sono espressi in €/MWh.
- ETS_t: serie storica giornaliera del prezzo dei certificati verdi, in inglese *European Union Allowances*, EUA (espresso in €/ton).

In aggiunta alle precedenti variabili, sono state utilizzati le seguenti variabili di calendario:

- T_t: variabile rappresentante il trend, ossia la dinamica di lungo periodo, dal giorno 1 al giorno t;

- DY_t: variabile indicante il giorno dell'anno, e rappresentante la periodicità annuale dei dati; vettore costituito dal susseguirsi della sequenza 1,...,365 per ogni anno (366 per il 2020, anno bisestile);
- DW_t: variabile indicante il giorno della settimana, e rappresentante la periodicità settimanale dei dati; vettore costituito dal susseguirsi della sequenza 1,...,7 per ogni settimana (da 1 per il lunedì a 7 per la domenica);
- $bank_t$: variabile dummy per le festività, escluse le domeniche, ufficialmente riconosciute come giorni non lavorativi, in inglese *bank holidays*; assume valore 1 se il giorno t è giorno festivo non lavorativo, 0 altrimenti.

Di seguito vengono presentate le variabili elencate in precedenza, con analisi bivariate della relazione marginale con la variabile risposta, per anno.

I consuntivi del fabbisogno nazionale

La prima variabile in analisi, D_t, rappresenta i consuntivi del fabbisogno nazionale di energia elettrica.

Dal 1° gennaio 2021 sono entrate in vigore le modifiche alla struttura zonale previste dalla Delibera 103/2019/R/eel dell'Ufficio Speciale Regolazione Euro-Unitaria (REU)[49]. I dati forniti da Terna sui consuntivi e sulla produzione da fonti rinnovabili, che vedremo in seguito, sono riferiti al periodo dal 2017 al 2021 e risultano tutti aggiornati alla nuova suddivisione in zone.

Si dispone delle serie dei consuntivi per ciascuna delle sette zone. La serie nazionale dei consuntivi di fabbisogno è ottenuta sommando (algebricamente) il dato di ogni serie zonale.

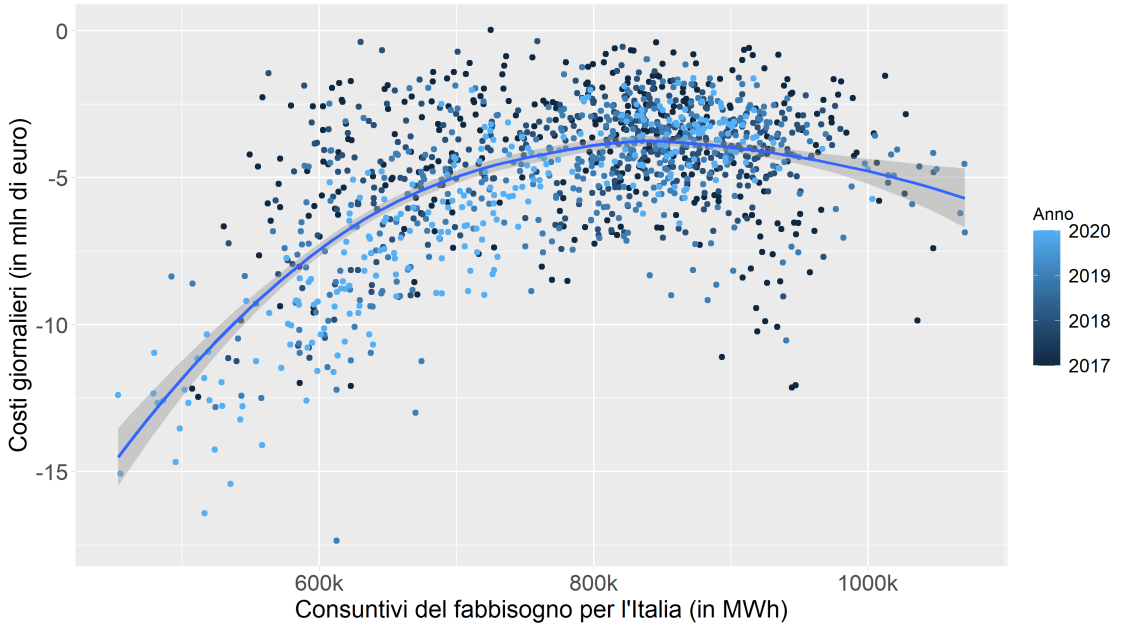


Figura 2.5: Relazione tra i consuntivi nazionali giornalieri e i costi di approvvigionamento giornalieri, suddivisi per anno.

Il grafico in Figura 2.5 rappresenta l'andamento dei costi giornalieri in MSD rispetto al valore dei consuntivi per il fabbisogno di energia: la curva stimata in modo non para-

[49] *Delibera 19 marzo 2019 n. 103/2019/R/eel* (2017)

metrico[50] è a forma parabolica. Si nota infatti, che per una domanda elevata di energia (coda destra), i costi aumentano come normale risposta di mercato. Tuttavia anche per valori particolarmente bassi della domanda, si assiste ad un aumento dei costi di approvvigionamento (coda sinistra). È qui che si materializzano le condizioni di potere di mercato: la domanda è poca su mercato dell'energia, ma il sistema non porta a rispettare molti dei vincoli di sicurezza. Le centrali commissionate non sono ben distribuite sul territorio ed è Terna che aggiusta il parco di produzione. In queste condizioni gli operatori hanno potuto capire quali sono state in passato le esigenze di Terna e quindi di arbitrare tra i due mercati, ME e MSD. Quando l'operatore sa che ha un'alta probabilità di essere accettato in MSD, perché Terna ha certe esigenze rilevate in passato e che ora l'operatore sa prevedere, l'operatore preferisce non uscire in ME e puntare su MSD, poiché il prezzo in MSD è di molto superiori a quello in ME, vista la sua natura locale e poco competitiva.
La transizione energetica punta a ridurre la domanda di energia, cercando al contempo di rispondervi con produzione rinnovabile, ma la vera sfida è la transizione verso questa situazione, facendo decrescere i costi, non aumentandoli come in questo caso. Per fare ciò, Terna come operatore di sistema deve risolvere le condizioni che portano a lasciare ad alcuni operatori potere di mercato, annullando le loro aspettative di essere accettati in MSD anziché in ME.

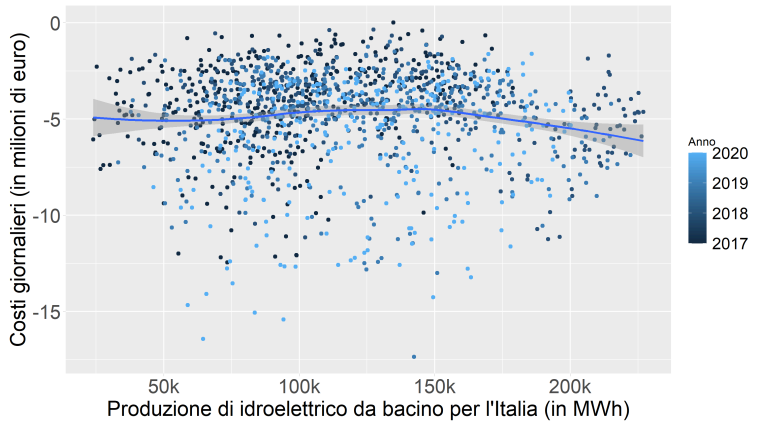
Si può notare quanto appena detto grazie alla suddivisione per anno in Figura 2.5: dal 2017, al quale corrisponde un colore più scuro, fino al settembre 2020, al quale corrisponde il blu chiaro, nella parte relativa a consuntivi bassi e costi elevati si trovano valori più recenti. Questo perché la domanda nazionale di energia degli ultimi anni è diminuita, e a questa riduzione sono corrisposti costi di approvvigionamento della riserva più elevati.

[50] Il metodo non parametrico impiegato per una prima stima a livello descrittivo della relazione tra la risposta e le variabili in analisi in questo capitolo è la regressione polinomiale locale, nota come *loess* ((Cleveland et al., 1992))

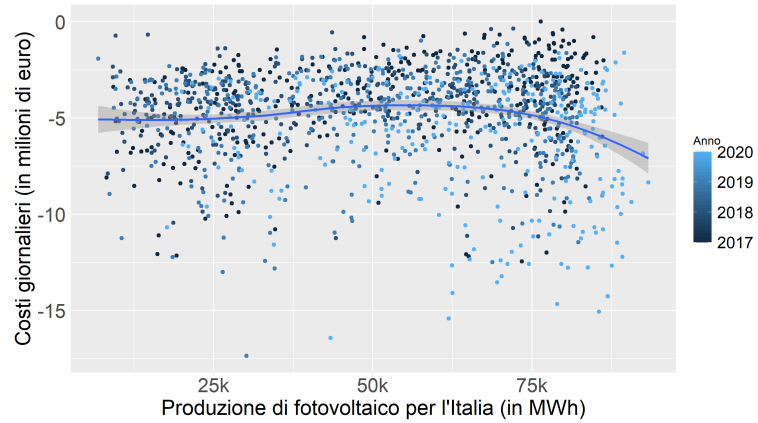
La produzione da fonti rinnovabili

Accanto ai dati del consuntivo del fabbisogno di energia, Terna ha reso disponibile anche i dati relativi alla produzione da alcune fonti rinnovabili tra le più importanti in Italia: idroelettrico da bacino, fotovoltaico ed eolico, nelle variabili $Hydro_t$, PV_t e $Wind_t$, rappresentate in Figura 2.6. Anche questi dati vengono raccolti con ripartizione in zone, per cui originariamente si dispone delle sette serie storiche per ogni zona, per ognuna delle tre variabili. Disponendo di costi a livello nazionale, anche in questo caso risulta più utile considerare per le tre fonti rinnovabili la produzione accorpata sull'intero territorio.

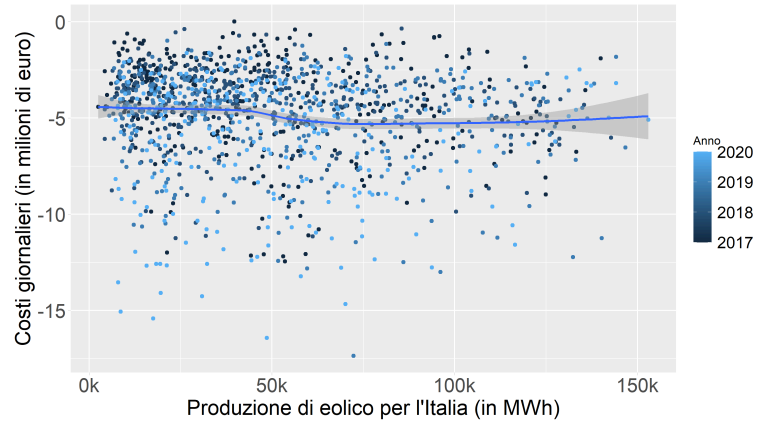
I costi di approvvigionamento di energia sono all'incirca costanti all'aumentare della produzione da idroelettrico, in Figura 2.6a, fotovoltaico, Figura 2.6b ed eolico, Figura 2.6c, ma per valori alti della produzione delle prime due i costi aumentano, mentre restano stabili per l'eolico.



(a) Idroelettrico da bacino



(b) Fotovoltaico



(c) Eolico

Figura 2.6: Relazione tra la produzione a livello nazionale di energia dalle tre fonti rinnovabili elencate e i costi di approvvigionamento giornalieri, suddivisi per anno.

La riserva secondaria, o *aFRR*

Un'altra variabile, fornita da Terna, e utile alla definizione dei costi per l'approvvigionamento di riserva è la quantità di riserva secondaria per i servizi di regolazione di frequenza/potenza, $aFRR_t$. La aFRR (*automatic Frequency Restoration Reserve*) corrisponde all'attuale riserva secondaria, nella nomenclatura europea[51].

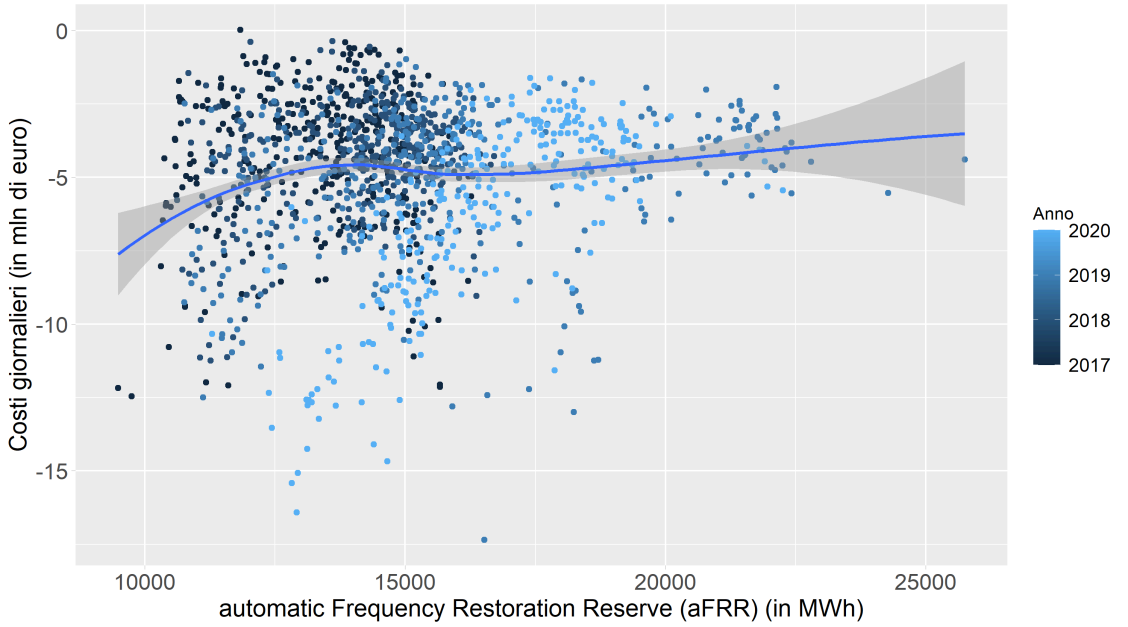


Figura 2.7: Relazione tra la riserva secondaria a salire e i costi di approvvigionamento giornalieri, suddivisi per anno.

In Figura 2.7 è mostrata la relazione tra i costi sostenuti da Terna in MSD e la quantità di riserva secondaria nazionale attivata "a salire", per sostenere il bilanciamento tra carichi e prelievi in seguito ad un calo della frequenza, ad esempio a causa dello scatto di un generatore, con un aumento della produzione immessa nella rete o con una riduzione della produzione prelevata. La riserva secondaria "a scendere" riguarda invece la predisposizione di unità abilitate a ridurre l'immissione o incrementare il prelievo, in seguito ad

[51] La riserva secondaria è approvvigionata per aggregati di zone (Continente, Sicilia e Sardegna) e Terna ha infatti fornito i dati relativi alla riserva secondaria, sia "a salire", che "a scendere", per i tre aggregati di zone. L'aggregato Continente costituisce l'aggregato più significativo ai fini del servizio di riserva secondaria, ma per completezza si considera la somma (algebrica) dei valori riferiti alle tre macro zone.

un aumento della frequenza nella rete, a causa per esempio di una congestione. Ai fini del calcolo dei costi risulta più rilevante l'attività di regolazione secondaria a salire, per cui di seguito sarà la serie di riferimento. La curva che rappresenta la relazione tra le due variabili è stimata in modo non parametrico. Essa mostra che per quantità minime di riserva secondaria a salire attivata, i costi di approvvigionamento sono elevati, ma all'aumentare della riserva secondaria attivata, i costi scendono leggermente per poi stabilizzarsi. Grazie alla suddivisione per anno è possibile notare come in anni più recenti, soprattutto nel 2020, la quantità di riserva secondaria a salire è aumentata, e ora quelle quantità a cui si era abituati a rispondere negli anni precedenti, corrispondono a costi più elevati, perché sono quantità basse per gli standard attuali, mentre quelle che un tempo erano quantità elevate, nel 2020 sono quantità ripagate con costi inferiori. Si assiste sostanzialmente ad una traslazione della relazione verso la movimentazione di quantità di riserva più elevate, come se fosse cambiata la situazione nel *background* di riferimento. In anni recenti infatti la diffusione delle fonti rinnovabili non programmabili ha ridotto la quantità di energia definita a consuntivo, come visto in Figura 2.5, al contempo richiedendo un dispiego maggiore della riserva per far fronte a necessità impreviste.

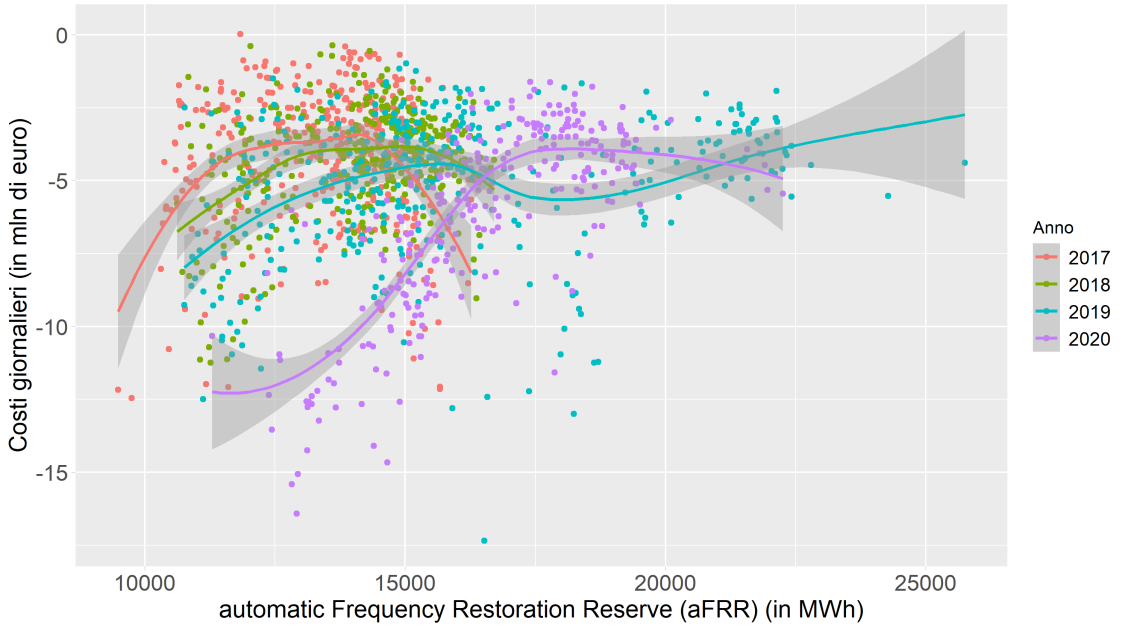


Figura 2.8: Stima della relazione tra la riserva secondaria a salire e i costi di approvvigionamento giornalieri per anno.

Per anno la relazione stimata è come appare in Figura 2.8. Nel 2017 (in rosso) e il 2018

(in verde) la curva è a forma di U rovesciata: un fabbisogno di secondaria elevato (coda destra) impatta sui programmi e come operatore di sistema, Terna è costretta a sostenere costi elevati per poterne disporre. L'aumento dei costi in relazione a livelli bassissimi di secondaria (coda sinistra), è un effetto secondario legato alla situazione in cui si materializza tale necessità, che dipende dalle condizioni del sistema. Il costo sale per valori molto bassi di secondaria perché ci si ritrova in condizioni di funzionamento del sistema di bassissimo fabbisogno.
Il 2019, come il 2020, ha un'iniziale flessione che tende a stabilizzarsi. Tuttavia, il 2019 presenta un andamento dei costi meno variabile, seppure i valori estremi dell'aFRR (maggiori di 22000 MWh) sono proprio stati registrati in quest'anno, anche se con costi contenuti rispetto a quantità inferiori di aFRR che però hanno richiesto il costo massimo registrato per essere movimentate. Il 2020 invece si caratterizza per una graduale riduzione dei costi all'aumentare dell'aFRR, ma con più valori elevati rispetto a quelli degli anni precedenti.

La riserva terziaria, o *RR*

Assieme alla riserva secondaria, si dispone anche della serie storica della quantità di riserva terziaria, sia "a salire" che "a scendere", attivata dal 2017 al 2021, per ogni giorno.

L'approvvigionamento di riserva terziaria è compito di Terna, che dispone, attraverso questa quantità, il ripristino delle riserve in seguito all'attivazione della secondaria.

Il fabbisogno di riserva terziaria (totale) a salire è costituito dalla somma del fabbisogno di riserva pronta a salire e del fabbisogno di riserva di sostituzione a salire (che comprende la riserva terziaria rotante a salire).

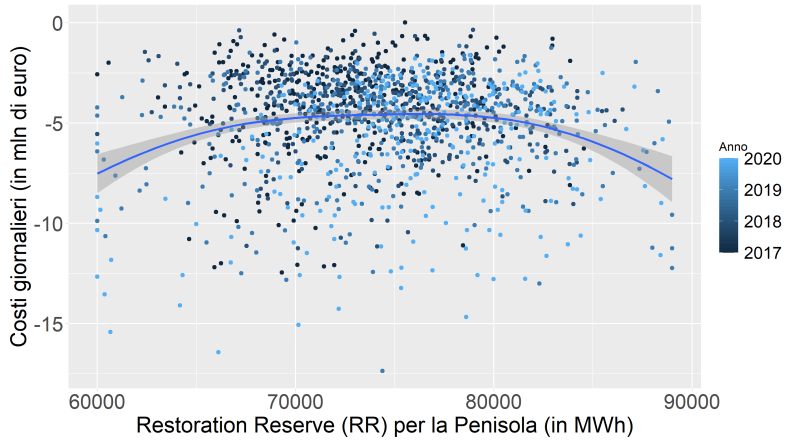
La riserva terziaria, così come la secondaria, è approvvigionata per aggregati di zone (Continente, Sicilia e Sardegna) e Terna ha infatti fornito i dati relativi alla riserva terziaria, sia "a salire", che "a scendere", per ognuna delle tre tipologie (ad eccezione della riserva pronta che è definita esclusivamente a salire) e per i tre aggregati di zone.

Per un'informazione più completa si considera la somma (algebrica) dei valori riferiti alle tre macro zone, ma verranno esposti anche i grafici relativi alla RR per le tre macro

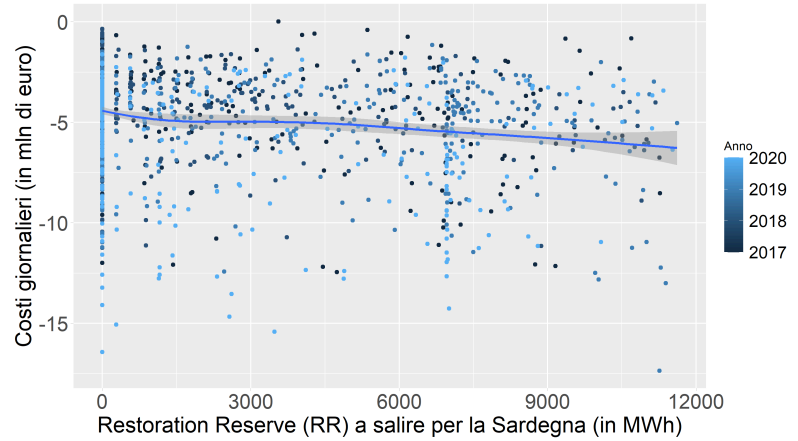
zone separatamente in Figura 2.10, utili ad una migliore comprensione del grafico della loro aggregazione.

In Figura 2.9 è stimata non parametricamente la relazione tra i costi e RR_t, la variabile della riserva terziaria (totale) a salire, indicata con la denominazione europea di *Restoration Reserve*, *RR*. La curva accenna ad un andamento parabolico, ma è da notare anche come nelle estremità aumenti molto la variabilità, per via della scarsità di osservazioni. Inoltre, grazie alla suddivisione per anno, è possibile notare che i costi maggiori sono associati ad anni più recenti, soprattutto al 2020, mentre la maggior parte di concentrano in una nuvola: infatti la curva nella parte centrale, in cui si trova la maggior parte dei dati, tende a stabilizzarsi.

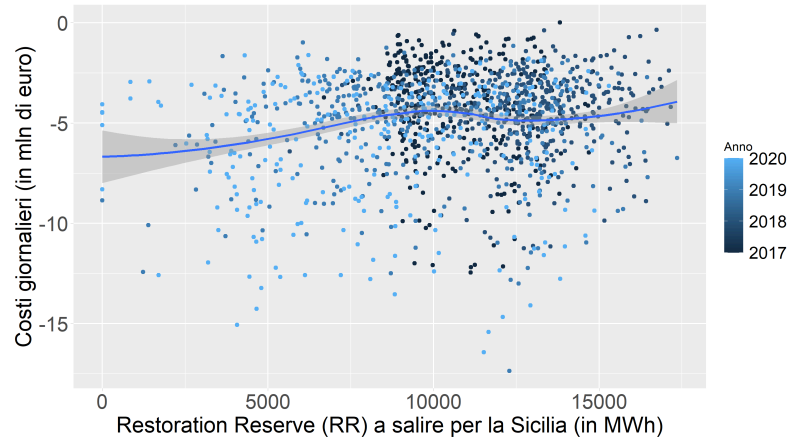
Figura 2.9: Relazione tra la riserva terziaria nazionale a salire e i costi di approvvigionamento giornalieri, suddivisi per anno.



(a) Penisola



(b) Sardegna



(c) Sicilia

Figura 2.10: Relazione tra la riserva terziaria a salire per i tre aggregati zonali e i costi di approvvigionamento giornalieri, suddivisi per anno.

I vincoli a rete integra

Un'altra variabile impiegata nelle analisi è il numero di unità produttive in servizio per garantire la regolazione di tensione, indicato con il nome di Vincoli a Rete Integra (VRI). All'aumentare delle unità produttive accese per rispondere al bisogno, aumentano i costi: la curva indica questa relazione direttamente proporzionale, ed è rappresentata da una sua stima non parametrica in figura 2.11.

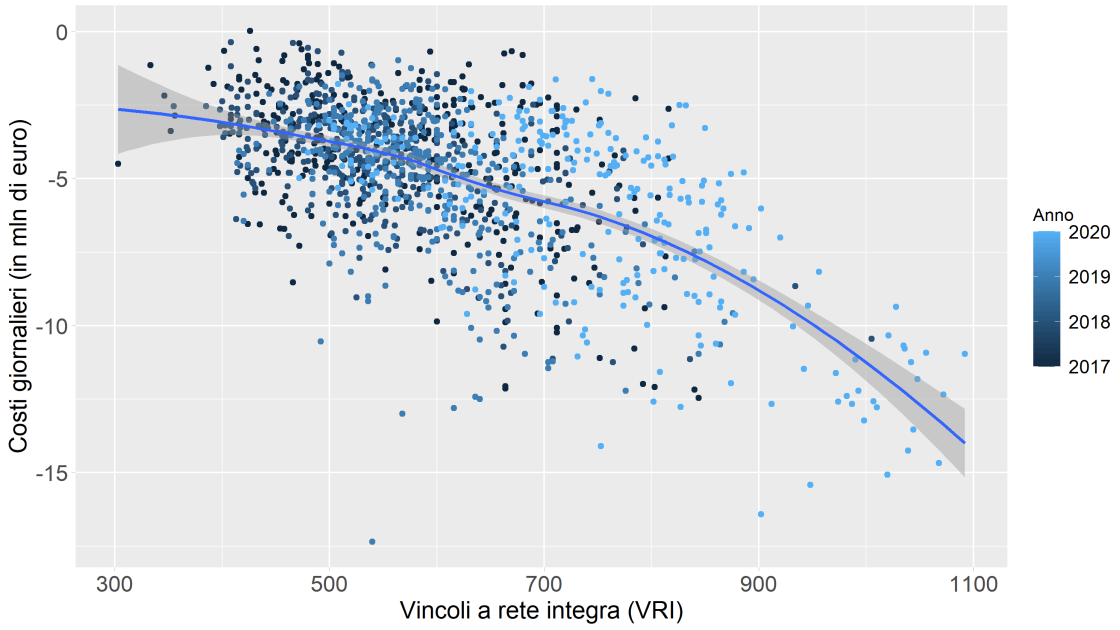


Figura 2.11: Relazione tra i vincoli a rete integra e i costi di approvvigionamento giornalieri, suddivisi per anno.

Il prezzo del gas nel mercato TTF

Il prezzo dell'energia elettrica in Italia è influenzato da diversi fattori, come la domanda, la stagione, le temperature, ma una componente fondamentale del prezzo è il costo per la produzione dell'energia stessa.
L'energia elettrica in Italia viene prodotta ancora per lo più da fonti tradizionali, anche se negli ultimi anni la produzione elettrica da carbone si è già ridotta di circa l'80% grazie alla crescita delle rinnovabili e al crescente prezzo della CO_2, che penalizza il carbone rispetto al gas. Quest'ultimo ne costituisce quasi il 45%, il carbone poco più del 6%, con la quota rinnovabili in rapido aumento.
Il gas è quindi il combustibile più utilizzato per la produzione elettrica in Italia, ed è dunque facilmente intuibile come il prezzo dell'energia elettrica sia strettamente legato a quello della principale materia prima utilizzata per produrla.
Lo stretto legame fra energia e gas non è dato solo dalla quantità di energia prodotta attraverso l'impiego del gas, ma anche dal ruolo che questo ha nella formazione del prezzo nel Mercato del Giorno Prima. Le diverse tipologie di centrali, avendo differenti caratteristiche tecnologiche, producono con programmi e costi diversi e, dunque, influenzano il prezzo in modo diverso. La produzione da fonti rinnovabili infatti, essendo intermittente e non programmabile, ha la priorità nella vendita a mercato, ma è il gas la risorsa che copre il fabbisogno residuo, determinando quindi molto spesso il prezzo marginale. Il costo dell'energia elettrica in MSD sostenuto da Terna è, per la natura poco liquida di MSD, molto maggiore del prezzo dell'energia definito nel mercato dell'energia, che è da suo canto strettamente connesso al costo del gas. Per questo può essere utile considerare l'andamento del prezzo del gas per la modellazione del costo di approvvigionamento di riserva.

In seguito è presentato il prezzo del gas formatosi sul mercato TTF (*Title Transfer Facility*), il mercato virtuale del gas naturale con sede in Olanda, uno dei principali mercati di riferimento per lo scambio del gas in Europa, grazie anche alla sua localizzazione centrale. I valori dell'indice sono calcolati come media aritmetica delle quotazioni giornaliere riferite al mese di fornitura, espressa in €/MWh. Solitamente è poi convertita in €/Smc, moltiplicando il valore in €/MWh per il fattore di conversione 0,0107 riferito a un potere calorifico pari a 0,03852 GJ/Smc. Nelle nostre analisi si considera il prezzo in €/MWh, l'unità di misura convenzionale di tutte le fonti di energia.

A questo punto potremmo chiederci per quale motivo non si faccia riferimento alla serie dell'indice del prezzo del gas italiano, definito nel Punto di Scambio Virtuale (PSV), il mercato italiano del gas gestito dal GME. Innanzitutto, occorre considerare che l'Olanda, grazie ai suoi giacimenti, è stata tra i primi paesi in Europa ad utilizzare il gas come combustibile, sviluppando un vero e proprio mercato all'ingrosso, punto di riferimento privilegiato anche grazie alla sua posizione centrale e di snodo. Le dinamiche del mercato che incidono sui prezzi del TTF vanno così a ripercuotersi sul nostro PSV. Inoltre, va considerato che la piattaforma italiana non è sufficientemente liquida e non può ancora costituire un riferimento di mercato adeguato, come invece la piattaforma TTF olandese, dove sono negoziati i prezzi di vendita e acquisto di gas tra i maggiori operatori europei.

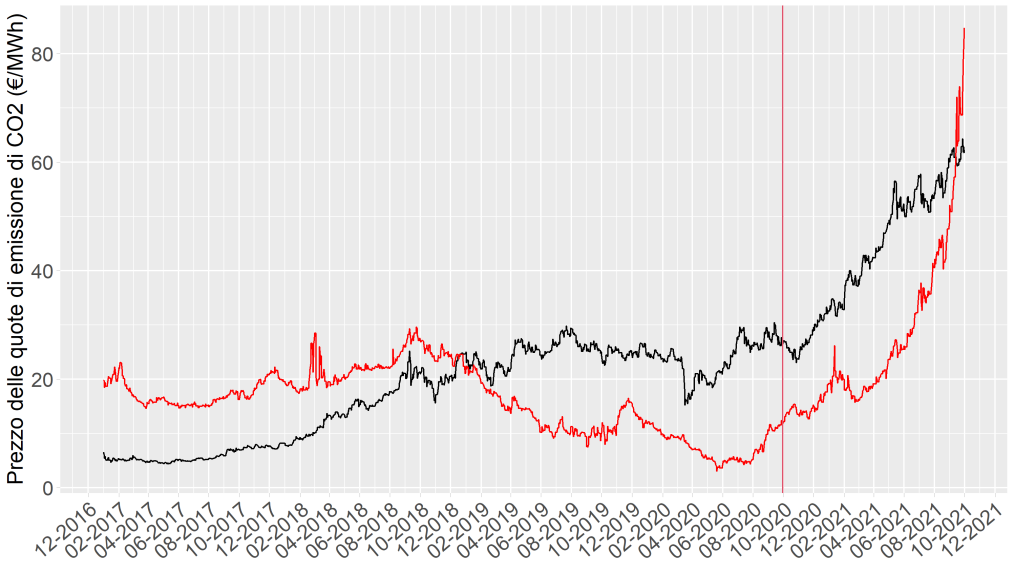


Figura 2.12: Serie storica del prezzo giornaliero del gas naturale sul mercato TTF e del prezzo delle emissioni di CO2 fino al 30 settembre 2021.

In Figura 2.12 è mostrata la serie storica relativa al prezzo del gas naturale sul mercato TTF, informazione contenuta nella variabile TTF_t, dapprima fino al 30 settembre 2021, successivamente con uno zoom sul periodo di stima, ovvero fino al 30 settembre 2020 (linea rossa) in Figura 2.13.

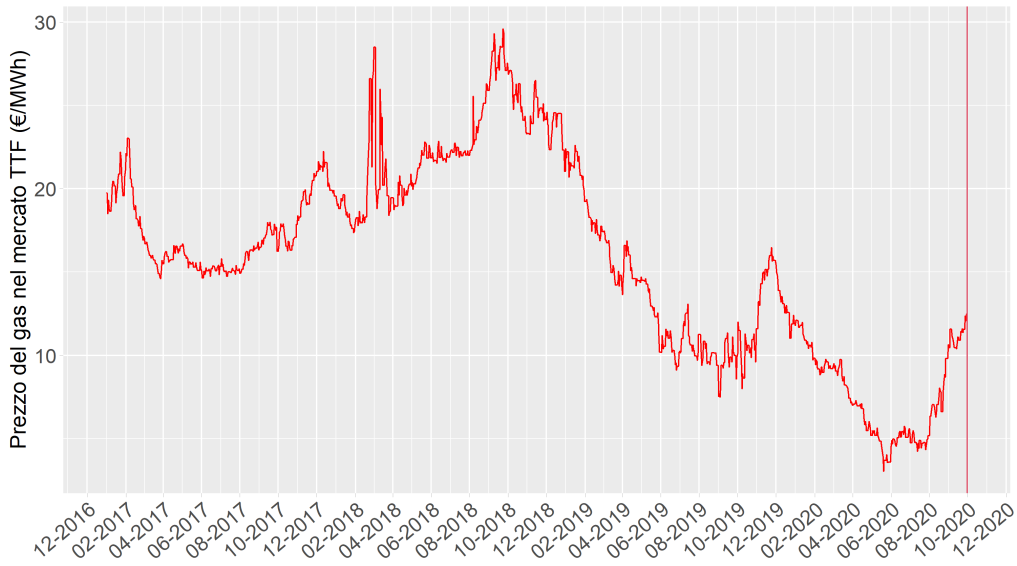


Figura 2.13: Serie storica del prezzo giornaliero del gas naturale sul mercato TTF fino al 30 settembre 2020.

Questa serie è rappresentata in colore rosso, accanto alla serie dei prezzi delle emissioni di CO_2, discusse in seguito. È evidente che entrambe le serie sono non stazionarie, ma soprattutto che dal 30 settembre 2020, il prezzo del gas è salito per arrivare ad un'impennata senza precedenti, così come quello delle quote di CO_2. A partire dall'estate del 2021 il prezzo del gas TTF è aumentato come mai prima, in seguito alla ripresa dell'economia e quindi dei consumi dopo la pandemia. Oltre ai consumi, incidono sul prezzo del gas altri fattori di natura geopolitica e congiunturali. In concomitanza con l'aumento autunnale dei consumi europei ad esempio, si è verificato un calo delle consegne di gas verso l'Europa da parte di Russia e Norvegia, con un incremento delle stesse verso i mercati asiatici più redditizi, che ha portato ad un aumento del prezzo del gas europeo. A fine febbraio 2022 inoltre, dopo l'invasione russa in Ucraina, i mercati energetici hanno subito degli scossoni che potrebbero implicare probabili aumenti nelle previsioni future dei prezzi del gas e dell'energia elettrica. In questa situazione di grande incertezza, almeno per quanto riguarda il gas, c'è da sottolineare anche la stagionalità, che comporta di solito una diminuzione del prezzo durante il periodo più mite in primavera ed estate.

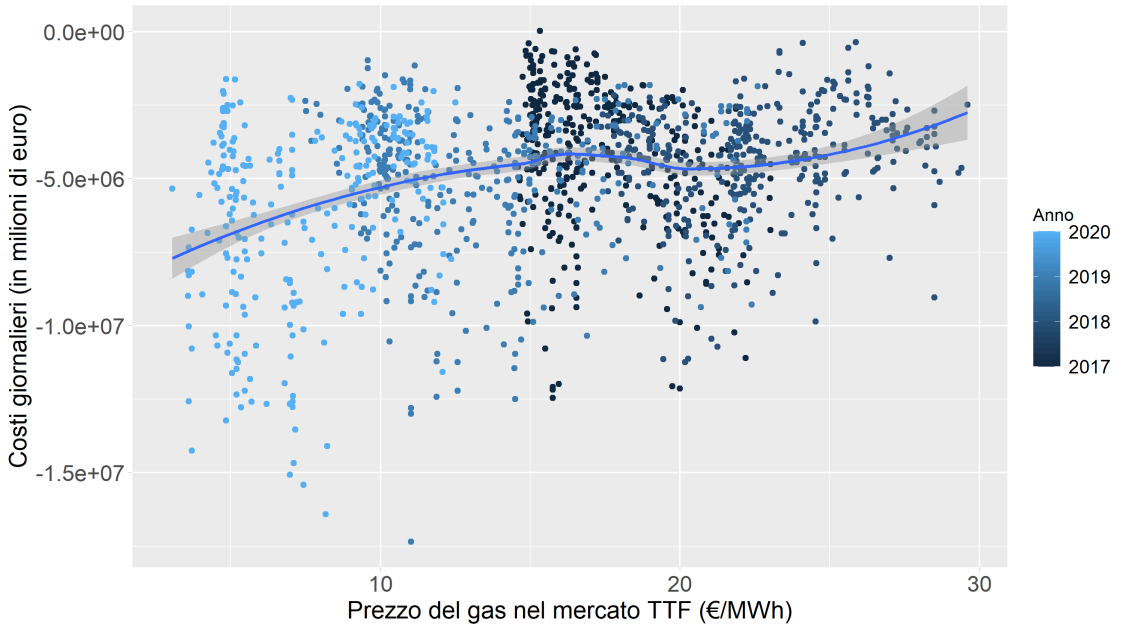


Figura 2.14: Relazione tra il prezzo del gas sul TTF e i costi di approvvigionamento giornalieri, suddivisi per anno, fino a settembre 2020.

Il grafico in Figura 2.14 mostra la relazione tra i costi di approvvigionamento di riserva e il prezzo del gas, suddivisi per anno. Questa analisi, come tutte le analisi presentate finora, riguarda solo il periodo di stima, in cui, come si vede della serie storica di Figura 2.13, i prezzi più bassi si registrano proprio nel 2020. Infatti, nella Figura 2.14, le osservazioni a cui sono associati prezzi inferiori del gas, sono quelle in blu chiaro, corrispondenti al periodo più recente. La relazione complessiva che emerge è che a prezzi bassi del gas si hanno maggior costi di approvvigionamento di riserva. Questa situazione nel 2020 era associata ad un crollo dei consumi che ha posto MSD in una condizione di minore competitività e maggiore potere di mercato da parte di alcuni operatori.

Il prezzo delle quote nel sistema EU ETS

L'aumento del prezzo del gas naturale osservato nell'ultimo periodo, ha quindi determinato, come concausa tra le principali, un notevole aumento del costo dell'energia. Ma insieme a tali rialzi, si è verificato l'aumento anche dei prezzi di quei permessi per le emissioni di anidride carbonica scambiati nel sistema ETS (le cosiddette "quote") dell'Unione europea. È possibile osservare l'andamento negli anni del prezzo di questi permessi, con riferimento alla Figura 2.12, in cui è rappresentata la serie storica della variabile ETS_t, fino al 30 settembre 2021 in Figura 2.12.

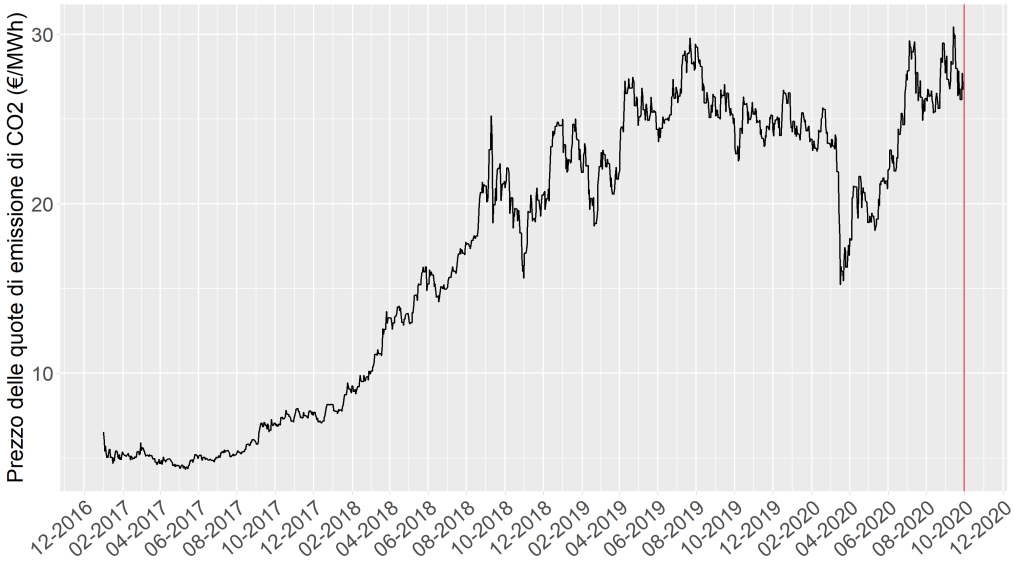


Figura 2.15: Serie storica del prezzo giornaliero delle quote di emissione di CO2 fino al 30 settembre 2020.

EU ETS è l'acronimo di *European Union Emissions Trading Scheme*, il sistema per lo scambio di quote emissione di gas a effetto serra dell'Unione Europea, riguardante settori industriali energivori, come operatori aerei, impianti termoelettrici industriali, manifatture e impianti di produzione, stoccaggio e trasporto di diverso tipo. Ad oggi, sono circa 1.200 gli impianti italiani coinvolti, di cui il 71% nel settore manifatturiero. Dal 1° gennaio 2005 gli impianti in Europa ad elevate emissioni di CO_2 non possono funzionare senza un'autorizzazione ad emettere gas serra. Ogni impianto autorizzato deve monitorare annualmente le proprie emissioni e compensarle con quote di emissione europee che possono essere comprate e vendute sul mercato.

Viene definito un sistema "*cap* & *trade*", perché fissa un tetto massimo (*cap*) al livello complessivo delle emissioni consentite a tutti i soggetti vincolati, ma permette ai partecipanti di acquistare e vendere sul mercato (*trade*) diritti a emettere CO_2 (quote) secondo le loro necessità, all'interno del limite stabilito.
In breve, l'EU ETS istituisce un mercato europeo per la compravendita di "quote di emissione" di CO_2: ne vengono assegnate alle aziende, ogni anno, in una certa quantità che si riduce via via nel tempo. Le aziende più inquinanti dovranno quindi acquistare altri permessi se vorranno continuare a emettere CO_2 senza incorrere in sanzioni; le aziende più "pulite", al contrario, hanno la possibilità di vendere le proprie quote inutilizzate. L'intero sistema mira a rendere sconveniente l'utilizzo di energia prodotta da fonti fossili (carbone, petrolio, gas naturale), incentivando il passaggio a forme di energia più pulite (come quelle rinnovabili).

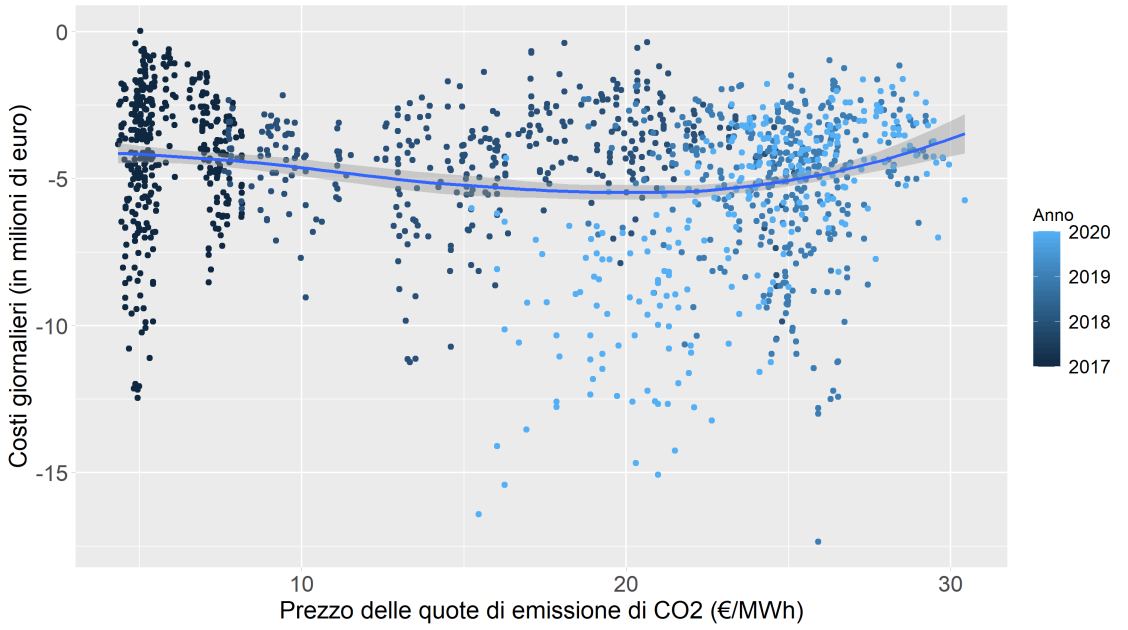


Figura 2.16: Relazione tra il prezzo delle quote di emissione di CO2 e i costi di approvvigionamento giornalieri, suddivisi per anno, fino a settembre 2020.

In Figura 2.16, è mostrata la relazione tra costi giornalieri di approvvigionamento della riserva, la nostra variabile di interesse, e il prezzo delle emissioni di CO_2, e si nota che al crescere del prezzo della CO_2, i costi prima aumentano leggermente e poi diminuiscono. I costi maggiori di approvvigionamento, relativi ad anni più recenti (in colore blu chiaro) si trovano in corrispondenza di prezzi più elevati. Negli ultimi anni infatti si è assistito ad

un aumento sia dei costi di approvvigionamento che dei prezzi delle quote di emissione di CO_2, quest'ultimo causato soprattutto dall'aumento dei prezzi del gas. Se il gas è diventa costoso, diventa più conveniente usare l'energia delle centrali a carbone, che però hanno bisogno di pagare più quote per compensare le loro emissioni, causando quindi l'aumento dei prezzi.

In conclusione, le analisi descrittive delle relazioni marginali mostrano come il fenomeno d'interesse, i costi di approvvigionamento di riserva di energia, sia complesso e influenzano in modo non lineare da molte variabili. In particolare, alcune variabili hanno un effetto periodico, altre sono fortemente non lineari, alcune hanno un effetto sempre non lineare ma meno marcato. Per riuscire a tenere conto della grande varietà di comportamenti riscontrati si sceglie la modellazione non parametrica, più flessibile e adatta a queste situazioni.

3

Strumenti statistici

Nel presente capitolo l'obiettivo è quello di offrire una panoramica degli strumenti statistici impiegati nella modellazione del rischio associato a costi di approvvigionamento di riserva acquistati da Terna in MSD per la messa in sicurezza del sistema elettrico nazionale. Per gestire la varietà di relazioni tra le variabili e il fenomeno di interesse, individuate nel capitolo 2, la modellazione più adeguata risulta essere quella non parametrica.

Iniziamo considerando la serie dei costi di approvvigionamento, indicata con c_t, e una serie di k variabili esplicative, possibili regressori per il rischio associato ai costi, $\mathbf{x_t} = (x_{1,t}, \ldots, x_{k,t})'$ per $t = 1, \ldots, n$. La complessa relazione tra la variabile dipendente c_t e le esplicative $\mathbf{x_t}$ è modellata tramite due approcci.

Nel primo approccio, il modello generale a cui si fa riferimento è

$$c_t = f(\mathbf{x_t}) + \varepsilon_t$$

dove $f(\mathbf{x_t}) = \mathbb{E}[c_t|\mathbf{x_t}]$ è la media condizionata dei costi c_t al valore assunto dalle variabili $\mathbf{x_t} = \{x_{1,t}, \ldots, x_{k,t}\}'$, e ε_t la parte residuale, che può assumere diverse caratteristiche, a seconda delle quali sono applicate diverse procedure. La media condizionata viene modellata in modo additivo non parametrico con un modello additivo generalizzato (GAM), scelto per la sua flessibilità.

I residui di tale modellazione

$$\varepsilon_t = c_t - \hat{c}_t = c_t - f(\mathbf{x_t}) = c_t - \mathbb{E}[c_t|\mathbf{x_t}]$$

rappresentano la deviazione dall'andamento medio dei costi, ovvero ciò che si discosta da quanto atteso. Per Terna questo rappresenta un rischio, più o meno grande in base all'entità della deviazione. I modelli di misura del rischio vengono applicati sui residui, al fine di osservarne il comportamento sulle code. È qui che si collocano deviazioni ampie dalla media, in positivo, sulla coda destra, in negativo, su quella sinistra, corrispondenti a osservazioni che, dopo essere state modellate dalla media sono ancora poco spiegate.

Nella modellazione dei residui sono considerate diverse procedure:

- a definire i modelli di CaR "marginali", tra cui il modello basato sul metodo del nucleo, con l'assunzione di residui omoschedastici, ovvero con varianza costante nel tempo: $\varepsilon_t \sim D(0, \sigma^2)$;
- a definire i modelli di CaR "condizionati", senza assumere omoschedasticità dei residui e capaci di cogliere diverse dinamiche nel tempo. Tra questi sono impiegati:

  1. modelli della classe GARCH, che definiscono una struttura per la varianza condizionata dei residui: $\varepsilon_t \sim D(0, \sigma_t^2(\mathbf{z_t}))$, dove $\mathbf{z_t} \subseteq \mathbf{x_t}$ indica l'insieme di covariate da cui dipende la varianza dei residui, che può non coincidere con l'insieme di tutte le variabili esplicative $\mathbf{x_t}$;
  2. modelli di regressione quantilica, per i quali il quantile della distribuzione dei residui q_t^ε può variare nel tempo. Saranno considerati due modelli di regressione quantilica, di cui uno è il CAViaR (Engle and Manganelli, 2004).

Da notare che non è stata specificata la distribuzione D dei residui: solitamente vengono definiti gaussiani, ma qui non è specificata poiché verranno utilizzati (anche) metodi non parametrici, grazie ai quali sarà possibile non specificare alcuna forma a priori.

Il secondo approccio opera anch'esso senza alcuna assunzione sulla dinamica temporale, rivolgendosi alla modellazione in un'unica fase del quantile condizionato. Con questo approccio, senza modellare in prima battuta la media condizionata dei costi e poi il

quantile dei residui, si lavora direttamente sul quantile condizionato dei costi. Il modello impiegato in questo caso è il Quantile-GAM (QGAM) (Fasiolo et al., 2021b), basato su regressione quantilica non parametrica con logica bayesiana.

In questo capitolo verranno descritti dal punto di vista teorico i modelli utilizzati. Si inizia con il modello additivo generalizzato, noto come *Generalized Additive Model* (GAM), sez. 3.1, impiegato per la modellazione della media condizionata nell'approccio in due fasi; per la seconda fase, relativa alla modellazione dei residui, sono utilizzati modelli di misura del rischio CaR (*Cost-at-Risk*), sez. 3.3.
Lavorando sulla distribuzione marginale dei residui, in sez. 3.4, con metodi non parametrici, si utilizza la modellazione basata sul metodo del nucleo, detta anche distribuzione *kernel*, par. 3.4.1.
Lavorando sulla distribuzione condizionata dei residui, in sez. 3.5, si utilizzano modelli della classe GARCH, par. 3.5.1 e modelli di regressione quantilica (QR), par. 3.5.2. Viene presentata anche un particolare specificazione del modello di regressione quantilica, il *Conditional Autoregressive VaR* (CAViaR) (Engle and Manganelli, 2004), par. 3.5.3.
Infine si presenta il modello *Quantile-GAM* (QGAM) (Fasiolo et al., 2021b) per il calcolo del rischio, relativo all'approccio di modellazione diretta del quantile condizionato, sez. 3.6.1.

In conclusione al capitolo, in sez. 3.7, sono analizzati dal punto di vista teorico anche i test di validazione del modello di misura del rischio, utilizzati in seguito: il test di Kupiec (1995), il test di Christoffersen (1998) e il test del quantile dinamico (2004).

3.1 Modelli additivi generalizzati

Nel capitolo 2 sono state descritte le variabili a disposizione: i costi di approvvigionamento di riserva, la variabile di interesse, che nel modello vengono indicati con $\mathbf{y}$, di natura quantitativa. Le altre variabili, potenziali predittori della variabile risposta, divengono nei modelli, mostrati in seguito, le covariate o variabili indipendenti, anch'esse tutte quantitative. Per questa tipologia di variabili i modelli a disposizione sono molti, alcuni più affidabili di altri, considerando l'elevata dimensionalità del problema.

Infatti, per modellare la complessità del fenomeno ideale è l'utilizzo di modelli non parametrici, che riescano ad adattarsi e a rappresentare relazioni non lineari come quelle rilevate nel nostro caso. Tuttavia, con molti dati, questi modelli mostrano debolezze strutturali, dovute al problema della cosiddetta "maledizione della dimensionalità", alla quale non sono soggetti modelli parametrici.

Quest'espressione, coniata da Richard Bellman (1961), e ripresa da Hastie (2009), nell'ambito dell'analisi di dati con elevata dimensionalità, fa riferimento ad un limite nella gestione di $k > 2$ dimensioni. Potrebbe sembrare logico che all'aumentare del numero p dei predittori utilizzati per stimare un modello aumenti la qualità della stima, tuttavia in generale a migliorare la stima di un modello, riducendo l'errore nell'insieme di stima, è l'aggiunta di covariate che sono effettivamente associate con la variabile risposta. Altrimenti, includere queste variabili, che costituiscono quindi rumore, non fa altro che aumentare la dimensionalità e il rischio di sovradattamento ai dati[52].
Quello che succede è che con l'aumento del numero di osservazioni, aumenta la loro sparsità nello spazio $\mathbf{R}^p$, come mostrato in Figura 3.1. Per compensare alla dispersione dei dati occorrerebbe un numero di punti dell'ordine di n^p, ma con la conseguente impossibilità di stimare accuratamente la funzione f che definisce tale superficie in $\mathbf{R}^p$,ovvero passante per gli intorni degli n^p punti.

La soluzione a questa problematica viene dall'utilizzo di tecniche per la riduzione della dimensionalità, o dall'imposizione di strutture, arrivando a individuare un compromesso

[52] Alle variabili non associate con la risposta, classificabili come rumore, vengono comunque associati coefficienti non nulli per non escludere la possibilità che emergano correlazioni con l'insieme di verifica

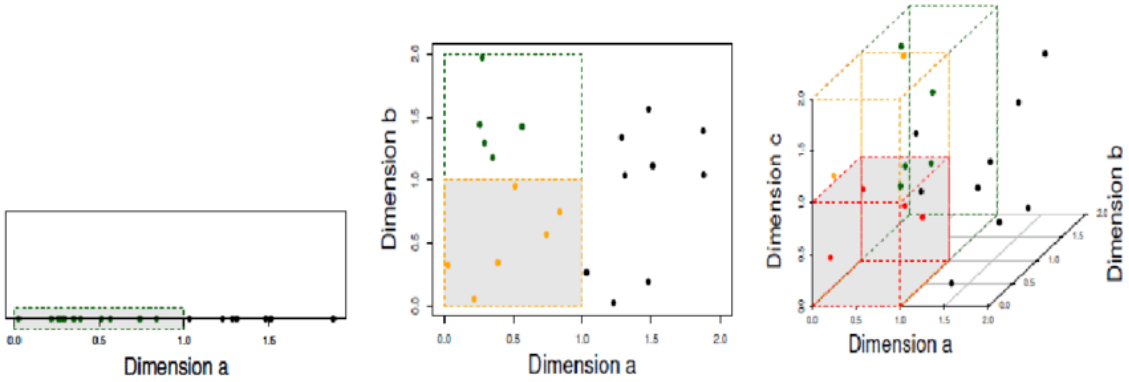


Figura 3.1: La maledizione della dimensionalità (a) 11 osservazioni in un segmento unitario (1 dimensione) (b) 6 osservazioni in un quadrato di lato unitario (spazio a 2 dimensioni) (c) 4 osservazioni in un cubo di lato unitario (3 dimensioni). Fonte[53]

tra modelli non parametrici, flessibili ma deboli con elevata dimensionalità, e strutture forti ma rigide, come i modelli parametrici.

I modelli additivi generalizzati, *Generalized Additive Models* (GAMs) (Hastie and Tibshirani, 1986), furono sviluppati originariamente nel tentativo di unire le caratteristiche di modelli lineari generalizzati e dei modelli additivi, *Additive Models* (AM), conferendo una struttura alla forma della f, ma lasciando ampi margini di flessibilità.

Tali modelli forniscono un *framework* a supporto dell'estensione del modello lineare standard, rendendo possibile l'applicazione di funzioni non lineari a ognuna delle covariate, mantenendo nel mentre l'additività. Così come i modelli lineari, l'additività dei GAM può essere applicata sia con una variabile risposta qualitativa (problemi di classificazione) che quantitativa (problemi di regressione).
Di fatto, i GAMs sono modelli lineari generalizzati, in cui la variabile risposta dipende in modo lineare da funzioni liscianti potenzialmente non lineari dei predittori, e l'interesse si concentra sull'inferenza in queste funzioni. La possibilità di introdurre automaticamente funzioni non lineari delle covariate conferisce grande flessibilità al modello, oltre a rendere l'adattamento più accurato.

In generale, la relazione descritta tra variabile risposta e covariate può così essere rappresentata:

$$g(\mathbb{E}[\mathbf{c_t}|\mathbf{x_t}]) = \alpha_0 + f_1(x_{1,t}) + f_2(x_{2,t}) + \cdots + f_p(x_{p,t}) \qquad (3.1)$$

[53] Parsons et al. (2004)

in cui α_0 è la media marginale della variabile risposta c_t , p è il numero di variabili esplicative ed ε_t è il termine di errore a media nulla $\mathrm{E}[\varepsilon_t] = 0$ e varianza costante $\mathrm{V}(\varepsilon_t) = \sigma_t^2$.
Classici esempi di funzione *link* sono la funzione logaritmica, la funzione logit, la funzione probit e la funzione identità:

$$\mathbb{E}[\mathbf{c_t}|\mathbf{x_t}] = \mu(\mathbf{x_t}) = \alpha_0 + \sum_{i=1}^{p} f_i(x_{i,t}) \tag{3.2}$$

Sarà questa la specificazione utilizzata nelle analisi per i costi di approvvigionamento della riserva di energia, assumendo che la loro distribuzione sia generata da una variabile casuale y gaussiana, e in questo caso la funzione legame $g(\cdot)$ è la funzione identità.
Affinché non vi siano problemi di identificabilità, deve vale le seguente condizione, per cui le f_i sono centrate attorno allo zero:

$$\mathbb{E}[f_j(x_j)] = 0 \ \ \forall j = 1, \ldots, p$$

L'inferenza nei GAMs non di concentra su parametri, bensì riguarda le funzioni $f_1, \ldots, f_p$, centrate sullo zero, in modo da ottenere una stima dell'effetto parziale (additivo) di quella variabile a cui sono associate sulla risposta, al netto dell'effetto stimato per le altre. Questo rende i GAMs modelli altamente interpretabili, rendendo anche più facile analizzare la struttura dei fenomeni, grazie alla possibilità di osservare l'effetto di ogni covariata sulla risposta, mantenendo fisse le altre. Dal punto di vista computazionale le funzioni vengono stimate con due possibili procedure, cui sono associati due diversi pacchetti di implementazione nel software `R`:

- procedura basata sull'algoritmo di *backfitting* per la stima delle funzioni *smooth*, implementata dal pacchetto `gam` Hastie, Tibshirani, James and Witten (2021), equivalente all'originale `gam` nel software S-Plus (Hastie, Tibshirani and Buja, 2021);
- procedura basata sull'algoritmo P-IRLS di *Penalized Iteratively Reweighted Least Squares*, che è una modifica dell'algoritmo IRLS utilizzato per i GLM. Tale algoritmo di stima viene applicato con il pacchetto `mgcv` sviluppato da Wood (2015), e in particolare sarà questa la metodologia di stima sfruttata nelle analisi presentate in seguito.

3.1.1 Stima del modello GAM

Questa procedura di stima delle funzioni $f_1, \ldots, f_p$, implementata nel pacchetto `mgcv` del software `R` (Wood, 2015), è impiegata per la stima dei GAMs, ma includendo, con questi modelli anche qualsiasi GLM con penalità quadratica e una grande varietà di altri modelli stimati da un approccio di verosimiglianza penalizzata quadraticamente.
I GAMs vengono infatti trattati in `mgcv` come modelli lineari generalizzati penalizzati (GLMs): ogni termine di lisciamento è rappresentato attraverso un insieme di funzioni base e viene a ciascuno associata una penalità; il peso di ciascuna nella verosimiglianza penalizzata è determinato da un parametro di lisciamento. L'algoritmo di stima messo a punto da Wood si basa su un procedimento iterativo sviluppato originariamente per i GLMs, con il nome *Iteratively Re-weighted Least Squares* (IRLS), e che diventa in `mgcv` P - IRLS, *Penalized IRLS* (Wood, 2000). In questo procedimento, ad ogni iterazione, il problema dei minimi quadrati viene sostituito dai minimi quadrati penalizzati, nel quale la selezione dei parametri di lisciamento avviene tramite diversi criteri: la convalida incrociata generalizzata, *Generalized Cross-Validation* (GCV), UBRE (*Un-biased Risk Estimator*), GACV (*Generalized Approximate Cross-Validation*) o un'approssimazione di Laplace della massima verosimiglianza ristretta (*Restricted Maximum Likelihood*, REML). Per tutti i metodi, valori più piccoli indicano modelli migliori.

Ci sono evidenze che l'ultima possa essere effettivamente la scelta migliore, dimostrandosi più robusta ma computazionalmente più costosa. La sfida computazionale risolta da `mgcv` è quella di riuscire a ottimizzare la selezione del lisciamento in modo efficiente e affidabile.

Originariamente i GAM vengono stimati tramite *backfitting*. Con questo algoritmo tutte le funzioni $f_i(\cdot)$ sono stimate usando lisciatori, tuttavia risulta difficile selezionare opportunamente il loro parametro di lisciamento. L'approccio tramite P-IRLS si basa sulla rappresentazione delle funzioni $f_i(\cdot)$ in *basis expansion*, descritta in 3.2, di dimensione moderata, e sulla penalizzazione della verosimiglianza tramite penalità quadratica per contrastare il sovradattamento.

Il modello additivo generalizzato di eq.3.2 viene riscritto come

$$\mathbb{E}[\mathbf{c}|\mathbf{X}] = \boldsymbol{\mu}(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta} \tag{3.3}$$

Nell'eq.3.3 $\mathbf{X}$ è la matrice di disegno $n \times p$ con una colonna di 1 e le basi, mentre $\boldsymbol{\beta}$ contiene la costante e i coefficienti delle basi. Per una trattazione più estesa si rimanda a

Wood and Augustin (2002). Assumendo $p < n$, β può essere stimato minimizzando la devianza penalizzata

$$D(\lambda, \beta) = ||y - \mathbf{x}\beta||^2 + \sum_j \lambda_j \beta' S_j \beta \tag{3.4}$$

con $\mathbf{S_j}$ matrice dei coefficienti noti, tale che $\beta'\mathbf{S_j}\beta$ sia la componente di penalizzazione composta da elementi dell'espansione in basi delle *splines*. Il parametro λ_j controlla il compromesso tra adattamento e lisciamento (*variance-bias trade off*), e una volta selezionato determina β_λ. La stima di λ è un processo delicato: se λ è troppo grande i dati risulterebbero troppo lisciati, mentre se λ è troppo piccolo il modello rischia di sovradattarsi ai dati. L'obiettivo è scegliere λ in modo che $\hat{f}_i(\cdot)$ sia il più possibile vicina ad $f_i(\cdot)$.

L'algoritmo P-IRLS implementa diversi criteri per la scelta di λ, enunciati in precedenza; il più semplice è la convalida incrociata generalizzata (GCV). La scelta ricade sul valore di λ che minimizza, rispetto a λ:

$$V_{gcv}(\lambda) = \frac{nD}{(n - DoF)^2} = \frac{n||y - \mathbf{x}\hat{\beta}_\lambda||^2}{(n - tr(F_\lambda)^2}$$

con D, la devianza, pari a $||y - \mathbf{x}\hat{\beta}_\lambda||^2$, n il numero di osservazioni, DoF, i gradi di libertà effettivi del modello pari a $tr(F_\lambda)$

Nel caso del criterio UBRE, la quantità da minimizzare è

$$V_{ubre}(\lambda) = \frac{D}{n} + \frac{2sDoF}{n - s}$$

con s è il parametro di scala. Da notare che di fatto UBRE è AIC riscalato, e può essere impiegato solo quando s è noto.

Vediamo quindi più nel dettaglio come avviene la stima delle funzioni che caratterizzano un GAM nell'algoritmo P-IRLS.

Sia μ_i la media definita in eq.3.3 e $V(\mu_i)$ la varianza di c_i, indichiamo con l'apice $[k]$ la stima alla k-esima iterazione, arrivando a definire $\beta^{[k]}$;

1. si ottiene la matrice diagonale dei pesi $\mathbf{W}$, dove

$$\mathbf{W_{ii}} = \frac{1}{V_i g'(\mu_i^{[k]})^2}$$

con $g(\cdot)$ funzione legame identità, e il vettore

$$z^{[k]} = \mathbf{X}^{[k]}\beta^{[k]} + \Gamma^{[k]}(y - \mu^{[k]})$$

dove $\Gamma^{[k]}$ è una matrice diagonale tale che $\Gamma^{[k]}_{ii} = g'(\mu^{[k]}_i)$;

2. si determina il valore di λ_i che minimizza il criterio GCV in questo caso, analogamente definito in precedenza:

$$\frac{||\sqrt{\mathbf{W}}(z - \mathbf{X}\beta^{new})||^2}{|tr(I - A)|^2}$$

con $A = \mathbf{x}(\mathbf{x}'\mathbf{W}\mathbf{x} + \sum_j \lambda_j \beta_j' \beta)^1 \mathbf{x}'\mathbf{W}$.

3. si calcola quindi la soluzione del seguente problema di minimizzazione,

$$\beta^{new} \leftarrow arg\ min||\sqrt{\mathbf{W}}(z - \mathbf{X}\beta)||^2 + \sum_j \lambda_j \beta' S_j \beta$$

finché la procedura non raggiunge la convergenza.

Nelle elaborazioni presentate in seguito, il modello additivo generalizzato viene stimato con questo algoritmo, specificando per la modellazione delle funzioni $f_i(\cdot)$ *splines* cubiche di regressione, *splines* di regressive adattive e *thin plate splines*.

Il parametro di lisciamento scelto per l'avvio dell'algoritmo di stima corrisponde alla dimensione della base usata per rappresentare il termine di lisciamento, ed è indicato con k. La sua scelta deve essere calibrata con attenzione, in modo che sia comunque più grande del valore che si ritiene necessario per l'approssimazione della funzione. Per approfondimenti, Wood (2015).

A definire la metodologia di lisciamento vi sono: la base utilizzata per rappresentare la funzione di lisciamento e la penalità usata per i coefficienti della base, in modo da controllare i gradi di libertà. Le *thin plate regression splines* forniscono in genere il più basso valore di errore quadratico medio, anche se computazionalmente onerose soprattutto per grandi dataset. Le *thin plate splines* permettono di specificare un numero qualsiasi di covariate, di selezionare l'ordine di penalità e non presentano nodi. Questa tipologia di *spline* è stata scelta nella modello GAM, per cui un approfondimento a riguardo è dato nel paragrafo 3.2.

Il metodo che solitamente viene dopo le *thin plate splines*, in termini di errore quadratico medio, è quello basato sulle *splines* di regressione cubiche, *cubic regression spline*. A differenza delle *splines* precedenti, questo metodo permette di lisciare solo una covariata, ma è meno oneroso e i parametri sono direttamente interpretabili, vantaggi indubbi quando si trattano dataset di grandi dimensioni. Anche questa tipologia di *spline* è stata impiegata all'interno del modello GAM scelto per la media condizionata, per cui se ne dà una spiegazione teorica in seguito, nel paragrafo 3.2.

3.2 Splines

I modelli GAM utilizzano le *splines*, di lisciamento e di regressione all'interno delle due diverse procedure di stima, rispettivamente il metodo basato sull'algoritmo di *backfitting* e quello di massimizzazione della verosimiglianza penalizzata con P-IRLS. È quindi al fine di offrire una maggiore completezza espositiva che verranno presentate le due tipologie di *splines* e il loro impianto teorico.

Il termine inglese *spline* ha un corrispettivo molto concreto nel mondo reale: indica un sottile pezzo di legno o di metallo che, inserito all'interno di ingranaggi, collega parti diverse di un macchinario per farle funzionare. Il ruolo della *spline* è lo stesso anche in matematica: è infatti una funzione, costituita da un insieme di polinomi raccordati tra loro, il cui scopo è interpolare un insieme di punti (nodi, *knots*) in un intervallo, in modo tale che la funzione sia continua almeno fino ad un dato ordine di derivate in ogni punto dell'intervallo. Di fatto si tratta di una funzione polinomiale a tratti, utile per approssimare funzioni (globalmente) di cui è noto solo il valore in alcuni punti.

Più precisamente, una funzione f è una *spline* di ordine d se scelti k valori sull'asse delle ascisse, detti nodi, $\xi_1, \xi_2, \ldots, \xi_K$, la funzione f passa esattamente per questi punti (ξ_j, c_j), ed è libera negli altri punti, quindi $f(\xi_j) = c_j$ con $j = 2, \ldots, K-1$; se in ognuno degli intervalli (ξ_j, ξ_{j+1}) con $j = 2, \ldots, K-1$ la funzione è approssimabile con un polinomio di grado $d = K - 1$ (K vincoli); e infine, se nei nodi la funzione è "liscia" ($K - 2$ vincoli, detti di continuità), ossia: continua, tale che $f(\xi_j^-) = f(\xi_j^+)$, con derivata prima continua, tale che $f'(\xi_j^-) = f'(\xi_j^+)$, con derivata seconda continua, tale che $f''(\xi_j^-) = f''(\xi_j^+)$. In tutto ciò $j = 2, \ldots, K-1$ e dove $f(\xi^-)$ e $f(\xi^+)$ indicano il limite da sinistra e da destra rispettivamente, della funzione $f(\cdot)$ nel punto ξ.

Queste caratteristiche rendono la *spline* una giustapposizione di polinomi, con vincoli di continuità nei punti di discontinuità. Ora, affinché tutti i parametri siano stimati in modo univoco occorre che il numero di parametri e di vincoli imposti coincida.

In un polinomio di grado d si hanno $d + 1$ parametri (intercetta e coefficienti dei monomi). La *spline* prevede l'utilizzo di $K - 1$ polinomi, passanti per i K nodi, per un totale di $(d + 1)(K - 1)$ parametri.

Per poter identificare i parametri in maniera univoca, il numero di parametri deve essere eguagliato dal numero di vincoli. Questi sono imposti, dalle condizioni definite in precedenza, pari a $K + 3(K - 2)$: K è il numero di vincoli relativo al numero di nodi che la

funzione interpola, $3(K-2)$ è il numero di vincoli sulle derivate e sul valore della funzione in ciascun nodo (3 vincoli) per tutti i punti "interni", cioè esclusi gli estremi, per cui valgono i vincoli di continuità.
Complessivamente, $(d+1)(K-1) - K - 3(K-2)$ risulta in $K(d-3) - d + 5$ gradi di libertà, ovvero parametri che, non governati da alcun vincolo, restano "liberi" e che, a seconda di come vengono fissati, definiscono tipologie differenti di *splines*.

Solitamente il grado del polinomio scelto è $d = 3$, per cui risultano 2 gradi di libertà. I vincoli quindi imposti sono vincoli sulla derivata seconda nei nodi estremi $f''(\xi_1) = f''(\xi_K) = 0$: nei due intervalli estremi le curve sono delle rette, che suppongono un andamento lineare della funzione oltre i confini. Tale *spline* prende il nome di *spline cubica naturale*.

Le funzioni definite a tratti possono risultare di scomoda gestione: per questo esiste una formulazione per cui una funzione *spline* può essere definita su tutto $\mathbb{R}$ attraverso l'espansione in basi, *basis expansion*, come combinazione lineare di opportune funzioni di base, a costituire una base di funzioni.
Una *spline* di ordine d con nodi ξ_j con $j = 1\ldots, K$ è un polinomio definito a tratti di ordine d, con derivate continue fino all'ordine $d-2$ ed è definita come:

$$f(x) = \sum_{j=1}^{K+4} h_j(x)\beta_j$$

con β_j parametri da stimare e con $h_j(\cdot)$ funzioni base definite da $\mathbb{R}$ in $\mathbb{R}$, $h_j(x) = x^{j-1}$ per $j = 1\ldots, d$, d polinomi per l'origine $(0,0)$ e $h_{d+l}(x) = (x-\xi_l)_+^{d-1}$ per $l = 1, ..., K$, K funzioni a soglia, o *threshold*. $(\cdot)_+$ è l'operatore che indica solo la parte positiva, per cui:

$$(x-\xi_1)_+ = \begin{cases} (x-\xi_1) & x > \xi_1 \\ 0 & \textit{altrimenti} \end{cases}$$

Con una scrittura più estesa:

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \ldots \beta_d x^d + \beta_{d+1}(x-\xi_1)_+^d + \beta_{d+2}(x-\xi_2)_+^{d+1} + \cdots + \beta_{d+K}(x-\xi_K)_+^{d+K}$$

Tale rappresentazione delle *spline* sotto forma di combinazione lineare di funzioni base, che la rende una funzione lineare nei parametri β_j, torna particolarmente comodo ad esempio nella risoluzione del problema centrale delle *splines* di regressione, presentate in seguito.

3.2.1 Splines di regressione

Con l'obiettivo di stimare la funzione f nel modello

$$c_i = f(x_i, \beta) + \varepsilon_i \text{ con } i = 1, \ldots, n$$

utilizzando per la sua approssimazione, una funzione *spline*, vengono introdotte le *splines* di regressione.

Il metodo delle *splines* di regressione consiste nel trovare la migliore approssimazione *spline* tramite la seguente regressione ai minimi quadrati:

$$arg\ \min_\beta \left\{ c_i - \sum_{i=1}^{n} h_j(x)\beta_j \right\}^2$$

e il risultato del problema di ottimizzazione risulta $\hat{c}_i = \sum_{i=1}^{n} h_j(x)\hat{\beta}_j$. Si noti che la rappresentazione come combinazione lineare semplifica l'applicazione dei minimi quadrati. Per implementare la regressione basata sulle *splines* occorre scegliere il grado del polinomio p, nonché il numero e la posizione dei nodi ξ_j.
La scelta del grado del polinomio ricade solitamente su $p = 3$; il vero elemento di criticità rimane la scelta dei nodi, nel numero ($K << n$) e nel posizionamento. A questo scopo, vengono impiegati metodi di stima-verifica, convalida incrociata o criteri di informazione come quello di Akaike (AIC). La scelta di K regola la complessità del modello: più K aumenta, più la funzione diventa variabile. La posizione dei nodi ne determina la forma. Una volta fissato K, l'approccio più semplice consiste nell'imporli equispaziati.

Le *splines* cubiche di regressione, definite dal comando `s(·,bf="cr")` in `gam`, sono delle *splines* di regressione, definite per K nodi. Se non diversamente specificati, i nodi vengono posizionati attraverso la griglia di tutti i valori assunti dalla covariata a cui si riferisce il termine, in modo che gli intervalli abbiano ampiezza uniforme. La *spline* è una funzione polinomiale cubica a tratti, corrispondenti agli intervalli tra nodi consecutivi, continua nei punti e fino alla derivata seconda. Con l'ultima aggiunta del vincolo sui nodi estremi che la derivata seconda assuma valore zero, la curva risultante è una *spline* cubica naturale passante per tutti i nodi, che, a differenza di una cubica non vincolata, estrapola linearmente oltre i nodi estremi, d'altronde come già detto.

Tensor product of splines

Il prodotto tensoriale di *splines*, in inglese *tensor product of splines* rappresenta una generalizzazione multidimensionale delle *splines* di regressione. Si consideri per semplicità il caso bivariato con $p = 2$. Il prodotto tensoriale si ottiene disponendo di una base di funzioni in $\mathbb{R}^p = \mathbb{R}^2$ e moltiplicando insieme le basi di funzioni unidimensionali relative a ciascuna variabile esplicativa $x - 1, ..., x_p$.
Con $x = (x_1, x_2) \in \mathbb{R}^2$, le basi di funzioni relative alla prima esplicativa sono $h_{1k}(x_1)$ con $k = 1, ..., k_1$ per un totale di k_1 basi, con $k_1 - 2$ nodi, per la seconda variabile sono $h_{2k}(x_2)$ con $k = 1, ..., k_2$ per un totale di k_2 basi, con $k_2 - 2$ nodi.
La base prodotto tensoriale di dimensione $k_1 \times k_2$ è $g_{jk} = h_{1j}(x)h_{2k}(x)$: non è lineare in quanto prodotto tensoriale di una coppia di basi lineari a tratti (o fino al nodo e retta oltre) e viene usata per rappresentare $g(x) = \sum_{j=1}^{k_1} \sum_{k=1}^{k_2} \theta_{jk} g_{jk}(x)$.

3.2.2 Splines di lisciamento

Le *splines* di lisciamento si differenziano da quelle di regressione innanzitutto per il fatto che non richiedono sia fatta una scelta sul numero e il posizionamento dei nodi. Se per le *splines* di regressione i nodi sono spesso scelti dal software in modo uniforme sul campo di variazione della variabile da lisciare, sulla base dei gradi di libertà stabiliti dall'utente, per le *splines* di lisciamento selezionano i nodi in modo indiretto, imponendo un vincolo di variazione più o meno lenta (*smoothness*) ad una generica funzione interpolante, con l'obiettivo di minimizzare $D(f, \lambda)$, risolvendo un problema di minimizzazione dei quadrati penalizzati.

$$D(f, \lambda) = \sum_{i=1}^{n} (c_i - f(x_i))^2 + \lambda \int_{-\infty}^{+\infty} [f''(t)]^2 \, dt$$

Il primo elemento è la funzione di perdita quadratica, che misura la varianza dei dati, e il secondo elemento è la penalizzazione con derivata seconda, che misura il grado di irregolarità della curva: più è grande, più è irregolare la curva.
Tra questi due elementi troviamo λ, il parametro di penalizzazione: se λ è piccolo viene esaltata la funzione di perdita, quindi si preferisce che la funzione passi vicino ai punti, ma il rischio è il possibile sovradattamento. Per $\lambda \to 0$, converge ad una *spline* interpolativa. Se invece λ è grande, viene esaltata la penalizzazione, quindi preferisco che la funzione sia

liscia; il rischio in questo caso è che la funzione si adatti troppo poco. Per $\lambda \to +\infty$ converge alla stima ai minimi quadrati.

La soluzione al problema di minimizzazione è la *spline* cubica naturale con nodi su tutti i singoli valori delle x, per un totale di n_0, combinazione lineare di tutte le funzioni di base $N(\cdot)$

$$\hat{f(x)} = \sum_{j=1}^{n_0} N_j(x)\theta_j = N\theta$$

dove N è la matrice la cui j-esima colonna contiene i valori di N_j in corrispondenza degli n_0 valori distinti di x. $\hat{f(x)} = \hat{y} = S_\lambda y$ è un lisciatore lineare.
λ definisce poi il numero effettivo dei nodi, ed è scelto con metodo *Leave-One-Out Cross-Validation*.

Thin plate splines

Le *thin plate splines* rappresentano una generalizzazione alla caso multidimensionale delle *splines* di lisciamento: esse utilizzano, come sostituto della derivata seconda della funzione $f(\cdot)$, il laplaciano. La funzione da penalizzare nel problema di minimizzazione di $D(f, \lambda)$ diventa:

$$\int\int_{\mathbb{R}^2} \left\{ \left(\frac{\partial^2}{\partial x_1^2} f(x) \right)^2 + 2 \left(\frac{\partial^2}{\partial x_1 \partial x_2} f(x) \right)^2 + \left(\frac{\partial^2}{\partial x_2^2} f(x) \right)^2 \right\} dx_1 dx_2$$

La soluzione al problema di ottimo:

$$f(x) = \hat{\beta}_0 + \hat{\beta}' x + \sum_{j=i}^{n} \alpha_j h_j(x)$$

con $h_j(x) = \eta(||x - x_j||)$ e $\eta(z) = z^2 log(z)$.

Le *thin plate splines* implementate nel modello `gam` con `mgcv`, sono definite lisciatori isotropici a basso rango. Con il termine "isotropico" si indica il fatto che la rotazione del sistema di coordinate delle covariate non cambia in esito al lisciamento. Con la terminologia "a basso rango" si vuole indicare il fatto che hanno molti meno coefficienti dei dati da lisciare: il lisciatore non viene adattato ad ogni punto e se 10 nodi bastano per approssimare 100 punti, è scelto questo.

3.3 Modelli di Costo-a-Rischio

Il cuore delle analisi condotte in questa tesi risiede nel tentativo di misurare il rischio che Terna debba sostenere costi per l'approvvigionamento di riserva di energia molto più alti di quelli medi, precisamente, cercando di capire come si comportano quei costi attorno al quantile 10%. Ci si riferisce alla coda sinistra della distribuzione della variabile risposta, dove si trovano valori dei costi in modulo molto grandi, preceduti dal segno meno.

I modelli impiegati a questo scopo sono definiti modelli di calcolo del CaR, cioè del Costo-a-Rischio, *Cost-at-Risk*, che riprendono, nel nome e nella struttura, la metodologia del VaR, Valore a Rischio, *Value-at-Risk*. La sua semplicità concettuale rende il VaR facilmente estendibile ad un ambito differente di applicazione. Il VaR infatti è solitamente applicato a portafogli finanziari, mentre con il CaR l'oggetto della valutazione è il rischio associato non a un portafoglio di titoli, bensì ai costi, sostenuti da Terna.

Una prima intuitiva definizione del CaR, è la seguente: dato un insieme di variabili $\mathbf{x_t}$, a definire possibili fattori influenti sui costi, il CaR è il massimo costo potenziale, espresso in euro, nel quale l'azienda può incorrere in un determinato orizzonte temporale h, con un determinato livello di confidenza $1-\alpha$, in condizioni normali di mercato. È pertanto una misura di rischio monetario basata su considerazioni probabilistiche.
Una seconda definizione del CaR pone l'accento proprio sulla sua natura probabilistica, definendo il CaR come il quantile α-esimo della distribuzione dei costi e guadagni ad h periodi, in condizioni normali di mercato. Lo scopo finale è quello di calcolare:

$$CaR_{t,h,\alpha} = q^c_{t,h,\alpha}(\mathbf{x_t}) \quad tale\ \ che \quad Pr(c_t \leq q^c_{t,h,\alpha}) = \alpha$$

dove $q^c_{t,h,\alpha}(\mathbf{x_t})$ indica il quantile α-esimo della distribuzione dei costi/guadagni c_t ad h giorni, funzione di $\mathbf{x_t}$, l'insieme delle variabili che concorrono al rischio.

Nel calcolo del CaR si fa riferimento quindi al concetto di "quantile" della distribuzione. Il quantile di livello α di una variabile casuale Y è definito (Pace and Salvan, 2001) come quel valore q_α tale che

$$Pr(Y \leq q_\alpha) = \alpha \qquad e \qquad Pr(Y \geq q_\alpha) = 1-\alpha$$

Nella definizione di quantile entra la definizione di funzione di ripartizione di Y, $F_Y(q_\alpha) = Pr(Y \leq q_\alpha)$. Quindi il quantile q_α è ricavabile dalla funzione di ripartizione di Y tramite il

suo inverso in α. Il quantile di livello α è infatti definito come il minimo valore y assunto dalla variabile Y tale per cui la funzione di ripartizione in y supera α.

$$q_\alpha = F_Y^{-1}(\alpha) = inf\Big\{y|F_Y(y) \geq \alpha\Big\} = Q_Y(\alpha)$$

L'inverso della funzione di ripartizione $F_Y^{-1}()$ è chiamata con $Q_Y()$ ed è definita funzione quantile, definita come $Q : (0,1) \to \mathbb{R}$.

Nelle analisi presentate in seguito si fa riferimento al quantile di livello $\alpha = 0.10$. Questo valore lascia quindi alla sua destra il 90% della distribuzione dei dati. Stiamo considerando infatti il quantile della coda sinistra, indicante il rischio che si debbano sostenere costi più elevati, in valore assoluto, poiché i costi sono una variabile a valori negativi. La coda destra rappresenta invece costi "positivi", ovvero guadagni, che per Terna non costituiscono un vero e proprio rischio da quantificare.

Il CaR al 10% con dati giornalieri (h=1), ad esempio, è il quantile della distribuzione dei costi/guadagni che lascia alla sua sinistra un'area pari al 10% del totale. In altri termini, è il massimo costo potenziale verificabile nell'orizzonte temporale di 1 giorno con livello di confidenza del 90%. Un $CaR_{h,\alpha}$=$CaR_{1,0.10}$ pari a 1 milione di euro, indica che entro la prossima giornata, vi è la possibilità di avere un costo fino a 1 milione di euro con una probabilità del 10%.

Il CaR dipende quindi dai seguenti elementi:

- il livello di confidenza α. Dopo esserci confrontati con Terna, è stato ritenuto opportuno scegliere un livello α pari al 10%. α è la probabilità che definisce quanto il costo è estremo rispetto alla distribuzione dei costi e guadagni, e dipende dall'"attitudine al rischio" dell'azienda, o meglio, da quanto l'azienda vuole esporsi a livello di costi. Il CaR aumenta al diminuire di α: se α diminuisce, passando dal 15% al 10% ad esempio, sto considerando la situazione in cui si presentano costi ancora più elevati, per calcolare qual è il livello di risorse che l'azienda deve predisporre per fa fronte a costi più sostenuti; si può dire che è un atteggiamento più conservativo, che tiene conto di rischi maggiori. Tuttavia nell'ambito dei costi per un'azienda è poco indicativo ridurre ancora α perché in questo modo si considererebbero situazioni molto estreme, per cui si avrebbe una sopravvalutazione del rischio reale.
- l'orizzonte temporale h: è stato scelto un orizzonte giornaliero ($h = 1$) e mensile ($h = 30$) per calcolare il rischio a 1 giorno e a 30 giorni. h rappresenta il periodo su cui si

misura il costo potenziale. Il CaR aumenta all'aumentare di h: più è ampio il periodo di riferimento, più è probabile che si verifichino eventi rischiosi;

- il metodo di stima della distribuzione dei costi, da cui dipende la stima del quantile α-esimo.

I modelli di CaR vengono applicati nell'ambito di un approccio in due fasi, lavorando sui residui del modello per la media condizionata $\mathbb{E}[c_t|\mathbf{x_t}]$.

Per la modellazione della distribuzione dei residui $\varepsilon_t = c_t - \mathbb{E}[c_t|\mathbf{x_t}]$, di cui oggetto di interesse è il quantile α-esimo, vengono adottate diversi modelli di CaR.
Come già anticipato, vi sono diverse possibilità: i modelli di calcolo del CaR sono basati sia sulla distribuzione marginale, come mostrato nel par. 3.4, che su quella condizionata, nel par. 3.5; nel primo caso si assume omoschedasticità, nel secondo si lascia cadere tale assunto e si modella la dinamica nel tempo della varianza condizionata, con un modello GARCH, o del quantile condizionato, con la regressione quantilica. Vediamo quindi nello specifico quali sono i modelli scelti e la teoria sottostante.

3.4 Modelli di CaR basati su distribuzione marginale

La modellazione del CaR basata su distribuzione marginale è la più semplice da applicare. Basandosi sulla distribuzione marginale del processo, assume l'omoschedasticità del processo generatore dei dati, cioè che il processo abbia varianza costante nel tempo, quindi $\varepsilon_t \sim D(0, \sigma^2)$. Le diverse applicazioni di CaR marginale si basano su una differente assunzione sulla distribuzione dei residui, e sulla scelta dell'approccio modellistico. In particolare si può optare per un approccio parametrico, per il quale la distribuzione vera dei costi è approssimabile ad una distribuzione parametrica, come la gaussiana, o la t-Student, oppure un approccio non parametrico, come il metodo del nucleo.

Il più semplice modello di CaR marginale è il CaR marginale gaussiano, modello parametrico che si basa sull'assunzione che la distribuzione dei residui sia normale. Indicando con $\mu(\mathbf{x_t}) = \mathbb{E}[c_t|\mathbf{x_t}]$ la media condizionata, in questo caso il CaR diventa:

$$CaR_{t,h,\alpha} = \mu(\mathbf{x_t}) + \Phi^{-1}(\alpha)\sigma$$

dove $\Phi(\cdot)$ è la funzione di ripartizione di una Normale standard.

Scegliendo una distribuzione diversa dalla normale, si definiscono i modelli di CaR non gaussiano. Questi modelli sono utili nel caso in cui i dati mostrassero caratteristiche distanti dalla normalità, come code pesanti, o leptocurtiche. Un esempio di CaR marginale non gaussiano è il CaR che assume distribuzione t-Student per i residui, per il quale:

$$CaR_{t,h,\alpha} = \mu(\mathbf{x_t}) + \mathcal{P}_{\nu}^{-1}(\alpha)\sigma$$

dove $\mathcal{P}(\cdot)$ è la funzione di ripartizione di una t-Student con ν gradi di libertà t_ν.

Altri modelli marginali sono i modelli non parametrici di simulazione storica, tra cui quello basato sulla statistica d'ordine e sulla distribuzione *kernel*.

I modelli di CaR basati su simulazione storica sono così detti perché utilizzano i dati storici osservati per costruire la distribuzione empirica dei costi e guadagni. Il CaR storico, al livello α, ad un orizzonte h, è semplicemente il quantile α–esimo della distribuzione

empirica dei costi/guadagni con orizzonte h. Questo approccio non richiede alcuna assunzione sulla forma analitica della distribuzione, infatti sono modelli non parametrici, e ciò consente di poterli sfruttare in situazioni particolarmente complesse.

Per quanto riguarda la stima mediante statistiche d'ordine, si considera la serie dei costi c_t, la cui funzione di ripartizione è indicata con

$$\hat{F}_C(u) = n^{-1} \sum_{i=1}^{n} 1\{c_{t,i} \leq u\}$$

in cui u rappresenta il valore soglia. Da tale definizione è possibile quindi passare a quella di funzione quantilica o funzione quantile,

$$q_\alpha = \mathcal{F}_C{}^{-1}(u) = inf\Big\{u | F_C(y) \geq \alpha\Big\}$$

che a livello campionario è definita con

$$\hat{q}_\alpha = \hat{\mathcal{F}}_C{}^{-1}(u) = inf\Big\{u | \tfrac{\sum 1(c_{t,i} \leq u)}{n} \geq \alpha\Big\}$$

Dunque per ottenere $\hat{q}_\alpha$ sulla base di n osservazioni bisogna considerare la statistica ordinata, di ordine $k = n\alpha$. Una volta ordinati i dati $c_{[1]}, c_{[2]}, \ldots, c_{[n]}$, dove $c_{[i]}$ rappresenta la i-esima statistica ordinata, la statistica $c_{[n\alpha]}$, ossia in posizione $n\alpha$, non è altro che il quantile di livello α cercato. Nel caso in cui il quantile $k = n\alpha$ non sia un valore intero, presi $k1 < k < k2$, il quantile di livello α può essere stimato tramite interpolazione:

$$\hat{q_\alpha} = \tfrac{k_2 - k}{k_2 - k_1} c_{t,[k_1]} + \tfrac{k - k_1}{k_2 - k_1} c_{t,[k_2]}$$

Il CaR in questo caso corrisponde a

$$CaR_{t,h,\alpha} = \mu(\mathbf{x_t}) + \hat{q_{h,\alpha}}$$

Il CaR marginale stimato non parametricamente con il metodo del nucleo, o distribuzione *kernel* è presentato assieme alla descrizione teorica di questo modello, contenuta nel par. 3.4.1.

3.4.1 CaR basato su metodo del nucleo

Il CaR a h giorni di livello α, basato su metodo del nucleo, o distribuzione *kernel*, modella la distribuzione marginale dei residui, sotto l'assunzione di omoschedasticità, in modo non parametrico.

Sia ε_t la sequenza di variabili indipendenti e identicamente distribuite dei residui del modello per la media condizionata: $\varepsilon_t \sim D(0, \sigma^2)$, con varianza costante nel tempo.
L'espressione del $CaR_{t,h,\alpha}$ è:

$$\begin{aligned} CaR_{t,h,\alpha} &= \mu(\mathbf{x_t}) + q^{\varepsilon}_{h,\alpha} \\ &= \mu + \sum_{i=1}^{p} f_i(x_{i,t}) + q^{\varepsilon}_{h,\alpha} \end{aligned} \tag{3.5}$$

con q^{ε}_{α} quantile α-esimo della distribuzione dei residui ε_t, calcolata non parametricamente con il metodo del nucleo.

La stima della densità dei residui $\hat{f}_b(\varepsilon_t)$, $t = 1, \ldots, n$ tramite metodo del nucleo è ottenuta scegliendo come funzione del nucleo gaussiano $K(u)$, tale che $K(u) \geq 0$ e $\int K(u)du = 1$, descritta in 3.10. Con b ampiezza di banda > 0, si ha:

$$\hat{f}_b(\varepsilon_t) = \frac{1}{n} \sum_{i=1}^{n} K\left(\frac{\varepsilon_t - \varepsilon_i}{b}\right)$$

Sia $\hat{F}_b(\varepsilon_t)$ la funzione di ripartizione relativa a $\hat{f}_b(\varepsilon_t)$. Il CaR a 1 giorno di livello α è pari a

$$CaR_{h,\alpha} = q^{\varepsilon}_{h,\alpha} = \hat{F}_b^{\,-1}(\alpha) \tag{3.6}$$

con $h = 1$ e $q^{\varepsilon}_{h,\alpha}$ quantile α-esimo della distribuzione dei residui ε_t, calcolata non parametricamente con il metodo del nucleo.

Il modello di CaR basato sulla distribuzione *kernel*, elaborato sui residui del GAM, è indicato in seguito come GAM-K.

Il metodo del nucleo

Per il metodo del nucleo, basato sulla funzione nucleo, o *kernel*, si considera il campione di osservazioni $c_1, \ldots, c_n$ dalla variabile di interesse c_t, la serie dei costi di approvvigionamento. Immaginiamo di disporre tutte le osservazioni su una retta reale e selezioniamo un'origine c_0. Si suddivide la retta reale in intervalli (*bins*) B_j di ampiezza b (*binwidth*): $B_j = [c_0 + (j-1)b, c_0 + jb)$, con $j \in \mathbb{Z}$. Per un generico c_t la stima della densità è data da

$$\hat{f_b}(c_t) = \frac{1}{nb} \sum_{i=1}^{n} \sum_{j} 1(c_i \in B_j) 1(c_t \in B_j) \tag{3.7}$$

In pratica, indicato con m_j il centro di ogni intervallo B_j, $\hat{f_b}(c)$ conta quante osservazioni (n_j) cadono nell'intervallo $[m_j - b/2, mj + b/2)$, quindi ad ogni $c \in B_j$ assegna la stessa frequenza di m_j, $f_j = n_j/n$.

La stima della densità $\hat{f_b}(c_t)$ dipende sia dalla scelta di b che dalla scelta di c_0. Se $c_0 = 0$ e si vuole stimare la densità di un punto $c_t \in B_j$, 3.7 può essere scritta più semplicemente come

$$\hat{f_b}(c) = \tfrac{1}{nb} \sum_{i=1}^{n} 1(c_i \in B_j)$$

Si consideri senza perdita di generalità, intervalli di ampiezza $2b$, cioè intervalli del tipo $[c_t - b, c_t + b)$. La stima di $f(c_t)$ si può scrivere come

$$\hat{f_b}(c_t) = \frac{1}{2nb} \sum_{i=1}^{n} 1\Big(c_i \in [c_t - b, c_t + b)\Big) \tag{3.8}$$

Si consideri ora la funzione K, chiamata funzione del nucleo uniforme, espressa in funzione di u, con $u = (c_t - c_i)/b$:

$$K(u) = \tfrac{1}{2} 1(|u| \leq 1)$$

Utilizzando la funzione $K(u)$, la stima della densità in 3.8 diventa

$$\hat{f_b}(c_t) = \frac{1}{nb} \sum_{i=1}^{n} K\Big(\frac{c_t - c_i}{b}\Big) = \frac{1}{nb} \sum_{i=1}^{n} \frac{1}{2} 1\Big(|\frac{c_t - c_i}{b})| \leq 1\Big) \tag{3.9}$$

Per ogni osservazione che cade nell'intervallo $[c_t - b, c_t + b)$ la funzione indicatrice assume valore 1 e dà un contributo al calcolo della frequenza. Questo contributo, con la funzione nucleo uniforme, non dipende da quanto c_i è vicino a c_t, ma solo dal fatto che sia nell'intervallo definito da b. Se si vuole che le osservazioni più vicine ad c_t diano un contributo maggiore si possono utilizzare altre funzioni nucleo, come il nucleo gaussiano o il nucleo di Epanechnikov. La funzione nucleo di Epanechnikov corrisponde a:

$$K(u) = \frac{3}{4}(1 - u^2)\mathbb{I}(|u| \leq 1)$$

mentre il nucleo gaussiano corrisponde a:

$$K(u) = \frac{1}{2\pi} exp\Big(- \frac{1}{2}u^2\Big) \tag{3.10}$$

Nelle analisi condotte in seguito sarà quest'ultima funzione quella utilizzata.

Utilizzando la funzione nucleo si perviene alla forma generale dello stimatore della densità con il metodo del nucleo, basato su un campione casuale $c_1, \ldots, c_n$. Tale stimatore assume la forma

$$\hat{f_b}(c_t) = \frac{1}{n}\sum_{i=1}^{n} K_b(c_t - c_i) \tag{3.11}$$

con $K(\cdot) = \frac{1}{b}K(\frac{\cdot}{b})$, con K funzione nucleo. Per avere l'intera densità bisogna calcolare tale funzione su una griglia di c_t. Il termine "funzione nucleo" (*kernel function*) si riferisce alla funzione ponderante $K(\cdot)$; l'espressione "stimatore della densità basato sul nucleo" (*kernel density estimator*) si riferisce alla formula 3.11.

3.5 Modelli di CaR basati su distribuzione condizionata

La modellazione CaR basata su distribuzione condizionata fa riferimento alla distribuzione condizionata del processo, consentendo il superamento dell'assunzione di omoschedasticità del processo generatore dei dati.

Infatti è contemplata l'eteroschedasticità dei residui, che possono avere una varianza che cambia nel tempo, funzione di un vettore di variabili $\mathbf{z_t}$. Come detto all'inizio del capitolo, si indica con $\mathbf{z_t} \subseteq \mathbf{x_t}$, un insieme di variabili che potrebbe coincidere con tutti i regressori del modello o rappresentarne una parte, da cui la varianza condizionata dei residui dipendono. Per questa classe di modelli quindi $\varepsilon_t \sim D(0, \sigma_t^2(\mathbf{z_t}))$. I modelli della classe GARCH definiscono una struttura per la varianza condizionata, risultando adatti a questo scopo.

Accanto alla modellazione della varianza condizionata, risulta sensato considerare attraverso opportuni modelli la dinamica temporale del quantile dei residui, che può, appunto, variare nel tempo. Allo scopo di valutare il suo andamento è stata considerata la regressione quantilica, in due proposte: una nostra specificazione, e quella nota come CAViaR, di Engle e Manganelli (2004).

I paragrafi successivi mostrano i diversi modelli di CaR accanto alla teoria sottostante i vari modelli impiegati: il par. 3.5.1 è dedicato al CaR basato sui modelli della classe GARCH, il par. 3.5.2 al CaR basato sulla regressione quantilica, e il par. 3.5.3 alla specificazione CAViaR.

3.5.1 CaR basato su modelli GARCH

Il CaR a h giorni di livello α, basato su modelli della classe GARCH, lavora sulla distribuzione condizionata dei residui, definendo mediante una struttura parametrica la dinamica della varianza condizionata.

Sia ε_t la sequenza di variabili a definire i residui del modello sulla media condizionata, per i quali i modelli CaR basati su distribuzione condizionata, assumono $\varepsilon_t \sim D(0, \sigma_t^2(\mathbf{z_t}))$. La varianza è funzione del tempo e di variabili $\mathbf{z_t}$ che variano nel tempo. Si cerca di modellare quindi residui eteroschedastici, e il modello GARCH lo fa definendo:

$$\varepsilon_t = \nu_t \sigma_t$$

$$con \quad \nu_t \sim IID(0,1) \qquad \varepsilon_t | \mathcal{I}_{t-1} \sim D(0, \sigma_t^2(\mathbf{z_t}))$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$$

$$con \quad \varepsilon_t = c_t$$

Da notare che abbiamo scelto di rappresentare un modello GARCH a media nulla perché avendo già modellato la media con il modello per la media condizionata la consideriamo azzerata per i residui.
L'espressione del $CaR_{t,h,\alpha}$ diventa:

$$\begin{aligned} CaR_{t,h,\alpha} &= \mu(\mathbf{x_t}) + q_{t,\alpha}^{\varepsilon} \\ &= \mu + \sum_{i=1}^{p} f_i(x_{i,t}) + q_{\alpha}^{\nu} \sigma_t \end{aligned} \tag{3.12}$$

con $q_{t,h,\alpha}^{\varepsilon}$ quantile α-esimo della distribuzione dei residui ε_t, indicizzato da t in quanto variabile nel tempo e il quantile q_{α}^{ν} che può essere stimato sia parametricamente che non. Se si sceglie di ipotizzare che le innovazioni siano approssimativamente normali, q_{α}^{ν} diventa z_{α}, quantile di livello α di una normale standard. Mentre σ_t è la varianza condizionata, funzione di sè stessa a riatrdo $t-1$ e dei residui ritardati.

Il modello di CaR basato su modelli della classe GARCH, elaborato sui residui del GAM, è indicato in seguito come GAM-GARCH.

Modelli della classe GARCH

I modelli della classe GARCH (*Generalized Autoregressive Conditional Heteroskedasticity* furono formulati da Tim Bollerslev (1986), come argomento della sua tesi di dottorato, sotto la supervisione di Robert Engle. Nascono come generalizzazione dei modelli ARCH, introdotti proprio da Engle (1982). Lo sviluppo di "metodi di analisi delle serie storiche economiche con volatilità variabile nel tempo" valse a Engle, insieme a Granger, il Premio Nobel per l'economia nel 2003.

Sia il modello ARCH che la sua generalizzazione permettono di modellare la varianza mediante un'opportuna struttura parametrica, per tenere conto della possibile eteroschedasticità, presente in dati con varianza non costante nel tempo. Quando l'eteroschedasticità è correlata serialmente, ovvero condizionata a periodi di elevata o modesta varianza, definiamo la serie storica "ad eteroschedasticità condizionata". Il GARCH permette di tenere conto di una dinamica più complessa della varianza condizionata nel tempo, a differenza del più semplice processo autoregressivo modellabile con un ARCH.

La formulazione del modello GARCH(p,q) è costituita da due equazioni, una per la media condizionata (3.13) e una per la varianza condizionata (3.14):

$$c_t = \mu + \varepsilon_t = \mu + \nu_t \sigma_t \tag{3.13}$$

$$con \quad \nu_t \sim IID(0,1) \qquad \varepsilon_t | \mathcal{I}_{t-1} \sim IID(0, \sigma_t^2)$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \tag{3.14}$$

$$con \quad \varepsilon_t = c_t - \mu$$

dove p è l'ordine dei termini ARCH $\varepsilon_t^2 = (c_t - \mu)^2$, q è l'ordine dei termini GARCH σ_t^2 e le innovazioni ν_t sono assunte indipendenti e identicamente distribuite con distribuzione da specificare, sulla base delle caratteristiche dei dati.

In ogni istante t, la varianza dipende dai valori passati attraverso ε_{t-i}^2, per $i = 1 \ldots, p$ e dalla varianza condizionata σ_{t-j}^2, per $j = 1, \ldots, q$. In particolare, nella varianza condizionata in equazione 3.14 si possono distinguere il livello base della varianza, ω, il contribu-

to della varianza dell'innovazione, $\sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2$ e il comportamento medio di lungo periodo, $\sum_{j=1}^{q} \beta_j \sigma_{t-j}^2$.

Un modello ARCH, a differenza del GARCH, non possiede i termini σ_t^2: è un modello appropriato quando la varianza dell'errore segue un processo autoregressivo (AR). Se il suo andamento segue un processo autoregressivo a media mobile (ARMA), il modello ARCH viene generalizzato dal GARCH grazie all'introduzione dei q termini σ_{t-j}^2, con $j = 1, \ldots, q$. Pertanto imponendo $q = 0$ nella formulazione del GARCH(p,q) si ottiene esattamente un modello ARCH(p).

I primi due momenti di c_t, condizionato e incondizionato sono, dato $\mathcal{I}_{t-1} = c_1, c_2, \ldots, c_{t-1}$ l'insieme informativo fino al tempo $t-1$:

$$\mathbb{E}[c_t|\mathcal{I}_{t-1}] = \mathbb{E}[\mu + \nu_t \sigma_t] = \mu$$

$$\mathbb{V}[c_t|\mathcal{I}_{t-1}] = \mathbb{E}[\mu + \nu_t \sigma_t] = \sigma_t^2$$

Mentre i momenti marginali si calcolano con il "metodo del valore atteso iterato":

$$\mathbb{E}[c_t] = \mathbb{E}[\mathbb{E}[c_t|\mathcal{I}_{t-1}]] = \mathbb{E}[\mu + \nu_t \sigma_t] = \mu$$

$$\mathbb{V}[c_t] = \mathbb{V}[\mathbb{E}(c_t|\mathcal{I}_{t-1})] + \mathbb{E}[\mathbb{V}(c_t|\mathcal{I}_{t-1})] = 0 + \mathbb{E}(\sigma_t^2)$$

e in relazione a quest'ultima si può dimostrare che

$$\mathbb{V}[c_t] = \frac{\omega}{1 - (\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2)}$$

Condizioni necessarie sui parametri affinché venga rispettata la positività del secondo membro dell'equazione 3.14 sono: $\omega > 0$, $\alpha_i \geq 0$, per $i = 1 \ldots, p$, e $\beta_j \geq 0$, per $j = 1, \ldots, q$. In aggiunta, la condizione $(\sum_{i=1}^{p} \alpha_i + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2) < 1$ definisce la stazionarietà del processo GARCH(p,q).

Modelli GARCH non gaussiani

La letteratura scientifica ha visto lo sviluppo di una grande varetà di modelli della classe GARCH, costruiti sulla base del modello di Bollerslev (1986) per rispondere ad una modellazione *ad hoc* di determinate caratteristiche dei dati.

Il GARCH con la formulazione classica riesce a modellare una serie c_t di osservazioni incorrelate e la correlazione delle osservazioni stesse al quadrato, c_t^2. Modella l'eteroschedasticità e gli effetti di *clustering*, dati dalla maggiore probabilità di permanere nello stesso stato di volatilità piuttosto che cambiarlo, per cui variazioni grandi tendono ad essere seguite da variazioni altrettanto grandi.

Ma la modellazione GARCH si è rivelata molto più flessibile, in seguito all'introduzione di modelli in grado di controllare situazioni anche estranee a quelle appena elencate. Ad esempio, alla modellazione di possibili correlazioni delle osservazioni, ci pensa ARMA-GARCH, agli effetti stagionali, SGARCH, alla persistenza nell'ACF dei quadrati, IGARCH o FIGARCH, per la correlazione incrociata tra c_t e c_t^2, il GARCH IN MEDIA o ancora il TGARCH per la modellazione di effetti asimmetrici.
Nella formulazione del GARCH in 3.14, le innovazioni sono definite indipendenti e identicamente distribuite, ma la distribuzione in sè non è definita. Solitamente le innovazioni ν_t hanno distribuzione normale, e in tal caso la distribuzione specificata è la normale, opzione di default nel comando `garchFit` di `R`. Tuttavia, a volte le innovazioni non hanno un andamento gaussiano: questo è il caso dei modelli GARCH non gaussiani, basati su una specificazione alternativa della distribuzione delle innovazioni, rispetto alla classica normalità. In particolare, questi modelli permettono di tenere conto di una certa "non normalità" residua o ad esempio, della leptocurtosi. È questo il caso che si presenta con i dati a nostra disposizione, per i quali è più appropriata una distribuzione a code pesanti, come la t di Student. Alcune distribuzioni specificabili con il comando `garchFit` sono, appunto, la t di Student, sia simmetrica che non, o anche la *Generalized Error Distribution* (GED) e la quasi massima verosimiglianza (QMLE).

Nel caso della scelta della t di Student simmetrica, il modello GARCH diventa:

$$c_t = \varepsilon_t = \sqrt{\frac{\upsilon - 2}{\upsilon}} \nu_t \sigma_t \tag{3.15}$$

$$con \quad \varepsilon_t | \mathcal{I}_{t-1} \sim t_\upsilon$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{p} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{q} \beta_j \sigma_{t-j}^2 \qquad (3.16)$$
$$con \quad \varepsilon_t = c_t$$

Quello che cambia è l'equazione 3.15, che riporta la standardizzazione per i gradi di libertà υ della t di Student t_υ: si moltiplicano le innovazioni per la radice del reciproco della varianza di una t-Student, in modo che $\mathbb{V}\left(\sqrt{\frac{\upsilon-2}{\upsilon}}\nu_t\right) = 1$, e quindi che $\mathbb{V}\left(\sqrt{\frac{\upsilon-2}{\upsilon}}\nu_t\sigma_t|\mathcal{F}_{t-1}\right) = \sigma_t^2$.

Poiché nella maggior parte delle applicazioni del modello GARCH, la media marginale μ è prossima allo zero, il caso appena esposto tiene conto di ciò e senza perdita di generalità si rappresenta il caso con $\mu = 0$, scelta fatta anche nel modello presentato in seguito.

Stima del modello GARCH

I parametri, vincolati alle condizioni in precedenza definite, sono stimati con il metodo della massima verosimiglianza, se la distribuzione delle innovazioni ν_t è nota, o, se la vera distribuzione di ν_t è ignota e deve essere ipotizzata, con la quasi massima verosimiglianza, "quasi" in riferimento al fatto che il modello potrebbe non essere correttamente specificato. Se invece il modello scelto è corretto la quasi massima verosimiglianza si riconduce al caso precedente. In entrambi i casi il procedimento da seguire è lo stesso.
Indicando con $\theta = (\alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)'$ il vettore dei parametri del modello GARCH(p,q), la funzione di verosimiglianza $\mathcal{L}(\theta)$ di un campione $c_1, \ldots, c_n$ esprime quanto è verosimile (probabile) osservare ogni valore del parametro θ, alla luce delle osservazioni $c_1, \ldots, c_n$.

Assumendo che i dati $c_1, \ldots, c_n$ siano indipendenti:

$$\mathcal{L}(\theta) = \mathcal{L}(\theta|c_1, \ldots, c_n) = f(c_1, \ldots, c_n|\theta) = \prod_{i=1}^{n} f(c_i|\theta) \qquad (3.17)$$

Se i dati $c_1, \ldots, c_n$ sono invece tra loro dipendenti, si fattorizza $f(c_1, \ldots, c_n|\theta)$ nella densità congiunta delle prime p osservazioni e nel prodotto delle $n-p$ densità condizionate di ogni osservazione dalla c_{p+1} alla c_n rispetto alle precedenti.

$$\mathcal{L}(\theta) = \mathcal{L}(\theta|c_1, \ldots, c_n) = f(c_1, \ldots, c_p|\theta) \cdot \prod_{i=p+1}^{n} f(c_i|\mathcal{I}_{i-1}; \theta) \qquad (3.18)$$

Mentre il secondo fattore è definito da assunzioni, quindi sempre noto con i GARCH, il primo fattore è spesso ignoto, motivo per cui la verosimiglianza in questo caso è sempre approssimata. L'approssimazione può essere svolta in diversi modi: se n è sufficientemente grande, si può direttamente trascurare il primo addendo. La verosimiglianza in 3.18 diventa:

$$\mathcal{L}(\theta) \propto \prod_{i=p+1}^{n} f(c_i|\mathcal{I}_{i-1};\theta)$$

3.5.2 CaR basato sulla regressione quantilica

Il CaR a h giorni di livello α, basato sulla regressione quantilica lavora sulla distribuzione condizionata dei residui, definendo mediante una struttura parametrica la dinamica della quantile di livello α.

Sia ε_t la sequenza di variabili a definire i residui del modello sulla media condizionata, per i quali i modello CaR basti su distribuzione condizionata, assumono $\varepsilon_t \sim D(0, \sigma_t^2(\mathbf{z_t}))$. La varianza è funzione del tempo e di variabili $\mathbf{z_t}$ che variano nel tempo. L'obiettivo è quello di modellare quindi residui eteroschedastici, e il modello GARCH lo fa definendo:

$$\varepsilon_t = \beta_{\alpha,0} + \sum_{i=1} \beta_{\alpha,i} s_{i,t} + \nu_t$$

dove i regressori $s_{i,t}$ possono essere valori ritardati di ε_t o altre variabili esogene. Sotto l'assunzione che $q_\alpha^\nu(\mathbf{s_t}) = 0$, il valore del quantile condizionato di ε_t è:

$$q_{t,\alpha}^\varepsilon = \beta_{\alpha,0} + \sum_{i=1} \beta_{\alpha,i} s_{i,t}$$

Ciò corrisponde ad una regressione quantilica su ε_t e non richiede assunti sulla distribuzione di ν_t (Koenker, 2005).

L'espressione del $CaR_{t,b,\alpha}$ diventa:

$$\begin{aligned} CaR_{t,b,\alpha} &= \mu(\mathbf{x_t}) + q_\alpha^\varepsilon(\mathbf{s_t}) \\ &= \mu + \sum_{i=1}^{p} f_i(x_{i,t}) + \beta_{\alpha,0} + \sum_{i=1} \beta_{\alpha,i} s_{i,t} \end{aligned} \tag{3.19}$$

Il modello di CaR basato sul modello di regressione quantilica, elaborato sui residui del GAM, è indicato in seguito come GAM-QR.

Regressione quantilica (QR)

La regressione quantilica, in inglese *quantile regression* (QR), è un modello di analisi di regressione impiegato per modellare i quantili condizionati di qualsiasi livello della distribuzione. La regressione quantilica è una valida alternativa alla regressione lineare utilizzata quando i requisiti per l'applicazione dei minimi quadrati ordinari (OLS) non sono soddisfatti. Infatti, anche l'impostazione del problema segue la stessa costruzione logica, ma può essere applicata quando si hanno valori anomali, poiché è una tecnica può robusta, o quando gli errori non sono distribuiti normalmente, situazione in cui risulta anche più efficiente, ad esempio.

È un sistema particolarmente comodo quando l'oggetto di interesse della modellazione sono valori sulle code della distribuzione, come nel nostro caso, utilizzata anche per stimare delle bande di confidenza per la variabile dipendente senza assumere per essa una particolare distribuzione condizionata. La regressione quantilica permette comunque di stimare anche il quantile di livello $\alpha = 0.5$, ovvero la mediana condizionata. Questa in particolare è calcolata minimizzando la somma degli scarti in valore assoluto, laddove il metodo dei minimi quadrati stimava la media condizionata minimizzando la somma degli scarti al quadrato. Vediamo ora quanto detto sotto forma di formulazione matematica.

Per una variabile casuale Y, la media è quel valore μ che minimizza la somma dei quadrati degli scarti:

$$\mu = \min_{\mu \in \mathcal{R}} \mathbb{E}_Y [Y - \mu]^2$$

La mediana è invece quel valore che minimizza il valore atteso degli scarti in modulo dalla mediana.

$$med = \min_{med \in \mathcal{R}} \mathbb{E}_Y |Y - med|$$

Sappiamo che il quantile α-esimo di una distribuzione $\mathcal{F}$ si può scrivere come

$$q_\alpha = \mathcal{F}_Y^{-1}(\alpha) = inf\Big\{y | F(y) \geq \alpha\Big\}$$

e a livello campionario

$$\hat{q_\alpha} = \hat{\mathcal{F}}_Y^{-1}(\alpha) = inf\Big\{y | \frac{\sum 1(Y \leq y)}{n} \geq \alpha\Big\}$$

e per ottenerlo bisogna ordinare le osservazioni dalla più piccola alla più grande.
Furono Koenker e Bassett (1978) a suggerire un cambiamento nell'impostazione teorica del problema della regressione quantilica, proponendo di sostituire il concetto di ordinamento con quello di ottimizzazione pesata di una funzione di perdita. Così, per la definizione di quantile α-esimo viene proposta la seguente formulazione:

$$\hat{q}_\alpha = \min_{\xi_\alpha \in \mathcal{R}} \mathbb{E}[\rho_\alpha(y - \xi_\alpha)]$$

La funzione $\rho_\alpha(\cdot)$ è detta *check function* ed è definita come segue

$$\rho_\alpha(z) = z(\alpha - 1(z < 0)) = \begin{cases} z\alpha & z \geq 0 \\ z(1-\alpha) & z < 0 \end{cases} \tag{3.20}$$

Questa formulazione più complessa per esprimere i quantili risulta particolarmente utile per passare direttamente all'ambito regressivo.

Si consideri il modello di regressione classico

$$y = \mathbf{x}'\beta + \varepsilon$$

con $\mathbf{x} = (x_1, \ldots, x_p)'$, $\beta = (\beta_1, \ldots, \beta_p)'$, basato sull'assunzione che la media condizionata degli errori, a media nulla, sia zero: $\mathbb{E}(\varepsilon) = \mathbb{E}(\varepsilon|\mathbf{x}) = 0$.
I parametri β sono calcolati con lo stimatore ai minimi quadrati $\hat{\beta}$

$$\hat{\beta} = \min_{\beta \in \mathcal{R}^p} \mathbb{E}_{Y|\mathbf{x}}[y - \mathbf{x}'\beta]^2$$

il quale permette di definire la media condizionata $\mathbb{E}[Y|\mathbf{x} = x] = \mathbf{x}'\beta$.

Si consideri ora il modello di regressione "quasi-standard", formulato in maniera analoga a quello classico

$$y = \mathbf{x}'\beta_\alpha + \varepsilon$$

con $\alpha \in (0, 1)$ indicante il livello del quantile α-esimo: per ogni valore di α è definito un vettore di parametri β, da cui il nome "quasi-standard", perché β dipende dal valore di α. In questo modello l'assunzione è fatta non sulla media condizionata, ma sul quantile α-esimo condizionato degli errori: $q_\alpha(\varepsilon|\mathbf{x}) = 0$. Il quantile condizionato di y diventa quindi

$$q_\alpha(y|\mathbf{x}) = \mathbf{x}'\beta_\alpha \tag{3.21}$$

che, con $\mathbf{x}$ indicizzata dal tempo a rappresentare un insieme di variabili $\mathbf{x_t} = x_{1,t}, \ldots, x_{p,t}$ e $\beta_\alpha = \{\beta_{\alpha,0}, \beta_{\alpha,1}, \ldots, \beta_{\alpha,p}\}$ diventa

$$q_\alpha(y|\mathbf{x}) = \mathbf{x_t}'\beta_\alpha = \beta_{\alpha,0} + \sum_{i=1}^{p} \beta_{\alpha,t} x_{i,t} \tag{3.22}$$

dove β_α è quel valore che risolve il problema di ottimizzazione definito dall'approccio basato sulla *check function* di Koenker e Bassett

$$\hat{\beta_\alpha} = \min_{\beta_\alpha \in \mathcal{R}^p} \mathbb{E}[\rho_\alpha(y - \mathbf{x}'\beta_\alpha)]$$

Il quantile condizionato viene ricavato come

$$\begin{aligned}\hat{\beta_\alpha} &= arg \min_{\beta_\alpha \in \mathcal{R}^p} \left\{ \sum_{i \in i | c_i \geq x_i'\beta_\alpha|} \alpha |c_i + x_i'\beta_\alpha| + \sum_{i \in i | c_i < x_i'\beta_\alpha|} (1-\alpha)|c_i + x_i'\beta_\alpha| \right\} \\ &= arg \min_{\beta_\alpha \in \mathcal{R}^p} \sum_{i=1}^{n} \rho_\alpha(y - x'\beta_\alpha)\end{aligned}$$

Lo stimatore del quantile condizionato è dato da $x_i\hat{\beta_\alpha}$.

A differenza dell'approccio basato sull'ordinamento, questo approccio basato sull'ottimizzazione può essere esteso al contesto di un modello di regressione, e dunque posso scrivere il quantile in funzione di alcuni regressori, sia variabili esogene che endogene ritardate. Per ogni quantile α il coefficiente stimato indica, per unità di variazione della x a cui è associato, di quanto varia il α-esimo quantile della y a parità di tutte le rimanenti covariate (Koenker, 2004a).
Tuttavia, a differenza del caso dei minimi quadrati, l'equazione di stima non può essere risolta in maniera esplicita, perché la *check function* non è differenziabile nell'origine. La stima è ottenuta mediante l'algoritmo del simplesso modificato, più in generale, attraverso algoritmi di programmazione lineare (Buchinsky and Hahn, 1998).

Lo stimatore OLS (*Ordinary Least Squares*) $\hat{\beta}$ è il miglior stimatore lineare non distorto sotto gli assunti del Teorema di Gauss-Markov[54]. Ma lo stimatore $\hat{\beta_\alpha}$ è non lineare e con

[54] Teorema di Gauss Markov: in un modello lineare in cui i disturbi hanno valore atteso nullo, sono incorrelati e omoschedastici, gli stimatori lineari corretti più efficienti sono gli stimatori ottenuti con il metodo dei minimi quadrati.

errori non gaussiani può essere anche più efficiente. Per questo la regressione quantilica è un ottimo strumento per la gestione di situazioni in cui gli assunti del metodo OLS non sono validi. Inoltre, un ulteriore vantaggio della regressione quantilica è che se si osservano diverse stime per diversi quantili, si può comprendere come cambia l'influenza delle covariate sulla variabile dipendente, nei vari punti della distribuzione quantile condizionata.

A livello di implementazione in R, la regressione quantilica è costruita con il pacchetto `quantreg`, sviluppato da Koenker, lo stesso autore della riformulazione teorica di tale modellazione. La funzione minimizza la somma pesata degli scarti in valore assoluto, formulata come un problema di programmazione lineare. Vi sono differenti poi algoritmi disponibili, a seconda del numero di osservazioni e di altre caratteristiche: il default ad esempio è una versione modificata dell'algoritmo di Barrodale e Roberts, ed è il migliore metodo con osservazioni nell'ordine delle centinaia. Per numerosità più elevate, fino a 10000,il metodo migliore è il metodo *interior point* di Frisch-Newton, il metodo di Frisch–Newton *approach after preprocessing* per $n >> p$, con p piccolo, e diversi altri, per i quali si rimanda a (Koenker, 2004b).

3.5.3 CaR basato sul modello CAViaR

Il CaR a h giorni di livello α, basato sulla regressione quantilica con specificazione CAViaR, lavora sulla distribuzione condizionata dei residui, definendo mediante una struttura parametrica specifica la dinamica della quantile di livello α.

Essendo il CAViaR una specificazione della regressione quantilica, valgono le considerazioni fatte in precedenza. Sia per il quantile α-esimo di ε_t

$$q^{\varepsilon}_{t,\alpha} = \beta_{\alpha,0} + \beta_{\alpha,1} q^{\varepsilon}_{t-1,\alpha} + \beta_{\alpha,2} \varepsilon^{+}_{t-1} + \beta_{\alpha,3} \varepsilon^{-}_{t-1}$$

L'espressione del $CaR_{t,h,\alpha}$ diventa:

$$\begin{aligned} CaR_{t,h,\alpha} &= \mu(\mathbf{x_t}) + q^{\varepsilon}_{t,\alpha} \\ &= \mu + \sum_{i=1}^{p} f_i(x_{i,t}) + \beta_{\alpha,0} + \beta_{\alpha,1} q^{\varepsilon}_{t-1,\alpha} + \beta_{\alpha,2} \varepsilon^{+}_{t-1} + \beta_{\alpha,3} \varepsilon^{-}_{t-1} \end{aligned} \tag{3.23}$$

Il modello di CaR basato sul modello CAViaR, elaborato sui residui del GAM, è indicato in seguito come GAM-CAViaR.

Il paragrafo successivo è dedicato a un approfondimento della teoria sottostante tale specificazione, introdotta da Engle and Manganelli (2004).

Conditional Autoregressive VaR (CAViaR)

Il modello CAViaR (*Conditional Autoregressive VaR*) (Engle and Manganelli, 2004) è una particolare specificazione della regressione quantilica, elaborata nell'ambito dei modelli VaR.

Il modello CAViaR può essere derivato dal modello di regressione quantilica, così per come è definito in equazione 3.21, ponendo $x_{i,t} = (\varepsilon_{t-1}, q_{\alpha,t-1})$ e con $q_\alpha(y|\mathbf{x}) = q_{\alpha,t}$ per semplicità notazionale. Pertanto

$$q_{\alpha,t} = \beta_{\alpha,0} + \sum_{i=1}^{p} \beta_{\alpha,i} x_{i,t} = \beta_{\alpha,0} + \beta_{\alpha,1}\varepsilon_{t-1} + \beta_{\alpha,2} q_{\alpha,t-1} \tag{3.24}$$

In questa equazione, $q_{\alpha,t}$ e i coefficienti $\beta_{\alpha,0}, \beta_{\alpha,1}$, parametro del termine ε_{t-1}, $\beta_{\alpha,2}$, parametro associato al termine autoregressivo sono indicizzati da α: essi dipendono da α perché il modello CAViaR è di fatto una regressione quantilica.

In generale, il modello CAViaR può essere definito con q termini autoregressivi $q_{\alpha,t-j}$ $j = 1, \ldots, q$ e con r termini ε_{t-i}, $i = 1, \ldots, r$.

$$q_{\alpha,t} = \beta_{\alpha,0} + \sum_{j=1}^{q} \beta_{\alpha,j} q_{\alpha,t-j} + \sum_{i=1}^{r} \beta_{\alpha,i} \varepsilon_{\alpha,t-i}$$

La logica del CAViaR nasce nell'ambito della modellazione VaR, nel tentativo di modellare il comportamento autoregressivo dei rendimenti finanziari, che esibiscono effetti di *clustering*, già accennati nel par. 3.5.1. La presenza di *cluster* è una caratteristica che anche un'altra tipologia di modelli cerca di spiegare: i modelli della classe GARCH. Il modello CAViaR mostra di fatto una certa somiglianza nella strutturazione con il GARCH.

Per semplicità, si consideri un GARCH(1,1), cioè con $p = q = 1$. Sappiamo che il modello GARCH è definito da due equazioni, una per la media e una per la varianza condizionata, così come definite in 3.13 e 3.14. L'equazione della varianza condizionata di un GARCH(1,1) è:

$$\sigma_t^2 = \omega + \alpha_1 \varepsilon_{t-1}^2 + \beta_1 \sigma_{t-1}^2$$

ed è immediato notare l'analogia con l'equazione del modello CAViaR in 3.24.

Inoltre, analogamente a quanto si fa per il modello GARCH, nella sua declinazione in T-GARCH (*Threshold*) GARCH (detto anche GJR-GARCH)[55], in cui si introduce una soglia, anche per il CAViaR esiste una specificazione di questo tipo, il CAViaR con modellazione a soglia. Utilizzato per le nostre analisi, si presenta come

$$q_{\alpha,t} = \beta_{\alpha,0} + \beta_{\alpha,1}\varepsilon_{t-1}^{-} + \beta_{\alpha,2}\varepsilon_{t-1}^{+} + \beta_{\alpha,3}q_{\alpha,t-1}$$

dove $\varepsilon_{t-1}^{-} = min\{\varepsilon_{t-1}, 0\}$ e $\varepsilon_{t-1}^{+} = max\{\varepsilon_{t-1}, 0\}$.

[55] Glosten, Jagannathan and Runkle, 1993

3.6 Modello QGAM per il calcolo del rischio

Il modello QGAM non richiede una procedura in due passaggi, bensì lavora direttamente sul quantile condizionato dei costi, che coincide proprio con il CaR ricercato nelle procedure sopra citate. È una generalizzazione del GAM ma riferita al quantile condizionato, quindi si tratta di una regressione quantilica non parametrica applicata ai costi:

$$c_t = \beta_{\alpha,0} + \sum_{i=1}^{p} f_{\alpha,i}(x_{i,t}) + \varepsilon_t$$

Sotto l'assunzione che il quantile di livello α dei costi c_t, condizionato al valore assunto dalle variabili $\mathbf{x_t}$ sia zero $q^{\varepsilon}_{t,\alpha}(\mathbf{x_t}) = 0$, l'espressione del $CaR_{t,h,\alpha}$ coincide con il quantile $q^{c}_{t,\alpha}$ dei costi:

$$CaR_{t,h,\alpha} = q^{c}_{t,\alpha} = \beta_{\alpha,0} + \sum_{i=1}^{p} f_{\alpha,i}(x_{i,t}) \tag{3.25}$$

dove le funzioni $f_{\alpha,i}$ hanno lo stesso significato che avevano nel modello GAM applicato alla media condizionata con il primo approccio, ma ora dipendono, tramite i loro parametri, dal livello del quantile α.

Il CaR calcolato con questo approccio è indicato con Q-GAM. La differenza con il GAM-QR risiede nel fatto che con il Q-GAM si modella direttamente il quantile condizionato di c_t, mentre con GAM-QR la regressione quantilica è applicata ai residui del modello per la media.

Nei paragrafi successivi è esposta più approfonditamente la teoria su cui si basa il modello QGAM, in riferimento all'articolo di Fasiolo et. al (2021b), con cui è stato presentato.

3.6.1 Quantile GAM

Il modello QGAM (Fasiolo et al., 2021b) è una novità abbastanza recente relativa alla modellazione del quantile condizionato. Viene infatti presentato nel 2021 da Fasiolo e Wood, l'autore del già descritto pacchetto `mgcv`, oltre che della teoria sottostante. Come evocato dal nome, il QGAM, *Quantile* GAM, è una generalizzazione del modello GAM riferita al quantile condizionato, anziché alla media condizionata.

I modelli additivi generalizzati sono modelli di regressione non lineare piuttosto flessibili, stimati dalla metodologia implementata in `mgcv`, che sfrutta elementi della logica bayesiana. Mentre però questi metodi per i GAMs sono basati sulla modellazione parametrica della distribuzione della variabile risposta, il modello QGAM, costruito in `R` con il pacchetto `qgam`, non presuppone l'utilizzo di alcuna assunzione parametrica.
Il QGAM è basato su una versione lisciata della funzione di perdita *pinball* (Koenker and Bassett, 1978), anziché sulla funzione di verosimiglianza. La sua assenza impedisce la diretta applicazione della regola di Bayes nell'aggiornamento della distribuzione a priori dei coefficienti di regressione.

Per ovviare alle tale problematica, la formulazione teorica più idonea è stata elaborata nel secondo articolo di Fasiolo, "*Fast Calibrated Additive Quantile Regression*" (Fasiolo et al., 2021a). Il termine "veloce", *fast*, si riferisce al fatto che il metodo presentato con questo articolo, e implementato in `qgam`, è il più efficiente metodo di regressione quantilica additiva dal punto di vista computazionale, proprio grazie al *framework* teorico su cui poggia.

In particolare, esso mira a selezionare il *learning rate*, la velocità di apprendimento, parametro di regolazione della funzione di perdita, in modo da raggiungere una copertura prossima a quella nominale. Viene quindi fornita una descrizione del QGAM partendo proprio dalla funzione di perdita *pinball*.

La funzione di perdita *pinball*

La funzione di perdita *pinball*, detta anche perdita di quantile, *quantile loss*, è un indicatore utilizzato per valutare l'accuratezza di una previsione quantilica.

Sia α il quantile desiderato, y il valore reale e z la previsione quantilica. Allora ρ_α^{pin}, la funzione di perdita *pinball*, potrà essere scritta come segue:

$$\rho_\alpha^{pin}(y) = \begin{cases} (y-z)\alpha & y \geq z \\ (z-y)(1-\alpha) & y < z \end{cases} \tag{3.26}$$

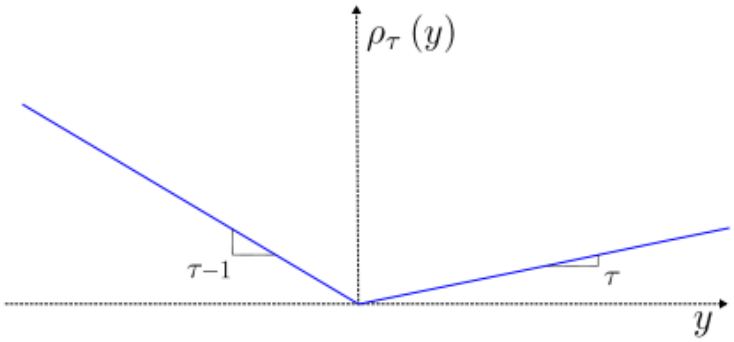


Figura 3.2: Funzione di perdita *pinball*. Fonte[56]

La funzione di perdita *pinball* rappresentata in Figura 3.2, dove τ coincide con α, è sempre positiva, è lineare a tratti, e in $(0, 0)$ ha derivate non continue. È chiamata così per via della sua forma, che ricorda la traiettoria della pallina di un flipper (*pinball* in inglese). Più ci allontaniamo da y, maggiore sarà il valore di $\rho_\alpha^{pin}(y)$. L'inclinazione è usata per riflettere lo squilibrio desiderato nella previsione quantilica. In particolare, se $y \geq z$, faccio pesare la differenza in quell'istante α, perché dovrebbe verificarsi $\alpha * 100$ volte. Se invece la previsione è maggiore del valore osservato de quantile, la differenza viene pesata per $1-\alpha$. Non dice se il quantile è stato stimato correttamente, bensì penalizza la differenza tra il valore reale e la previsione quantilica.

I migliori modelli quantilici hanno una bassa perdita *pinball*. Per confrontare l'accuratezza di due diversi modelli di previsione è sufficiente calcolare la perdita *pinball* me-

[56] Biau and Patra (2010)

dia in ogni modello, su un insieme di serie temporali abbastanza grande (nell'ordine delle centinaia) da assicurarci che la differenza osservata sia statisticamente rilevante.

Vediamo ora come la *pinball loss* entra nella modellazione del quantile condizionato con logica bayesiana, presentandone la strutturazione teorica.

Formulazione teorica del QGAM

Sia Y una variabile casuale continua con distribuzione condizionata $p(y|\mathbf{x})$, dove $\mathbf{x}$ è un vettore di covariate di dimensione p.
La funzione $q_\alpha(\mathbf{x}) = F^{-1}(\alpha|x)(\cdot)$ definisce la distribuzione condizionata dei quantili di $p(y|\mathbf{x})$, con $\alpha \in (0,1)$ livello del quantile e $F^{-1}(\cdot)$ funzione di ripartizione condizionata cumulata inversa di Y.
Nella regressione quantilica (Koenker and Bassett, 1978), i quantili condizionati sono modellati individualmente, senza specificare un modello per $p(y|\mathbf{x})$. La stima diretta del quantile è ottenuta sfruttando la definizione alternativa di quantile elaborata da Koenker e Bassett:

$$q_\alpha(\mathbf{x}) = arg\min_{\xi_\alpha} \mathbb{E}\{\rho_\alpha(y - \xi_\alpha)|\mathbf{x}\} \tag{3.27}$$

dove $\rho_\alpha(\cdot)$ è la versione riscalata della funzione di perdita *pinball* (3.26), definita nel seguente modo:

$$\rho_\alpha(z) = \begin{cases} (\alpha - 1)\frac{z}{\sigma} & z < 0 \\ \alpha\frac{z}{\sigma} & z \geq 0 \end{cases}$$

Il parametro di scala $\sigma > 0$ può potenzialmente dipendere dalle covariate $\mathbf{x}$.

Nella regressione quantilica additiva generalizzata di `qgam`, la funzione di perdita *pinball* è sostituita da una sua versione lisciata, la funzione di perdita ELF, *extended log-f loss* (Fasiolo et al., 2021a), che permette di gestire la discontinuità presente in 3.26 e definita come segue:

$$\tilde{\rho}_\alpha(z) = (\alpha - 1)\frac{z}{\sigma} + \lambda log(1 + e^{\frac{z}{\lambda\sigma}}) \tag{3.28}$$

Per $\lambda \to 0^+$ si riconduce alla funzione di perdita *pinball*. Se il parametro $\lambda > 0$ è adeguatamente scelto, la funzione di perdita ELF conduce ad una stima dei quantili più accurata rispetto alla funzione di perdita *pinball*.

Il QGAM modella il quantile condizionato imponendo una struttura additiva, del tipo:

$$q_\alpha(\mathbf{x}) = \sum_{j=1}^{J} f_j(\mathbf{x})$$

dove le funzioni f_j sono *splines* costruite utilizzando l'espansione in basi, cosicché il j-esimo elemento possa essere scritto come:

$$f_j(\mathbf{x}) = \sum_{k=1}^{K_j} b_k^j(\mathbf{x})\beta_k^j$$

In questa rappresentazione $b_1^j, \ldots, b_{K_j}^j$ sono le funzioni base della *spline* modellante il j-esimo effetto e $\beta_1^j, \ldots, \beta_k^j$ sono i corrispondenti coefficienti di regressione. Le basi sono note e fisse, i coefficienti vanno stimati. Da notare che il quantile q_α dipende sia da β che da $\mathbf{x}$, e che viene presentato come $q_\alpha(\beta)$ o $q_\alpha(\mathbf{x})$ a seconda dei contesti, per comodità notazionale.

Nella logica bayesiana, la complessità delle $f_j(\cdot)$ è controllata usando la distribuzione a priori dei coefficienti di regressione $p(\beta)$. Nel seguito si assume che la $p(\beta)$ è una distribuzione a priori impropria gaussiana, centrata in 0 e con matrici di precisione (matrici inverse della matrice di varianza e covarianza) semi-definite positive $\mathbf{S}^\gamma$, definite in base a un parametro γ selezionato opportunamente.

Indicata con $p(\beta)$ la distribuzione a priori dei coefficienti di regressione β, funzione dei parametri di lisciamento γ, si assume che $p(y|\beta, \mathbf{x})$, indicata per comodità con $p(y|\beta)$, sia la vera distribuzione condizionata della variabile risposta. Si assuma per il momento che γ sia fissato, quindi che $p(\beta)$ sia nota. Dal Teorema di Bayes deriva la formula per inferire la distribuzione di probabilità a posteriori di β, sulla base dei dati y: questa è ottenuta moltiplicando la funzione di verosimiglianza $p(\mathbf{y}|\beta) = \mathcal{L}(\beta)$ per la probabilità a priori, normalizzato (diviso) per la probabilità dei dati $p(y)$, come in equazione 3.29.

$$p(\beta|\mathbf{y}) = \frac{p(\mathbf{y}|\beta)p(\beta)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\beta)p(\beta)}{\int p(\mathbf{y}|\beta)p(\beta)d\beta} \propto p(\mathbf{y}|\beta)p(\beta) \qquad (3.29)$$

Il limite di tale approccio all'applicazione della logica bayesiana per il QGAM risiede, come già annunciato in precedenza, nell'impossibilità di utilizzare la regola di Bayes per

l'aggiornamento della $p(\beta)$ nella distribuzione a posteriori $p(\beta|y)$, in quanto la regressione quantilica si basa sulla funzione di perdita ELF e non su un modello probabilistico per $p(y|\beta)$, per cui non si dispone di questo elemento. Il *framework* di aggiornamento della conoscenza pregressa, *belief updating* sviluppato da Bissiri et al. nel 2016, sfrutta la funzione di perdita ELF, consentendo di superare tale limite, con la formulazione di una pseudo-verosimiglianza. Si ha quindi la seguente formula di aggiornamento:

$$p(\beta|y) \propto \tilde{p}_\alpha\{y - q_\alpha(\beta)\}p(\beta)$$

dove la funzione $\tilde{p}_\alpha\{y - q_\alpha(\beta)\}$ rappresenta la funzione di densità di probabilità (PDF) della distribuzione ELF, estensione della distribuzione log-f di Jones (2008):

$$\tilde{p}_\alpha\{y-\mu\} = \frac{e^{-\tilde{p}_\alpha\{y-\mu\}}}{\int e^{-\tilde{p}_\alpha\{y-\mu\}}dy} = \frac{e^{(1-\alpha)\frac{y-\mu}{\sigma}}(1+e^{\frac{y-\mu}{\lambda\sigma}})^{-\lambda}}{\lambda\sigma Beta[\lambda(1-\alpha),\lambda\alpha]}$$

In questa formula λ determina il grado di lisciamento della funzione di perdita, $\frac{1}{\sigma}$ la velocità di apprendimento, che determinano anche il peso della pseudo-verosimiglianza basata su funzione di perdita e a priori rispettivamente.

I QGAMs basati su ELF, implementati da `qgam`, vengono stimati con una procedura di ottimizzazione in tre passaggi annidati, e non sequenziali, per cui per l'ottimizzazione successiva è necessario che quella precedente sia risolta e conclusa.

1. Stima del massimo a posteriori (MAP) dei coefficienti di regressione, primo passaggio di ottimizzazione, al livello più interno della procedura annidata;
2. Selezione dei parametri di lisciamento e della funzione ELF, tramite un'espressione chiusa;
3. Selezione del *learning rate* con una calibrazione bayesiana, risolta numericamente, ultimo passaggi di ottimizzazione.

Per un maggiore approfondimento circa l'algoritmo di ottimizzazione si rimanda a (Fasiolo et al., 2021a).

3.7 Validazione del modello

Una volta calcolato il *Cost-at-Risk* occorre verificare se il modello è appropriato, ovvero se effettivamente la percentuale di sforamenti è pari a α%. Per una serie di costi di lunghezza n ci si attende infatti di osservare esattamente un numero di violazioni pari a $n \cdot \alpha$.
Fondamentale per una corretta valutazione del rischio, è che il CaR di livello α venga superato α volte su 100 casi, proprietà definita di "copertura non condizionata" (*unconditional coverage*): la frequenza attesa delle violazioni osservate del CaR dovrebbe essere esattamente uguale a α. Se la probabilità incondizionata della violazione è significativamente più alta, in risposta ad un numero di sforamenti più elevato di quello atteso, allora il modello di CaR sta sottovalutando il livello di rischio. Se invece il numero di sforamenti osservati è più piccolo della copertura desiderata, il modello di CaR è troppo prudente e sta sopravvalutando il livello di rischio reale.

La procedura di verifica del corretto funzionamento del CaR prende il nome di *backtesting*, che confronta la storia dei costi e guadagni verificati con le corrispondenti stime del CaR. Dunque, affinché un modello per il CaR possa essere considerato un "buon modello" è necessario che tale soddisfi le seguenti caratteristiche:

1. la percentuale di sforamenti osservati del CaR non dovrebbe essere diversa dalla percentuale attesa α%; in altre parole, la copertura nominale e quella osservata non devono differire significativamente;

2. gli sforamenti del CaR, relativi ad uno stesso livello di copertura α, devono devono distribuirsi uniformemente nel tempo: l'indipendenza delle violazioni caratterizza un processo senza memoria, e l'eventuale presenza di violazioni ravvicinate o raggruppate è sintomo di un modello che non cattura adeguatamente la dinamica della volatilità.

Se entrambe le proprietà sono soddisfatte il modello possiede la proprietà della "copertura condizionata" (*conditional coverage*). Ciò equivale ad assumere che le violazioni si distribuiscono come una Bernoulli con probabilità α. I test statistici adoperati per la verifica di tali proprietà e utilizzati per la validazione dei modelli presentati in seguito sono il test di Kupiec (1995), il test di Christoffersen (1998) e il test DQ del quantile dinamico (*dynamic quantile*) di Engle e Manganelli (2004).

3.7.1 Test di Kupiec

Il test proposto da Kupiec (1995) verifica la prima proprietà, ovvero che la copertura osservata e quella vera siano uguali. Tale verifica non è condizionata a ulteriori ipotesi, perciò il test viene detto anche test di copertura "non condizionata". Se l'ipotesi nulla è corretta, ovvero copertura osservata e nominale coincidono, allora la probabilità di osservare k sforamenti in un campione di n osservazioni (tasso di sforamenti pari a $p = k/n$) è definita da una variabile casuale con distribuzione di Bernoulli con probabilità p di successo (cioè di sforamento) in ogni istante t, con distribuzione binomiale considerando la sua somma su tutti i t.

Sia $I_{t,\alpha}$, $t = 1, \ldots, n$ la variabile casuale dicotomica che definisce gli sforamenti, *hit* in inglese, in ogni istante, con valore 1 se il CaR di livello α nel giorno t viene sforato, 0 altrimenti.

$$I_{t,\alpha} = \begin{cases} 0 & X_t \leq CaR_{t,\alpha} \\ 1 & X_t > CaR_{t,\alpha} \end{cases} \tag{3.30}$$

Sia $S_\alpha = \sum_{t=1}^{n} I_{t,\alpha}$ la variabile che definisce la probabilità del numero di sforamenti in n prove di Bernoulli. La probabilità di osservare k sforamenti su un campione di n osservazioni è

$$Pr(S_\alpha = k) = \binom{k}{n} p^k (1-p)^{n-k}$$

dove $\binom{k}{n}$ rappresenta il coefficiente binomiale.

Data la seguente forma è possibile ottenere la log-verosimiglianza e quindi la statistica log-rapporto di verosimiglianza. Nello specifico, in questo contesto il test d'ipotesi sarà

$$\begin{cases} H_0 & : p = \alpha \\ H_1 & : p \neq \alpha \end{cases}$$

dove $\hat{p} = k/n$ è la proporzione di sforamenti osservati sul campione e α il livello di significatività fissato. La statistica del log-rapporto di verosimiglianza è definita come

$$W_K = -2log\left(\frac{L(\alpha)}{L(p)}\right) = 2log\left\{\frac{\hat{p}^k(1-\hat{p})^{n-k}}{\alpha^k(1-\alpha)^{n-k}}\right\}$$

con distribuzione asintotica $W \sim \mathcal{X}_1^2$ sotto H_0.

Se il p-value calcolato è piccolo, minore di 0.05 (o 0.10), l'ipotesi nulla H_0 viene rifiutata ad un livello del 5% (o 10%). Si deve in questo caso concludere che il modello CaR non è sufficientemente accurato, perché copertura osservata e nominale sono significativamente diverse.
Viceversa, se il p-value risultante è maggiore del valore soglia scelto, l'ipotesi nulla H_0 non viene rifiutata e si può definire come adeguato il modello CaR verificato.

In generale, i test sono esposti a due tipi di errori. Gli errori del primo tipo si concretizzano nel rifiutare H_0 quando è corretta, quelli del secondo tipo, nell'accettare H_1 quando è falsa. Esiste un *trade-off* tra i due tipi di errore; per finalità di valutazione del rischio si è interessati alla capacità del test di minimizzare l'errore del secondo tipo, ovvero alla sua potenza. È stato dimostrato che la potenza statistica del test di Kupiec (definita come il complemento a uno dell'errore del secondo tipo) è piuttosto bassa. In generale il test di Kupiec richiede un campione composto da un numero elevato di dati per poter generare risultati veramente affidabili. Infatti il test di Kupiec si focalizza unicamente sul numero di eccezioni e non considera la loro distribuzione temporale, tanto che un modello che alterna periodi in cui il CaR è sottostimato, ovvero in cui vi è un numero elevato di sforamenti, a periodi in cui il CaR è sovrastimato, in cui gli sforamenti sono pochi, potrebbe risultare accettabile. Quindi anche di fronte a sforamenti ravvicinati, il test di Kupiec potrebbe affermare che il modello di CaR analizzato è un buon modello.

Questo limite del test di Kupiec è superato dal test di Christoffersen.

3.7.2 Test di Christoffersen

Il test di Christoffersen (1998) permette di verificare che le proprietà di copertura non condizionata e di indipendenza degli sforamenti valgano contemporaneamente. Infatti, in un buon modello, che reagisce correttamente a nuova informazione, la probabilità che si verifichi uno sforamento nel giorno t dovrebbe essere indipendente da eventuali eccezioni registrate il giorno $t-1$. Il test di Christoffersen prende il nome di test di copertura condizionata (*conditional coverage*), proprio perché si verifica che la copertura osservata coincida con la nominale condizionatamente alla verifica dell'indipendenza degli sforamenti.

Sia $I_{t,\alpha}$, $t = 1, \ldots, n$ la variabile casuale dicotomica che definisce gli sforamenti, presentata in 3.30.

Il test di Christoffersen verifica l'ipotesi nulla che gli sforamenti I_t siano indipendenti contro l'alternativa di dipendenza markoviana del primo ordine, per cui I_t dipende solo dal I_{t-1}.

SI definisce con Π_1 la matrice di transizione

$$\Pi_1 = \begin{bmatrix} 1-\pi_{01} & \pi_{01} \\ 1-\pi_{11} & \pi_{11} \end{bmatrix} = \begin{bmatrix} \pi_{00} & \pi_{01} \\ \pi_{10} & \pi_{11} \end{bmatrix}$$

con $\pi_{ij} = Pr(I_t = j | I_{t-1} = i)$, per $i,j = \{0,1\}$ la probabilità di transizione dallo stato i allo stato j. Quindi $\pi_{11} = Pr(I_t = 1 | I_{t-1} = 1)$ rappresenta la probabilità che uno sforamento in $t-1$ sia seguito da un altro sforamento in t, e il suo complemento a 1 è $\pi_{10} = Pr(I_t = 0 | I_{t-1} = 1) = 1 - \pi_{11}$, la probabilità che una violazione in $t-1$ sia seguita da una non violazione in t. π_{11}, la probabilità che non vi siano violazioni in $t-1$ nè in t, e il suo complemento a 1 è $\pi_{01} = Pr(I_t = 1 | I_{t-1} = 0) = 1 - \pi_{11}$, la probabilità che si verifichi una violazione in t senza che questa si sia verificata in $t-1$. Sia n_{ij} il numero di transizioni.

Sotto l'ipotesi di dipendenza markoviana del primo ordine, la verosimiglianza degli I_t è:

$$\mathcal{L}_1(\pi_{i,j} | I_1, \ldots, I_n) = (1-\pi_{01})^{n_{00}} \pi_{01}^{n_{01}} (1-\pi)^{n_{10}} \pi_{11}^{n_{11}} \qquad (3.31)$$

Si può dunque ottenere lo stimatore per π_{ij} calcolando la derivata prima di della logverosimiglianza $log\mathcal{L}_1(\pi_{i,j} | I_1, \ldots, I_n)$, ottenendo n_{ij}/π_{ij}. Eguagliando a zero la precedente quantità e tenendo conto che le probabilità devono sommare a 1, si giunge a:

$$\hat{\pi}_{ij} = \frac{\pi_{ij}}{\sum\limits_{j \in \{0,1\}} n_{ij}}$$

per $i \in \{0,1\}$, che è quindi il valore che massima la 3.31. La matrice di transizione stimata $\hat{\Pi}_1$ diventa:

$$\hat{\Pi}_1 = \begin{bmatrix} \frac{n_{00}}{n_{00}+n_{01}} & \frac{n_{01}}{n_{00}+n_{01}} \\ \frac{n_{10}}{n_{10}+n_{11}} & \frac{n_{11}}{n_{10}+n_{11}} \end{bmatrix}$$

Nel caso invece di indipendenza degli sforamenti, si ha $\pi_{ij} = Pr(I_t = j | I_{t-1} = i) = Pr(I_t = j) = \pi_j$ con $i,j = \{0,1\}$, cioè $Pr(I_t = 1) = \pi$ e $Pr(I_t = 0) = 1 - \pi$.

Modificando sulla base di questo la 3.31, si ottiene

$$\mathcal{L}_2(\pi_{i,j}|I_1,\ldots,I_n) = \pi^{n_{01}+n_{11}}(1-\pi)^{n_{10}+n_{00}} \tag{3.32}$$

In questo caso, il valore che massimizza la 3.32 è

$$\hat{\pi} = \frac{\sum\limits_{i\in\{0,1\}} n_{i1}}{\sum\limits_{i\in\{0,1\}}\sum\limits_{j\in\{0,1\}} n_{ij}}$$

Il test W_C per la sola ipotesi di indipendenza degli sforamenti è:

$$W_C = -2log\Big(\frac{\mathcal{L}_2(\hat{\pi})}{\mathcal{L}_1(\hat{\pi}_{ij})}\Big) \sim \mathcal{X}_1^2$$

mentre quello per la verifica congiunta dell'indipendenza e della copertura W_{KC} diventa:

$$W_{KC} = -2log\Big(\frac{\mathcal{L}_2(\alpha)}{\mathcal{L}_2(\hat{\pi})}\Big) - 2log\Big(\frac{\mathcal{L}_2(\hat{\pi})}{\mathcal{L}_1(\hat{\pi}_{ij})}\Big) = -2log\Big(\frac{\mathcal{L}_2(\alpha)}{\mathcal{L}_1(\hat{\pi}_{ij})}\Big) =$$
$$= -2log\Big(\frac{\pi_{01}^{n_{01}}(1-\pi_{01})^{n_{00}}\pi_{11}^{n_{11}}(1-\pi_{11})^{n_{10}}}{\alpha^k(1-\alpha)^{n-k}}\Big) \sim \mathcal{X}_2^2$$

Per la proprietà dei logaritmi:

$$W_{KC} = W_C + W_K$$

Il test di Christoffersen è una metodologia più efficiente e completa rispetto al test di Kupiec, in quanto tiene conto del problema dell'indipendenza e inoltre, la scomposizione in due componenti permette di evidenziare le cause che conducono al rifiuto del modello.

Nella verifica congiunta, se il p-value calcolato è piccolo, minore di 0.05 (o 0.10), l'ipotesi nulla H_0 viene rifiutata ad un livello del 5% (o 10%), a favore di H_1, per la quale gli sforamenti risultano legati da una dipendenza markoviana del primo ordine. Si deve in questo caso concludere che il modello CaR non è sufficientemente accurato, perché gli sforamenti non sono indipendenti o perché la copertura nominale e quella osservata non coincidono, o entrambi i casi. Viceversa, se il p-value risultante è maggiore del valore soglia scelto, l'ipotesi nulla H_0 non viene rifiutata e si può definire come adeguato il modello CaR verificato.

Tuttavia questo test dipende dalla frequenza con cui si verificano sforamenti successivi. Se nel periodo considerato riguardano eventi rari, il test ha scarsa potenza. Per questo occorre fare attenzione che il periodo di stima sia abbastanza lungo e contenga un'ampia varietà di situazioni differenti.

3.7.3 Test del quantile dinamico

Il test DQ del quantile dinamico (*Dynamic Quantile*), fu proposto da Engle e Manganelli (2004) e introdotto nella letteratura della regressione quantilica, dopo essere stato derivato indipendentemente da Chernozhukov (1999).

Venne sviluppato per rispondere alla necessità di controllare la dipendenza degli sforamenti, poiché può capitare che gli sforamenti vadano a creare raggruppamenti (*cluster*) mentre la copertura (non condizionata) osservata rimane pari a quella nominale.

Sia $I_{t,\alpha}$ la variabile che indica gli sforamenti definita come in 3.30.
Il test DQ è basato su una regressione lineare di $I_{t,\alpha}$ su un insieme di covariate $\mathbf{z_{t-k}}$ proveniente dall'insieme informativo fino al tempo $t-1$. Più precisamente, come mostrato in 3.33, può includere ritardi della variabile degli sforamenti $I_{t-i,\alpha}$, ma anche ritardi della variabile a cui si riferisce il CaR (c_t), sue trasformazioni (come il quadrato c_t^2) o la previsione un passo in avanti del CaR ($CaR_{t-k|t-k-1,\alpha}$). Tale test infatti verifica che la variabile degli sforamenti sia incorrelata con il suo passato e con altre variabili ritardate prese dall'insieme informativo a disposizione. Ciò può essere rappresentato con:

$$\begin{aligned} I_{t,\alpha} = \delta + \sum_{k=1}^{K} \beta_k I_{t-k,\alpha} + \\ + \sum_{k=1}^{K} \gamma_k g\left[I_{t-k,\alpha}, I_{t-k-1,\alpha}, \ldots, z_{t-k}, z_{t-k-1}, \ldots\right] + \varepsilon_t \end{aligned} \tag{3.33}$$

Il sistema di ipotesi del test del quantile dinamico, in riferimento alla regressione di eq.3.33, corrisponde a:

$$\begin{cases} H_0 & : \delta = \beta_1 = \cdots = \beta_k = \gamma_1 = \cdots = \gamma_k = 0 \\ H_1 & : \textit{almeno una} \neq 0 \end{cases}$$

Per verificare ciò il test utilizza la statistica

$$W_{DQ} = \frac{I^T Z (Z^T X)^{-1} Z^T I}{\alpha(1-\alpha)} \sim \mathcal{X}^2_q$$

dove Z è la matrice delle variabili esplicative nella regressione lineare su $I_{t,\alpha}$ e I il vettore dei $I_{t,\alpha}$. Sotto l'ipotesi nulla, Engle e Manganelli hanno mostrato che la statistica W_{DQ} segue una distribuzione $\mathcal{X}^2_q$ dove $q = rank(Z)$ (Engle and Manganelli, 2004).

4

Rischio associato ai costi di approvvigionamento

In questo capitolo sono riportate le applicazioni dei modelli descritti nel capitolo 3 ai dati forniti da Terna e presentati nel capitolo 2.

Ciò che è emerso dall'analisi descrittiva delle relazioni marginali delle variabili esplicative rispetto alla variabile d'interesse è innanzitutto una notevole complessità degli effetti sulla risposta. Alcune variabili hanno mostrato avere un impatto periodico, come quelle temporali, che hanno rivelato la presenza di una componente settimanale, ma anche annuale. Inoltre si è notata la presenza di un trend temporale non lineare. È questa la tipologia di relazione maggiormente osservata: alcune variabili mostrano una relazione a volte parabolica, altre volte più complessa, altre meno marcatamente, ma sempre non lineare.

Si è quindi scelto di applicare un modello non parametrico, che sommasse congiuntamente gli effetti parziali dei regressori per l'effetto finale sulla risposta, quindi un modello di tipo additivo. In particolare, è stato adattato un modello additivo generalizzato (GAM), scelto proprio per la sua flessibilità.

I modelli GAM sono stati utilizzati sia nella nell'approccio in due fasi, sia nella modella-

zione del quantile condizionato in una fase. In tal caso, il modello adattato è un *Quantile-GAM* (QGAM), anch'esso non parametrico, con struttura additiva. Si tratta infatti di un'estensione del GAM alla modellazione del quantile condizionato.

Lo scopo finale della modellazione è il calcolo del rischio a 1 giorno e a 30 giorni. Nell'ambito del primo approccio, per poter esprimere una valutazione del rischio associato ai costi che Terna sostiene in MSD si utilizzano modelli di CaR, o Costo-a-Rischio. Più precisamente, come visto nel capitolo 3, il CaR a 1 e a 30 giorni, viene applicato sui residui della modellazione della media condizionata. Questo metodo infatti permette di modellare il quantile di livello α della distribuzione dei residui, con diversi modelli.

Nell'ambito del secondo approccio basato sul QGAM, il CaR coincide direttamente con il quantile della distribuzione dei costi condizionato al valore assunto dai regressori.

Per l'applicazione dei modelli, l'intero dataset di dati giornalieri dal 1/1/2017 al 30/9/2021, è stato suddiviso in due sottoperiodi: l'insieme di stima, periodo *in-sample*, dal 1/1/2017 al 30/9/2020, per un totale di 1369 osservazioni, e l'insieme di verifica, periodo *out-of-sample*, dal 1/10/2020 al 30/9/2021, che corrisponde agli ultimi 365 giorni di osservazioni.
Il periodo di stima è stato ulteriormente suddiviso in un insieme di prova dei modelli, dal 1/1/2017 al 30/9/2019, in cui vengono identificati e stimati i modelli, e un insieme di validazione dei modelli stimati, dal 1/10/2019 al 30/9/2020, in cui viene valutata la performance del CaR in un insieme di dati diverso da quello di verifica. Questa procedura permette di verificare la capacità predittiva dei singoli modelli, confrontandoli anche tra loro: il confronto è fatto per tutti i modelli sotto le stesse condizioni.

In seguito vengono riportati le modellazioni fatte, suddividendole in quelle relative al CaR a 1 giorno, in 4.1 e al CaR a 30 giorni, in 4.2, i due obiettivi finali di calcolo. Ognuna delle due sezioni presenta un paragrafo relativo alle applicazioni sull'insieme di stima dei modelli, le analisi *in sample* e sull'insieme di verifica, *out of sample*.

La selezione del modello tiene conto della significatività dei parametri, del loro adattamento dal punto di vista grafico per l'interpretabilità, e vengono sottoposti alla validazione tramite test statistici e calcolo della copertura. In particolare, la strategia adoperata per la validazione dei modelli è mista e coinvolge la significatività dei parametri, l'analisi dei residui, l'errore assoluto medio (MAE) nel periodo di stima, e il confronto della copertura osservata con quella nominale nel periodo di verifica. Accanto a questa vengono impiegati test come il test di Kupiec (1995) della copertura incondizionata, di Christoffersen (1998)

della copertura condizionata e il test di Engle e Manganelli (2004), che considera il quantile dinamico e può essere interpretato come un test di bontà di adattamento complessivo del metodo CaR stimato.

4.1 CaR a 1 giorno

In questa sezione sono descritti i modelli selezionati per la valutazione del rischio a 1 giorno con i due approcci metodologici, con paragrafi corrispondenti ai passi della procedura di stima (*in-sample*) 4.1.1, e di verifica (*out-of-sample)* 4.1.2.

4.1.1 Stima e validazione del modello

Il primo passo della procedura di modellazione prevede la stima del modello, realizzata con i dati del periodo di stima 1/1/2017 - 30/9/2019, e una prima validazione, effettuata utilizzando il periodo 1/10/2019 - 30/9/2020.

Per quanto riguarda il primo approccio, si inizia stimando il modello GAM per la media condizionata. Il modello additivo generalizzato scelto come migliore per la stima dei dati contiene variabili di calendario come il trend (T_t), il giorno dell'anno (DY_t), il giorno della settimana (DW_t) e la variabile relativa ai giorni non lavorativi escluse le domeniche ($bank_t$); include anche alcuni ritardi della variabile risposta, come i costi del giorno precedente (c_{t-1}) e di sette giorni prima (c_{t-7}), la quantità di energia prodotta con l'idroelettrico ($Hydro_t$) e con l'eolico ($Wind_t$), ma anche la quantità di riserva secondaria ($aFRR_t$), terziaria (RR_t) e i vincoli a rete integra (vri_t).
Il modello GAM viene applicato per la modellazione della media dei costi c_t condizionata al valore delle variabili esplicative, indicate nel loro complesso come $\mathbf{x_t}$, per cui, all'interno del modello per i costi definito come:

$$c_t = \mu_t + \varepsilon_t = \mathbb{E}[c_t|\mathbf{x_t}] + \varepsilon_t$$

si ha

$$\begin{aligned}\mu_t \quad &= \beta_0 + f_1(T_t) + f_2(DY_t) + f_3(DW_t) + bank_t + f_4(c_{t-1}) + f_5(c_{t-7}) + \\ &+ f_6(Hydro_t) + f_7(Wind_t) + f_8(aFRR_t) + f_9(RR_t) + f_{10}(vri_t) \end{aligned} \tag{4.1}$$

Le funzioni $f_i(\cdot)$ dell'eq.4.1 sono funzioni *splines*, ognuna applicata alle singole variabili $x_{i,t}$. Per ognuna si sono scelti manualmente la tipologia di *spline* e il parametro di lisciamento, k, osservando il loro adattamento ai dati prova dopo prova. Nella tabella 4.1 si riportano le opzioni scelte per il modello.

Variabile	Descrizione	Tipologia di *spline*	k
T_t	Trend	f_1: *Spline* di regressione adattiva	
DY_t	Componente annuale	f_2: *Spline* di regressione adattiva	
DW_t	Componente settimanale	f_3: *Spline* di regressione cubica	7
$bank_t$	Festività di calendario		
c_{t-1}	Costi del giorno precedente	f_4: *Thin plate spline*	2
c_{t-7}	Costi del settimo giorno antecedente	f_5: *Thin plate spline*	2
$Hydro_t$	Quantità di energia da idroelettrico	f_6: *Thin plate spline*	2
$Wind_t$	Quantità di energia da eolico	f_7: *Thin plate spline*	2
$aFRR_t$	Quantità di riserva secondaria	f_8: *Thin plate spline*	3
RR_t	Quantità di riserva terziaria	f_9: *Thin plate spline*	3
vri_t	Vincoli a rete integra	f_{10}: *Thin plate spline*	3

Tabella 4.1: Funzioni *spline* utilizzate nel modello GAM per ogni variabile, con valore scelto per k.

Si noti che per le *spline* di regressione adattive il parametro di lisciamento, ovvero la dimensione della base di *spline*, specificati con k, è di *default* pari a 40, restando libero per l'adattamento ai dati. La variabile relativa alle festività, escluse le domeniche, ufficialmente riconosciute come giorni non lavorativi, $bank_t$, essendo dicotomica, viene inserita nel modello GAM senza una funzione di lisciamento.

I gradi di libertà stimati sono: $12,14$, $6,87$, $4,79$ per le tre variabili di calendario, $1,91$, $1,70$ per i costi ritardati, $1,00$, $1,83$ per le due variabili relative alle fonti rinnovabili, $1,97$, $1,92$ per le riserve e $1,69$ per i vincoli a rete integra. I gradi di libertà rappresentano il numero di volte in cui la funzione cambia direzione, discostandosi dalla linearità. Gradi di libertà pari all'unità corrispondono a un modello lineare di primo ordine. Per cui tra le funzioni stimate, solo la produzione di idroelettrico viene stimata da una funzione lineare, tutte le altre presentano gradi di libertà superiori.

Nelle figure da 4.1 a 4.5, sono riportati i grafici degli effetti parziali dei regressori stimati con il modello GAM (in nero), con le loro bande di variabilità (in rosso).
Si tratta della stima delle varie funzioni del modello GAM definito in 4.1 e validato, al momento, sull'insieme di validazione. Le $f_i(x_{i,t})$, ognuna funzione del regressore $x_{i,t}$, con $i = 1, \ldots, 10$, modellano in modo non parametrico la relazione tra il singolo regressore e la variabile risposta, e sono riportate sullo sfondo di quelli che sono i grafici della relazione bivariata tra il regressore e la variabile risposta (punti in grigio).

I grafici sono tutti sulla stessa scala: il range sull'asse delle y è quello dei valori assunti dai

costi di approvvigionamento, in milioni di euro, mentre sull'asse delle x il range è quello dei valori delle variabili indicate. Le stime presentate in questi grafici sono in questo modo confrontabili con le relazioni marginali mostrate nel capitolo 2: in quel momento erano state presentate con una stima non parametrica della relazione, che trova ora un'adeguata modellazione a livello congiunto. È bene osservare infatti che, a differenza di quanto fatto a livello marginale e puramente descrittivo, qui la stima dei vari effetti parziali è realizzata a livello congiunto dal modello additivo generalizzato per la media. Si nota inoltre che l'insieme delle variabili scelte come regressori per i costi non coincide con l'insieme di tutte le variabili fornite: la selezione è stata fatta tenendo conto della significatività dei parametri nel GAM.

Vengono riportati gli effetti parziali sui costi delle variabili di calendario, dei costi ritardati, della produzione di energia idroelettrica da bacino ed eolica, della quantità di riserva secondaria, terziaria e dei vincoli a rete integra. Sono il risultato grafico dell'output del modello GAM.

Dall'osservazione dei grafici si nota che la modellazione non parametrica realizzata dal GAM riesce a catturare l'andamento non lineare delle variabili. A livello interpretativo, si ricorda che essendo la variabile costi a valori negativi, l'andamento crescente delle funzioni rappresentate indicano una riduzione dei costi, viceversa l'andamento decrescente ne indicano l'aumento.

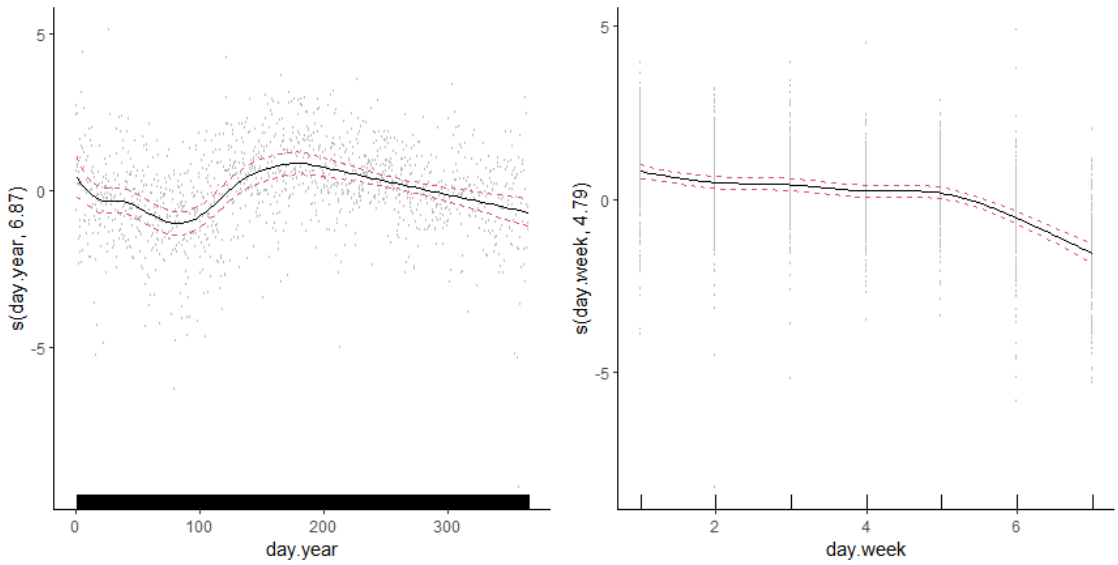


Figura 4.1: Effetto parziale di alcuni regressori sulla media dei costi: variabili di calendario relative alla componente annuale (a sinistra) e settimanale (a destra).

In Figura 4.1 sono riportate gli effetti parziali di alcune variabili di calendario. la componente annuale e della settimanale. Il grafico a sinistra, relativo alla componente annuale, riflette l'andamento mostrato dal boxplot di Figura 2.3: la funzione che stima tale componente mostra un andamento non lineare prima decrescente, fino all'incirca al valore 100, corrispondente al mese di aprile, poi crescente, nei mesi estivi e infine di nuovo decrescente, relativo all'aumento dei costi nei mesi autunnali.

La componente settimanale, grafico a destra di Figura 4.1 è stimata da una funzione dall'andamento non lineare. Nei giorni feriali, da 1, per il lunedì a 5 per il venerdì, la funzione è all'incirca costante, per poi aumentare nel fine settimana. Questo andamento riflette quanto osservato a livello empirico in 2.2, dove i costi nei giorni festivi, aumentano sia in media che a livello di volatilità.

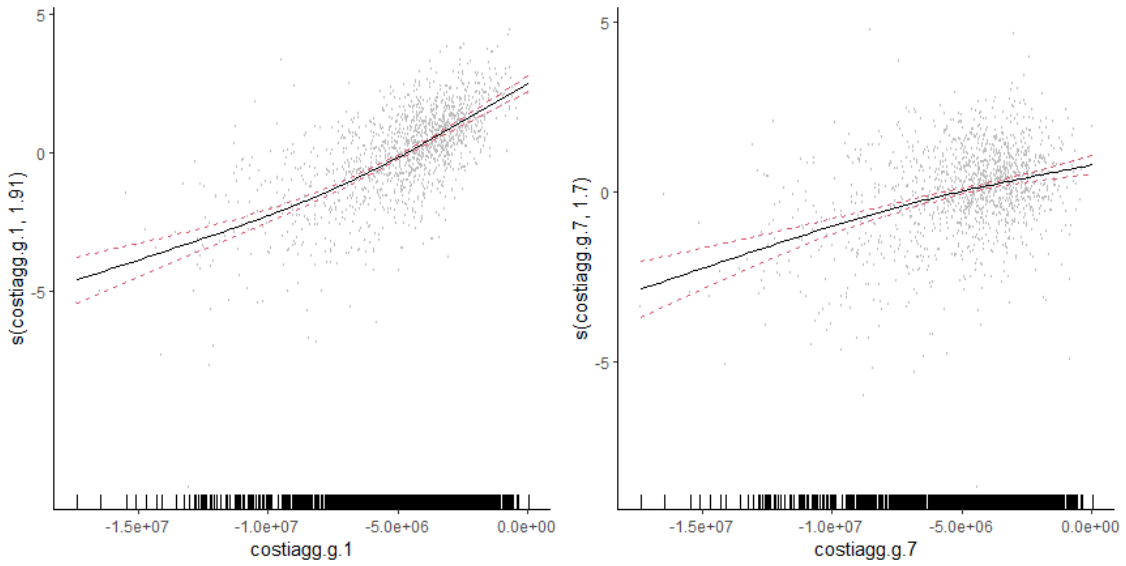


Figura 4.2: Effetto parziale di alcuni regressori sulla media dei costi: costi al ritardo 1 e 7.

Seguono i due grafici relativi alla relazione stimata tra i costi di approvvigionamento al tempo t e i suoi ritardi $t-1$ e $t-7$, in Figura 4.2. L'andamento nei due grafici è simile e indica che costi maggiori nel giorno precedente, così come nel settimo giorno antecedente, corrispondono a costi maggiori nel giorno t, viceversa, con relazione crescente, alla riduzione dei costi ritardati si ha la riduzione dei costi al tempo t. Se l'andamento è lo stesso, la relazione stimata è più marcatamente non lineare per i costi in $t-1$ rispetto ai costi in $t-7$, con bande di variabilità più ampie, quindi maggiore incertezza per valori grandi dei costi.

Il grafico in Figura 4.3 rappresenta l'effetto parziale stimato dei vincoli a rete integra, il numero di unità produttive in esercizio per la garanzia di sicurezza del sistema. La funzione di lisciamento della relazione tra questa variabile e la risposta indica che all'aumentare delle UP accese in risposta al bisogno di garantire sicurezza al sistema, aumentano i costi di approvvigionamento di riserva. L'aumento dei costi all'aumentare delle necessità di Terna si verifica sia per effetto della maggiore domanda del servizio, sia per le condizioni di ridotta competitività che portano all'aumento dei prezzi.

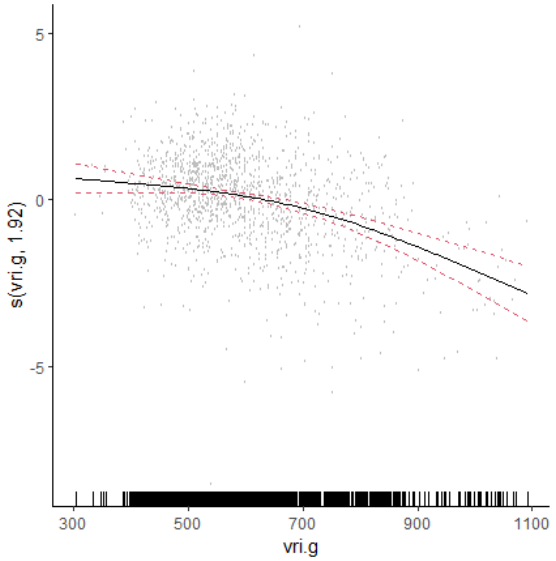


Figura 4.3: Effetto parziale di alcuni regressori sulla media dei costi: vincoli a rete integra.

In Figura 4.4 sono mostrati gli effetti parziali della quantità di energia prodotta da fonti rinnovabili come l'idroelettrico da bacino e l'eolico corrispondenti ai primi due grafici. All'aumentare della produzione di idroelettrico i costi aumentano: l'idroelettrico da bacino è una tipologia di risorsa energetica che viene attivata per rispondere ad aumenti della domanda, in particolare, grazie alla possibilità di attivarle e disattivarle in pochi minuti con l'immediata apertura delle saracinesche idrauliche, sono impiegate nella risposta a picchi di fabbisogno. La quantità di energia prodotta con l'idroelettrico da bacino, corrisponde pertanto a situazioni di elevata domanda, che accompagnano costi più elevati anche in MSD. L'effetto parziale stimato riflette anche in questo caso quanto osservato empiricamente in Figura 2.6a.

La stima dell'effetto della produzione di energia elettrica da eolico sui costi ha anch'essa un aspetto non lineare, e all'aumentare della quantità prodotta da eolico si ha un sensibile

aumento dei costi, oltre che un aumento della variabilità associata a questa stima. Ciò è principalmente dovuto alla scarsità di osservazioni in quest'area del grafico. Anche in questo caso la stima approssima la relazione osservata in Figura 2.6c.

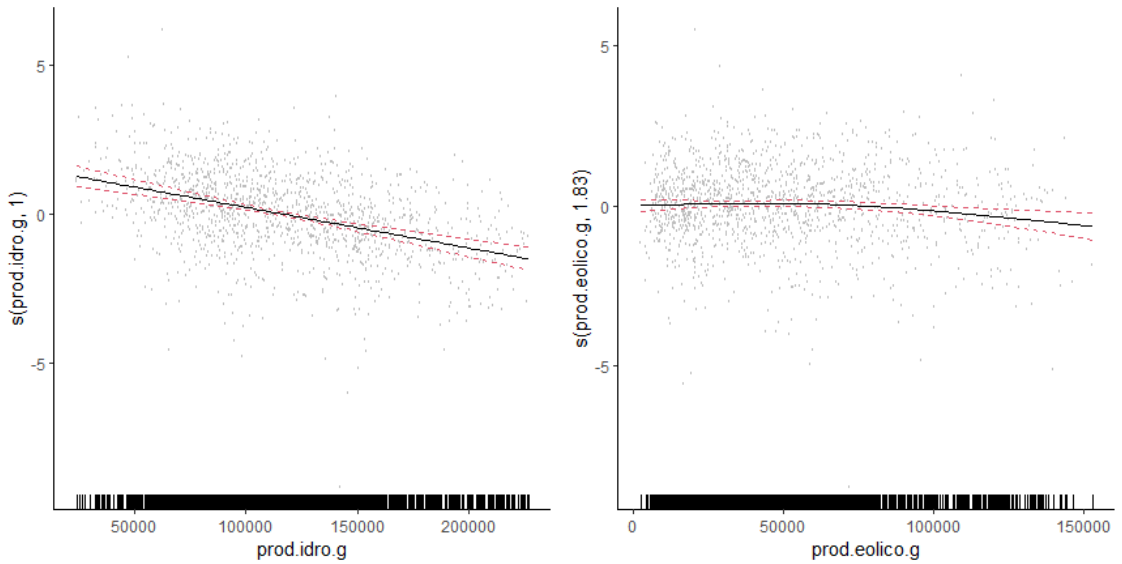


Figura 4.4: Effetto parziale di alcuni regressori sulla media dei costi: variabili relative alle fonti rinnovabili.

I due grafici successivi di Figura 4.5 mostrano infine l'effetto parziale sui costi della quantità di riserva secondaria nazionale attivata "a salire". La sua attivazione avviene per sostenere il bilanciamento tra carichi e prelievi in seguito ad un calo della frequenza, ad esempio a causa dello scatto di un generatore, con un aumento della produzione immessa nella rete o con una riduzione della produzione prelevata. Innanzitutto si nota che la funzione presenta un andamento non lineare e ai due estremi aumenta l'incertezza connessa alla stima per via della scarsità di osservazioni. La relazione sembra suggerire che all'aumentare della quantità di riserva secondaria diminuiscano i costi, ma questo andamento è legato ad un cambiamento della gestione della riserva verificatosi negli anni, come era stato visto in Figura 2.8. Infine viene presentato l'effetto della riserva terziaria nazionale, che lisciando l'andamento leggermente parabolico mostrato in Figura 2.9, acquisisce un andamento leggermente crescente.

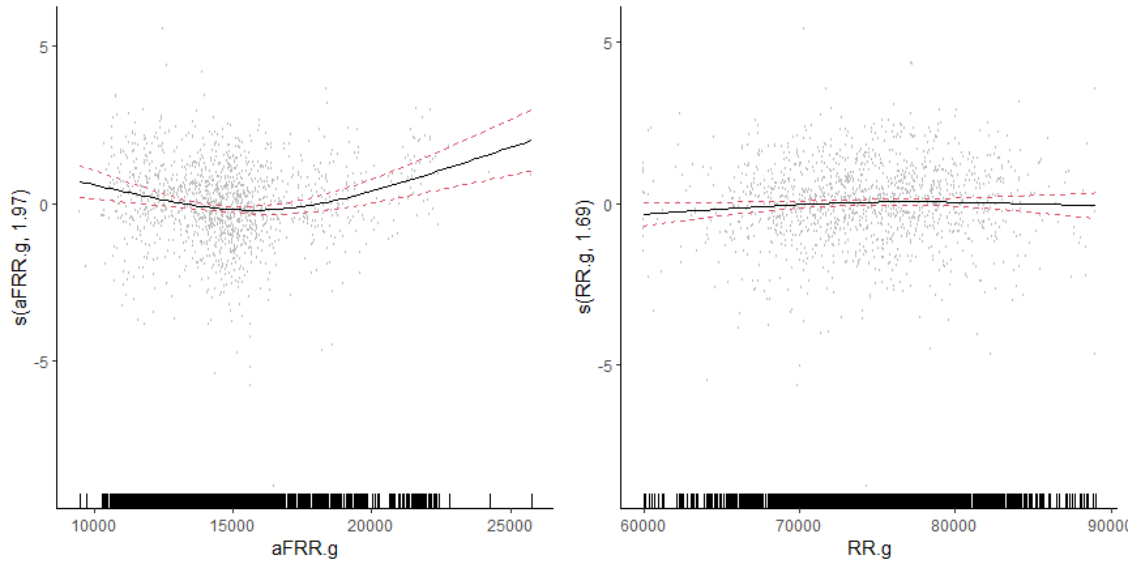


Figura 4.5: Effetto parziale di alcuni regressori sulla media dei costi: variabili relative alle due tipologie di riserva, secondaria e terziaria.

Il modello in 4.1 per la media condizionata stima $\hat{\mu}_t$. Sono quindi calcolabili i residui del modello GAM rispetto ai costi $\hat{\varepsilon}_t = c_t - \hat{\mu}_t$, sui quali vengono applicati i modelli di calcolo del CaR definiti nella sezione 3.3.

- Il modello di CaR basato su distribuzione *kernel*, il GAM-K, modella la distribuzione marginale di $\hat{\varepsilon}_t$ con un approccio non parametrico basato su nucleo gaussiano.
- Nel modello di CaR basato sulla classe GARCH, il GAM-GARCH, viene applicato un GARCH$(1,1)$ a media zero sui residui stimati. In riferimento alla definizione di GARCH(p,q) con $p = q = 1$ in 3.13 e 3.14 si ha

$$\hat{\varepsilon}_t = \nu_t \sigma_t$$

la cui varianza condizionata σ_t è stimata con

$$\hat{\sigma}_t^2 = \hat{\omega} + \hat{\alpha}_1 \hat{\varepsilon}_{t-1}^2 + \hat{\beta}_1 \hat{\sigma}_{t-1}^2$$

dove $\hat{\omega} = 1420.0 \cdot (10)^8, \hat{\alpha}_1 = 0.0997, \hat{\beta}_1 = 0.8266$.

La stima è stata ottenuta con il metodo QMLE, cioè di quasi massima verosimiglianza, evitando assunzioni sulla distribuzione condizionata delle innovazioni. Pertanto il quantile condizionato di $\hat{\varepsilon}_t$ è basato sul quantile empirico delle innovazioni stimate $\hat{\nu}_t$.

- Il modello di CaR basato sulla regressione quantilica, il GAM-QR, è stato definito scegliendo come regressori del quantile condizionato i residui stimati di ritardo 1 e 7, $\hat{\varepsilon}_{t-1}$ e $\hat{\varepsilon}_{t-7}$ e la media mobile sugli ultimi 7 giorni dei residui al quadrato, indicata con $mm\hat{\varepsilon}^2_{t-1}$. La stima del quantile condizionato di $\hat{\varepsilon}_t$ diventa

$$\hat{q}^{\hat{\varepsilon}}_{0.10,t} = -1.31 \cdot (10)^6 + 0.019917\hat{\varepsilon}_{t-1} + 0.15851\hat{\varepsilon}_{t-7} - 2.25 \cdot (10)^{-7} mm\hat{\varepsilon}^2_{t-1}$$

- Si considera anche la specificazione del modello di regressione quantilica introdotta da Engle and Manganelli (2004), il modello CAViaR. La dinamica del quantile condizionato di $\hat{\varepsilon}_t$ stimato con il modello GAM-CAViaR è data da

$$\hat{q}^{\hat{\varepsilon}}_{0.10,t} = 0.2293 + 0.8997\hat{q}^{\hat{\varepsilon}}_{0.10,t-1} - 0.1195\hat{\varepsilon}^{+}_{t-1} + 0.2144\hat{\varepsilon}^{-}_{t-1}$$

dove $\hat{\varepsilon}^{+}_t$ e $\hat{\varepsilon}^{-}_t$ indicano valori positivi e negativi rispettivamente di $\hat{\varepsilon}_t$

Tutte le variabili che rientrano nei modelli rappresentati sono significative al 5%.

Per quanto riguarda il secondo approccio, di modellazione diretta in una fase del quantile condizionato con il QGAM, un primo modello applicato per la stima del quantile condizionato è lo stesso definito per la media condizionata in eq.4.1, ovvero:

$$\begin{aligned} \hat{q}^{c}_{0.10,t} \quad &= \beta_0 + f_1(T_t) + f_2(DY_t) + f_3(DW_t) + bank_t + f_4(c_{t-1}) + f_5(c_{t-7}) \\ &\quad + f_6(Hydro_t) + f_7(Wind_t) + f_8(aFRR_t) + f_9(RR_t) + f_{10}(vri_t) \end{aligned} \tag{4.2}$$

Le variabili che impattano sulla media condizionata, hanno un effetto chiaramente anche sul quantile condizionato al 10% della stessa variabile risposta, ma differente. Infatti, dai risultati ottenuti validando il modello nell'insieme di stima, il QGAM definito in 4.2, con le stesse variabili e funzioni di tabella 4.1, non risulta essere un buon modello, sia in relazione alla significatività degli elementi del modello, che alla copertura osservata.

Relativamente alla significatività delle variabili in particolare, si è osservato che $aFRR_t$ e RR_t le due variabili relative alla quantità di riserva attivata, non sono significative, mentre

$Wind_t$ è significativa solo per alti livelli della produzione. Il modello QGAM costruito includendo solo le variabili significative, indicato con QGAM_{sig}, corrisponde a:

$$\begin{aligned}\hat{q}^{c}_{0.10,t} &= \beta_0 + f_1(T_t) + f_2(DY_t) + f_3(DW_t) + bank_t \\ &\quad + f_4(c_{t-1}) + f_5(c_{t-7}) + f_6(Wind_t) + f_7(vri_t)\end{aligned} \tag{4.3}$$

Tuttavia, dal punto di vista della copertura questa è maggiore di quella attesa. La percentuale di sforamenti risulta infatti pari al 6.3%, un valore troppo basso rispetto al livello nominale del 10%, e tale elemento ci porta a dover riconsiderare il modello QGAM.

Dopo una serie di prove, si è giunti a definire il miglior modello QGAM per la performance, valutata sulla base della copertura, nell'insieme di stima. Si tratta del più semplice:

$$\hat{q}^{c}_{0.10,t} = \beta_0 + f_1(DY_t) + f_2(DW_t) + bank_t + f_4(c_{t-1}) + f_5(c_{t-7}) \tag{4.4}$$

Tale modello viene indicato con QGAM_{cov}, dove il pedice "*cov*" sta per *coverage*, essendo questo modello scelto per la sua copertura. La sua percentuale di sforamenti risulta pari al 10.4%, un ottimo risultato se confrontata con quella attesa del 10%.

Quando questi modelli vengono utilizzati per la stima del $CaR_{t,1,0.10}$ nell'insieme di validazione, portano ad avere i risultati mostrati in tabella 4.2. I modelli sono valutati sulla base del livello osservato α_{oss}, del valore del test di Kupiec (Kup), di Christoffersen (Chris) e del test del quantile dinamico (DQ) di Engle e Manganelli.

Modello	α_{oss}	**Kup**	**Chris**	**DQ**
GAM-K	0.1123	0.3404	0.0079	0.8355
GAM-GARCH	0.1062	0.665	0.906	0.980
GAM-QR	0.1069	0.6658	0.6487	0.9999
GAM-CAViaR	0.0876	0.423	0.305	0.9799
QGAM_{sig}	0.063	0.012	0.036	0.257
QGAM_{cov}	0.1041	0.7948	0.7396	0.9999

Tabella 4.2: Livello di copertura osservato α_{oss} rispetto a quello nominale pari al 10% nell'insieme di validazione (365 osservazioni dal 1/10/2019 al 30/9/2020) e valore del p-value dei test di Kupiec (Kup), Christoffersen (Chris) e del quantile dinamico (DQ).

Si nota che, ad eccezione del GAM-K e $QGAM_{sig}$, per tutti i modelli i test non rifiutano l'ipotesi nulla risultando tutti molto conservativi. I livelli di α_{oss} migliori sono quelli del GAM-GARCH e del $QGAM_{cov}$, con una copertura di circa l'89% per entrambi. Poiché il modello $QGAM_{cov}$ si basa sull'esclusione di variabili significative, una strategia non usuale e ottimale dal punto di vista statistico, si considera come miglior modello il GAM-GARCH e il $QGAM_{cov}$ come seconda scelta.

4.1.2 Verifica della capacità predittiva

Le analisi condotte finora hanno riguardato l'insieme di stima, fino al 30/9/2020. Sui modelli fissati dalla stima e una prima validazione, si esegue la verifica vera e propria, testandoli nel periodo dal 1/10/2020 al 30/9/2021. Idealmente, non dovremmo considerare quei modelli che non funzionano adeguatamente nell'insieme di stima, ma per un avere una conferma vengono utilizzati anch'essi nella previsione. I diversi modelli di $CaR_{t,1,0.10}$ stimati e i risultati dei test di validazione vengono riportati in Figura 4.6 e in Tabella 4.3.

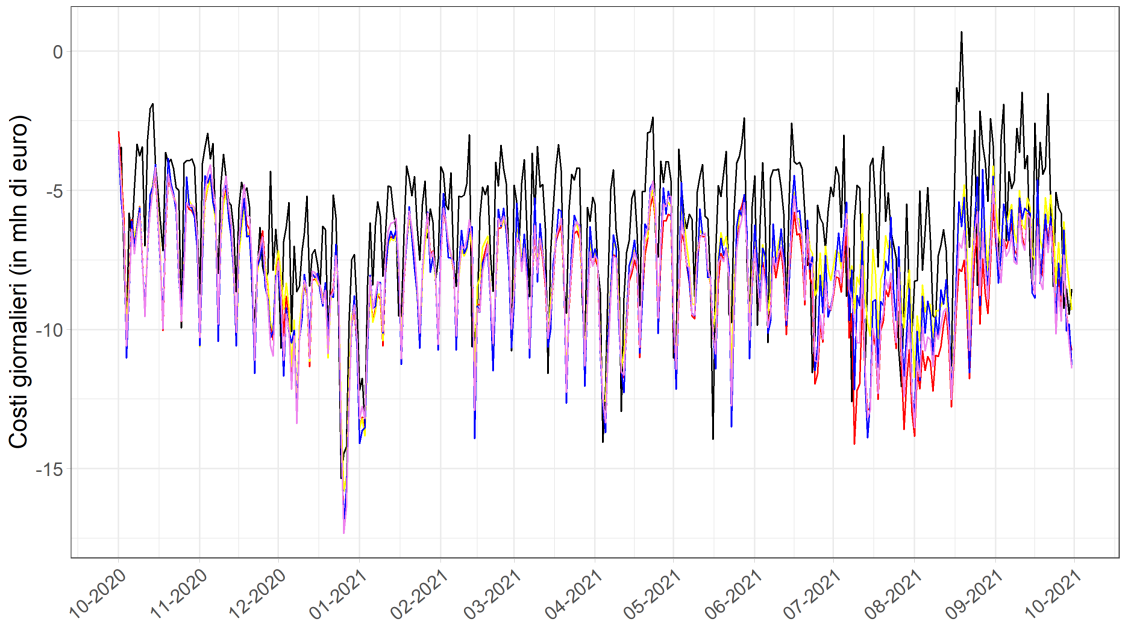


Figura 4.6: Serie storica dei costi giornalieri, in milioni di euro, per il periodo di verifica dal 1/10/2020 al 30/9/2021, in nero, e applicazione dei modelli CaR a 1 giorno stimati: GAM-K (in giallo), GAM-GARCH (in rosso), GAM-QR (in blu) e GAM-CAViaR (in viola).

La figura 4.6 mostra le serie dei $CaR_{t,1,0.10}$ calcolati con i tutti i vari modelli. A una prima occhiata sembrano abbastanza simili tra loro, e per la maggior parte del periodo di verifica risultano sovrapposti. Tuttavia, il loro diverso comportamento nel periodo compreso tra inizio luglio 2021 e settembre 2021, segna le differenze a livello di copertura mostrate in tabella 4.3.

Modello	α_{oss}	**Kup**	**Chris**	**DQ**
GAM-K	0.1671	< 0.001	< 0.001	0.1570
GAM-GARCH	0.1178	0.2686	0.0554	0.9999
GAM-QR	0.1479	0.004	0.005	0.9745
GAM-CAViaR	0.1452	0.0065	< 0.001	0.7914
$QGAM_{sig}$	0.1510	0.002	0.001	0.684
$QGAM_{cov}$	0.1534	0.0015	< 0.001	0.7728

Tabella 4.3: Livello di copertura osservato α_{oss} rispetto a quello nominale pari al 10% nell'insieme di verifica (365 osservazioni dal 1/10/2020 al 30/9/2021) e valore del p-value dei test di Kupiec (Kup), Christoffersen (Chris) e del quantile dinamico (DQ), per i modelli di CaR a 1 giorno.

Tutti i modelli, anche quelli che si erano mostrati più conservativi nella prima validazione, mostrano un numero di violazioni del caR maggiore di quello atteso. In particolare, la copertura osservata del $QGAM_{cov}$ risulta essere del 86.5%, contro una copertura attesa del 90%, risultato che conferma i limiti di tale modello, che esclude variabili significative nell'insieme di stima ai fini della copertura.

Il modello migliore risulta essere il GAM-GARCH. La sua copertura osservata è pari al 88%, con una percentuale di sforamenti dell'11.78% a fronte del livello nominale del 10%. In Figura 4.7 viene mostrata la serie del $CaR_{t,1,0.10}$ calcolato con il modello GAM-GARCH (in rosso), rispetto al valore dei costi giornalieri e della media condizionata, in nero e verde rispettivamente.

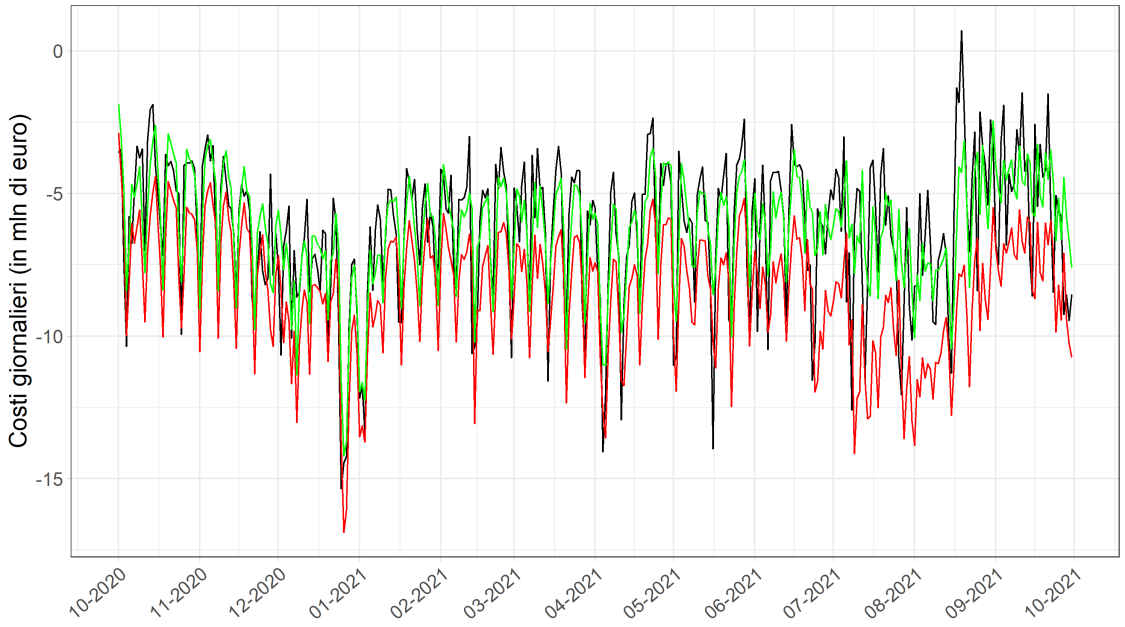


Figura 4.7: Serie storica dei costi giornalieri, in milioni di euro, per il periodo di verifica dal 1/10/2020 al 30/9/2021 (in nero), applicazione del modello sulla media condizionata (in verde) e del modello di CaR a 1 giorno con GAM-GARCH (in rosso).

La validità della copertura osservata del GAM-GARCH viene verificata con i test di Kupiec, Christoffersen e del quantile dinamico, così come per tutti gli altri modelli. Anche in questa situazione il modello GAM-GARCH dà buoni risultati, mostrati in tabella 4.5: il test di Kupiec non rifiuta l'ipotesi di uguaglianza tra il livello nominale α e α_{oss}, il test di Christoffersen, con un p-value vicino alla soglia del 5%, pari a 0.0554, non rifiuta l'ipotesi di indipendenza degli sforamenti, anche se ad un livello di confidenza più preciso potrebbe rivelare una relazione di dipendenza. Il test del quantile dinamico, che verifica una possibile relazione lineare tra gli sforamenti e i suoi ritardi e ritardi dei costi, e del CaR, rifiuta l'ipotesi di assenza di questo tipo di relazione.

Il grafico a sinistra in Figura 4.8 rappresenta l'andamento nel periodo di verifica della variabile dicotomica degli sforamenti I_t, che assume valori 1 se in t il CaR è stato violato, 0 altrimenti. Tale grafico è utile per valutare se la distribuzione degli sforamenti è uniforme, caratteristica fondamentale di un buon modello CaR. Per il CaR basato su GAM-GARCH in analisi si può dire che ciò sia rispettato.
Il grafico a destra in Figura 4.8 mostra invece l'ampiezza degli sforamenti del CaR, in milioni di euro. Gli sforamenti corrispondono alle barre nella parte inferiore del grafico: questi

sono i valori tali per cui $c_t - CaR_{t,\alpha} < 0$, ovvero $c_t < CaR_{t,\alpha}$, per cui I_t assume valore 1. Si ricordi la definizione di I_t di eq.3.30.

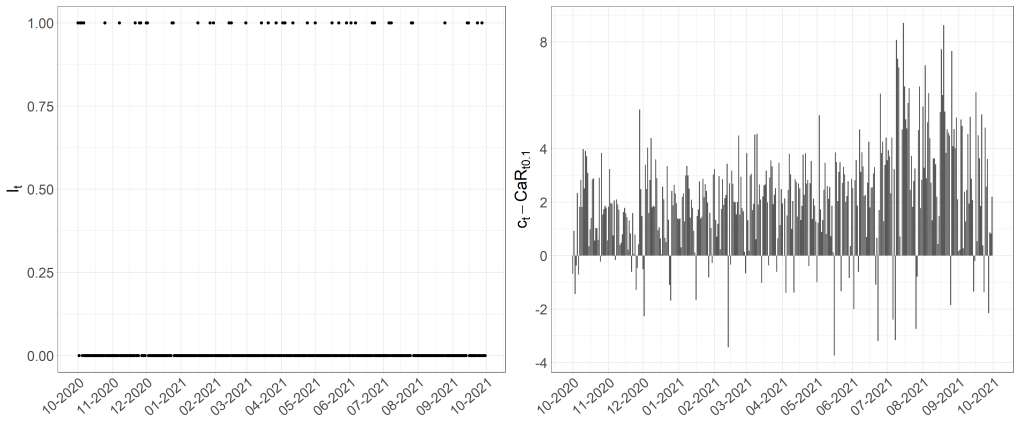


Figura 4.8: Sequenza (a destra) e ampiezza (a sinistra) degli sforamenti del $CaR_{t,1,0.10}$ calcolato con modello GAM-GARCH osservati nel periodo di verifica dal 1/10/2020 al 30/9/2021.

Infine, in tabella 4.4, sono riportate alcune statistiche descrittive relative al $CaR_{t,1,0.10}$.

Media $\mathbf{CaR_{t,30,0.10}}$	**Mediana** $\mathbf{CaR_{t,30,0.10}}$	$\mathbf{Q_{0.25}}$ $\mathbf{CaR_{t,30,0.10}}$	$\mathbf{Q_{0.75}}$ $\mathbf{CaR_{t,30,0.10}}$
-8.411	-8.082	-9.822	-6.821

Tabella 4.4: Alcune statistiche riassuntive della distribuzione del $CaR_{t,1,0.10}$: media, mediana, quantile al 25% e al 75%, espresse in milioni di euro.

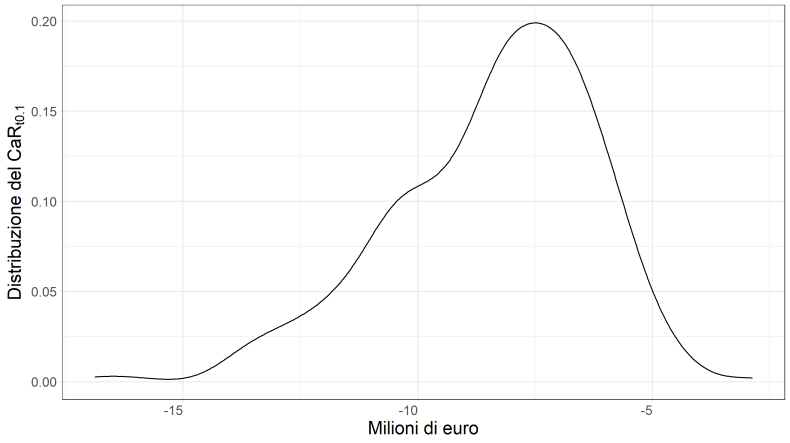


Figura 4.9: Distribuzione del $CaR_{t,1,0.10}$ calcolato con modello GAM-GARCH.

4.2 CaR a 30 giorni

In questo paragrafo viene presentato il calcolo del rischio a 1 mese. Assumendo che un mese sia costituito da 30 giorni, il problema è equivalente al calcolo del CaR a $h = 30$ giorni. Questa procedura è molto più complessa di quella del calcolo del CaR a 1 giorno, e vi sono diverse metodologie perseguibili.

La prima, la più intuitiva, consiste nel considerare la somma cumulata dei costi giornalieri sui 30 giorni con finestre fisse, cioè sommando di 30 in 30 i costi giornalieri e applicando i modelli di CaR visti in precedenza per il CaR a 1 giorno sulla serie dei costi mensili così ottenuti. Tuttavia, questo metodo comporta la riduzione del numero di osservazioni di un fattore di scala pari a 30, per cui si passerebbe da una serie di 1734 osservazioni giornaliere, ad una serie di 57 osservazioni su base mensile, troppo poche per poter stimare correttamente i modelli.

La seconda possibilità è quella di utilizzare finestre di 30 giorni consecutivi, ma non fisse, bensì mobili e parzialmente sovrapposte (*rolling windows*). Tali finestre sono ottenute spostandole di 1 giorno alla volta: in tal caso finestre consecutive avrebbero $h - 1$, cioè 29 giorni in comune, ma si potrebbe scegliere di spostarle anche di 15 giorni alla volta, ad esempio, e in tal caso condividerebbero meno osservazioni. L'utilizzo di finestre con elementi in comune comporta tuttavia l'introduzione di una forte autocorrelazione spuria, da modellare ulteriormente.

La terza proposta consente di superare le problematiche legati alle prime due sfruttando l'idea della simulazione, per calcolare il quantile d'interesse sulla distribuzione simulata dei residui del modello sulla media ad h giorni. Dal momento che il miglior modello per il CaR a 1 giorno risulta essere il GAM-GARCH, il metodo che, adattato al nostro contesto, consente il calcolo del CaR a 30 giorni, è la simulazione storica filtrata (Barone Adesi, 2015).

4.2.1 Stima e validazione del modello

Analogamente alla definizione del CaR a 1 giorno, l'espressione generale del CaR a 30 giorni è

$$CaR_{t,30,\alpha} = \hat{\mu}_{(30)|t} + q_{\alpha}^{\hat{\varepsilon}_{(30)|t}} \tag{4.5}$$

dove $\hat{\mu}_{(30)|t}$ rappresenta la previsione della media dei costi sostenuti nei successivi 30 giorni, $q_{\alpha}^{\hat{\varepsilon}_{(30)|t}}$ rappresenta il quantile di livello α della distribuzione degli errori di previsione 30 giorni in avanti $\hat{\varepsilon}_{(30)|t}$. Per calcolare il $CaR_{t,30,\alpha}$ è stata impiegata la seguente procedura:

1. calcolo di $\hat{\mu}_{(30)|t}$: in ogni t, viene calcolata la previsione a 30 giorni della media condizionata per i costi giornalieri, utilizzando un modello additivo generalizzato (GAM). Per fare ciò, vengono dapprima calcolate in t, con $t = 1, \ldots, n$ le previsioni del costo medio in $t + i$ con i dati fino a $t + i - 1$, $i = 1, \ldots, 30$, $\hat{\mu}_{t+i}$, usando un GAM.
   La somma di questi costi giornalieri previsti genera la previsione in t del costo medio totale a $t + 30$ giorni: $\hat{\mu}_{(30)|t} = \sum_{i=1}^{30} \hat{\mu}_{t+i}$.

2. calcolo di $q_{\alpha}^{\hat{\varepsilon}_{(30)|t}}$, il quantile di livello α della distribuzione degli errori a 30 giorni relativi alla previsione della media condizionata.
   A tal scopo, si assume che i residui del modello sulla media condizionata ε_t seguano un processo GARCH(1, 1), quindi:

   - utilizzando il modello GARCH stimato sui residui del GAM fino al tempo t, si simulano 1000 traiettorie $\varepsilon_{t+1}^{(j)}, \ldots, \varepsilon_{t+30}^{(j)}$, $j = 1, \ldots, 1000$, secondo la procedura di Barone Adesi descritta in sez.4.2.1. Ciascuna traiettoria rappresenta una sequenza di costi che non rientra nel modello di calcolo della media condizionata, essendo stato generato in seguito;
   - per ogni traiettoria, si considera la somma dei valori simulati a 30 giorni: $\hat{\varepsilon}_{(30)|t}^{(j)} = \sum_{i=1}^{30} \hat{\varepsilon}_{t+i}^{(j)}$. Ciò permette di disporre delle singole realizzazioni dell'errore di previsione a 30 giorni, per $j = 1, \ldots, 1000$;
   - i 1000 valori simulati di $\hat{\varepsilon}_{(30)|t}^{(j)}$ ci permettono di calcolare il quantile di livello α della loro distribuzione, $q_{\alpha}^{\hat{\varepsilon}_{(30)|t}}$.

Per il calcolo del $CaR_{t,30,\alpha}$ il modello scelto è definito con lo stesso approccio in due fasi scelto per il $CaR_{t,1,\alpha}$. Il modello per la media condizionata con la migliore performance nell'insieme di validazione è lo stesso modello usato per il $CaR_{t,1,0.10}$ e definito in 4.1, a meno della variabile sui vincoli a rete integra vri_t:

$$\begin{aligned}\mu_{t+i} &= \beta_0 + f_1(T_t) + f_2(DY_t) + f_3(DW_t) + bank_t + f_4(c_{t-1}) + f_5(c_{t-7}) + \\ &\quad + f_6(Hydro_t) + f_7(Wind_t) + f_8(aFRR_t) + f_9(RR_t) \end{aligned} \tag{4.6}$$

Il GAM così definito viene applicato sui costi giornalieri, stimando $\hat{\mu_{t+i}}$, i quali verranno poi sommati sui 30 giorni, stimando $\hat{\mu}_{(30)|t}$.

I residui $\hat{\varepsilon}_t$ del modello per la media condizionata sono quelli su cui viene stimato il modello GARCH, ipotizzando una struttura del tipo $\hat{\varepsilon}_t = \nu_t \sigma_t$, e fornendo la seguente stima della varianza condizionata σ_t

$$\hat{\sigma}_t^2 = \hat{\omega} + \hat{\alpha}_1 \hat{\varepsilon}_{t-1}^2 + \hat{\beta}_1 \hat{\sigma}_{t-1}^2$$

dove $\hat{\omega} = 1678.59 \cdot (10)^8$, $\hat{\alpha_1} = 0.1203$, $\hat{\beta}_1 = 0.7964$.

Metodo della simulazione storica filtrata

Il metodo della simulazione storica filtrata, in inglese *Filtered Historical Simulation* (FHS) (Barone Adesi, 2015), fu originariamente introdotto per il calcolo computazionale del VaR ad $h > 1$ giorni in ambito finanziario.

In origine, il calcolo del VaR si basava sulla stima delle matrici di varianza-covarianza dei rendimenti del portafoglio, che però fu mostrato essere necessaria solo nel calcolo delle posizioni ottimali. Nell'ambito del calcolo del rischio questo approccio si rivela inefficiente (Barone Adesi & Giannopoulos, 1996). Le assunzioni su cui si basa questa modellazione sono forti e quasi mai attese: è richiesta la linearità del portafoglio, e quindi che possano esservi relazioni solo lineari tra gli asset, e presuppone la conoscenza della distribuzione di probabilità dei rendimenti. Da evidenze empiriche tuttavia i rendimenti possono essere incorrelati, ma non indipendenti, ed è difficile definire correttamente la loro distribuzione effettiva.

Barone Adesi sviluppò un metodo che consentì di distaccarsi da tali ipotesi, utilizzando tecniche statistiche già esistenti come i metodi Monte Carlo, nel caso parametrico, o il

ricampionamento *bootstrap*, nel caso non parametrico (o semiparametrico). Il suo lavoro del 1996 definì le basi di quello che ancora oggi è uno dei metodi più utilizzati per la stima del VaR ad h giorni: l'approccio delle simulazioni storiche filtrate. Tale procedura impiega i modelli GARCH per filtrare i dati e rendere i residui IID, andando poi a utilizzare tali residui per generare scenari con tecniche *bootstrap*, tenendo conto sia della non-normalità che della loro eteroschedasticità.

Tale metodo, esteso al nostro contesto di calcolo del CaR, consente di stimare la distribuzione dei costi a 30 giorni senza ipotizzare l'indipendenza, ma una struttura di dipendenza di tipo GARCH. La procedura di FHS è la seguente:

1. ipotizzando che la serie dei residui ε_t, $t = 1, \ldots, n$ segua una processo GARCH$(1,1)$, si stimano i parametri del GARCH$(1,1)$ utilizzando questi dati;

2. in $t = n$ si simulano gli ε_t per $n+1, \ldots, n+h$ volte, secondo lo schema:

$$\begin{aligned}
\sigma^2_{n+1} &= \hat{\omega} + \hat{\alpha}\varepsilon^2_n + \hat{\beta}\sigma^2_n \Rightarrow \hat{\varepsilon}_{n+1} = \nu^*_1 \sigma_{n+1} \\
\sigma^2_{n+2} &= \hat{\omega} + \hat{\alpha}\hat{\varepsilon}^2_{n+1} + \hat{\beta}\sigma^2_{n+1} \Rightarrow \hat{\varepsilon}_{n+2} = \nu^*_2 \sigma_{n+2} \\
&\ldots \\
\sigma^2_{n+h} &= \hat{\omega} + \hat{\alpha}\hat{\varepsilon}^2_{n+h-1} + \hat{\beta}\sigma^2_{n+h-1} \Rightarrow \hat{\varepsilon}_{n+h} = \nu^*_h \sigma_{n+h}
\end{aligned}$$

   dove ν^*_i sono variabili campionate indipendentemente da una distribuzione parametrica o da quella empirica di $\hat{\nu}_t$;

3. nella logica finanziaria in cui è stato applicato per la prima volta questa procedura, $\hat{\varepsilon}^{(j)}_{(30)|t} = \sum_{i=1}^{30} \hat{\varepsilon}^{(j)}_{t+i}$ rappresentava la previsione del rendimento di un portafoglio ad $h = 30$ giorni, mentre nel nostro caso, coincide con una realizzazione della somma degli errori giornalieri fatti nella previsione in $t = n$ 30 giorni in avanti;

4. iterando M volte la procedura, si giunge ad avere M valori simulati $\hat{\varepsilon}^{(j)}_{(30)|t}$, dalla cui distribuzione poter derivare il quantile α-esimo degli errori $h = 30$ giorni in avanti.

4.2.2 Verifica della capacità predittiva

Il $CaR_{t,30,0.10}$ per $t = 1, \ldots, n$ calcolato nel periodo di verifica, dal 1/10/2020 al 30/9/2021, è mostrato in Figura 4.10, in rosso, mentre la serie dei costi mensili è riportata in nero. In verde è riportata la stima della media condizionata realizzata con il GAM, sulla quale viene costruito il modello del CaR a 30 giorni.

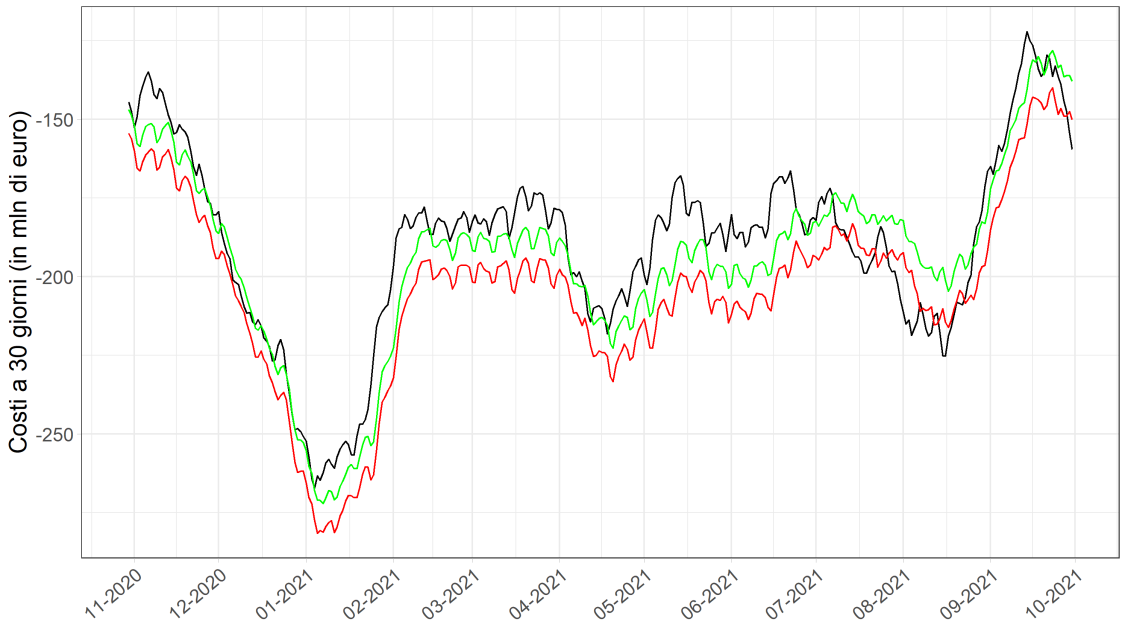


Figura 4.10: Serie storica dei costi a 30 giorni, in milioni di euro, per il periodo di verifica dal 1/10/2020 al 30/9/2021 (in nero), applicazione del modello sulla media condizionata (in verde) e del modello di CaR a 30 giorni (in rosso).

Si nota che la stima del $CaR_{t,30,0.10}$ sta quasi sempre sotto la serie dei costi, e della media prevista a 30 giorni: in certi periodi si discosta poco, come nel periodo fino a inizio 2021, dopodiché inizia a discostarsi leggermente, e continua così per buona parte del periodo di verifica, fino a luglio 2021. È a partire da questo momento, e fino a inizio settembre 2021, che il CaR sottostima i costi reali, invertendo la propria posizione e andandosi a collocare sopra la serie dei costi, per poi tornare a dare una misura del rischio sensata. Il CaR sembra essere un buon modello, per come è stato stimato sulle variabili a disposizione, che hanno sempre consentito di valutare il rischio correttamente. Il comportamento anomalo nel periodo sopra citato si ritiene possa essere dovuto a interventi esogeni, non modellabili in quanto fuori dal nostro controllo. Inoltre, come evidente dal grafico degli sforamenti, in

Figura 4.11 sempre in questo periodo si osservano sforamenti raggruppati, segnale di un evento non spiegabile con le informazioni a nostra disposizione.

In tabella 4.5, sono riportati i risultati legati alla modellazione GAM-GARCH per il $CaR_{t,30,\alpha}$. Tale modello consente di avere una livello osservato α_{oss} del 10.7%, praticamente uguale al livello nominale del 10%. Questo risultato viene valutato con i test di Kupiec, Christoffersen e del quantile dinamico. Il test di Kupiec della copertura non condizionata valida il modello, con un p-value pari a 0.665, che porta a non rifiutare l'ipotesi nulla di copertura osservata uguale a quella attesa. Il test di Christoffersen di copertura condizionata invece, rivela la presenza di sforamenti non indipendenti. Infatti, il p-value quasi 0 ad esso associato conduce al rifiuto dell'ipotesi nulla di indipendenza delle violazioni.

Modello	α_{oss}	**Kup**	**Chris**	**DQ**
GAM-GARCH	0.107	0.665	< 0.001	0.825

Tabella 4.5: Livello osservato α_{oss} rispetto a quella nominale del 10% nell'insieme di verifica (365 osservazioni dal 1/10/2020 al 30/9/2021) e valore del p-value dei test di Kupiec (Kup), Christoffersen (Chris) e del quantile dinamico (DQ), per il modello di CaR a 30 giorni.

Il motivo appare evidente, dal grafico di Figura 4.11, riferito agli sforamenti del CaR nel periodo di verifica. Il grafico riporta sull'asse delle ordinate la variabile $I_{t,\alpha}$, che assume valore 1 se al tempo t si è osservata una violazione del CaR, 0 altrimenti. Gli sforamenti sono concentrati prevalentemente in una finestra di 45 giorni, probabilmente a causa di un evento esogeno. Ricordando la modellazione del CaR a 1 giorno, anche a livello giornaliero questo periodo aveva creato problemi di copertura.

Il test del quantile dinamico non esclude la presenza di una relazione di tipo lineare tra la variabile relativa agli sforamenti e i suoi ritardi, nonché c_{t-1} e di $CaR_{t-1,\alpha,30}$. Il modello potrebbe essere migliorato da questo punto di vista grazie a informazioni aggiuntive sulla dinamica dei costi.

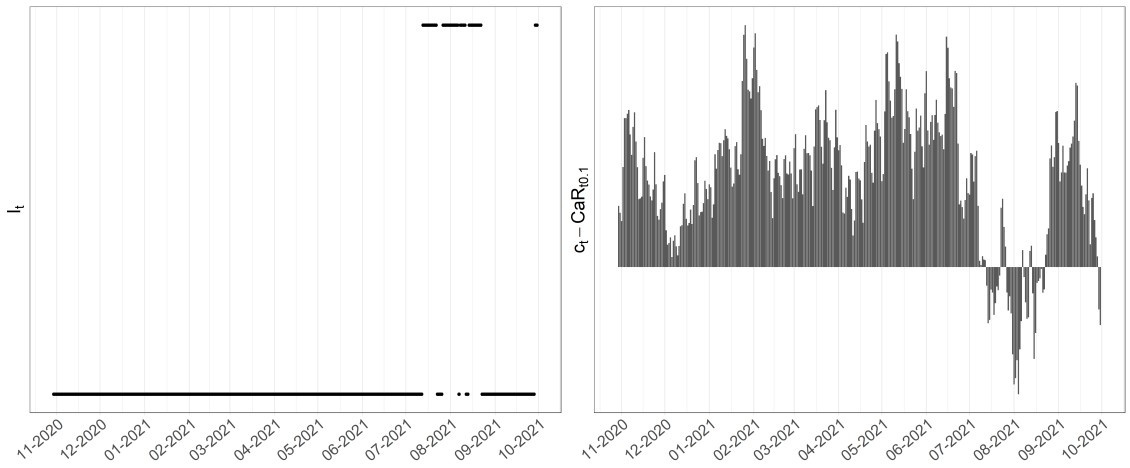


Figura 4.11: Sequenza (a destra) e ampiezza (a sinistra) degli sforamenti del $CaR_{t,30,0.10}$ osservati nel periodo di verifica dal 1/10/2020 al 30/9/2021.

Infine, in tabella 4.6, sono riportate alcune statistiche descrittive relative al $CaR_{t,30,0.10}$.

Media $\mathbf{CaR_{t,30,0.10}}$	**Mediana** $\mathbf{CaR_{t,30,0.10}}$	$\mathbf{Q_{0.25}}$ $\mathbf{CaR_{t,30,0.10}}$	$\mathbf{Q_{0.75}}$ $\mathbf{CaR_{t,30,0.10}}$
-203.999	-201.303	-214.542	-192.276

Tabella 4.6: Alcune statistiche riassuntive della distribuzione del $CaR_{t,30,0.10}$: media, mediana, quantile al 25% e al 75%, espresse in milioni di euro.

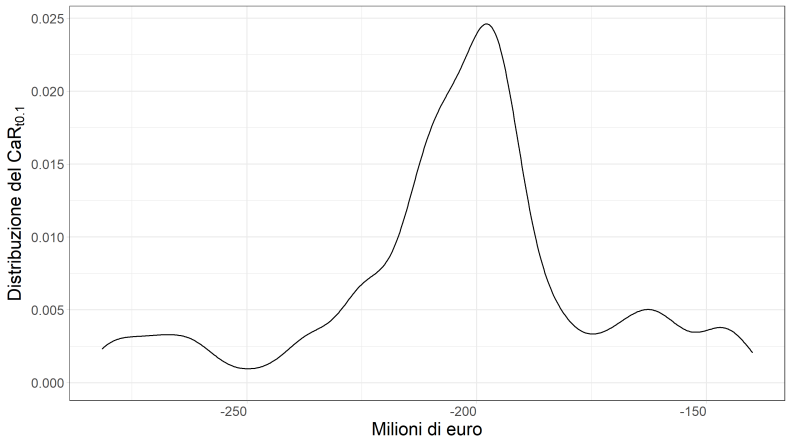


Figura 4.12: Distribuzione del $CaR_{t,30,0.10}$ calcolato con modello GAM-GARCH.

Conclusioni

I costi di approvvigionamento di riserva di energia rappresentano un fenomeno complesso da modellare perché influenzato da molteplici variabili, in modo soprattutto non lineare, e soggetto all'azione di regolamentazioni del mercato elettrico.

Nel primo capitolo sono state descritte le complicate dinamiche di tale mercato, proprio per fornire un inquadramento della problematica affrontata in seguito. Dopo la descrizione della filiera elettrica, sono stati presentati gli innumerevoli protagonisti del mercato e le evoluzioni a cui hanno assistito. Sono stati esaminate a fondo le caratteristiche del Mercato del Giorno Prima e del Mercato Infragiornaliero, che costituiscono il Mercato dell'Energia, poiché è dai programmi in esito a tali mercati che dipendono le azioni di Terna nel Mercato per il Servizio di Dispacciamento. Infatti, in questa piattaforma, Terna si approvvigiona dei necessari servizi per garantire che i programmi del Mercato dell'Energia siano attuabili entro limiti di sicurezza. Tali servizi prendono il nome di servizi ancillari, disposti con il fine ultimo del bilanciamento in tempo reale. In particolare, è nel Mercato per il Servizio di Dispacciamento che Terna si occupa di predisporre la riserva di energia, servizio fondamentale per poter intervenire a ripristino degli squilibri di rete. E per fare ciò sostiene dei costi, la variabile risposta delle nostre modellazioni.

Accanto alla serie dei costi giornalieri sostenuti in MSD per l'approvvigionamento di riserva, Terna ha fornito altre variabili, tra cui, i consuntivi di fabbisogno di energia, variabili relative all'attivazione di riserva e all'energia proveniente da fonti rinnovabili, tema al centro dell'evoluzione del mercato elettrico. Peculiarità di questo mercato è che si basa sulla realizzabilità fisica dei programmi, per questo sono state considerate variabili relative ai vincoli a rete integra. Sono state indagate anche le dinamiche temporali dei costi, la relazione con le serie storiche dei prezzi del gas e dei certificati verdi, poiché fortemente legate all'andamento del prezzo dell'energia. Ne è emerso che i costi di approvvigionamento sono influenzati in modo non lineare da tutte queste variabili, con dinamiche cicliche per le variabili di calendario.

Alla definizione dei costi che Terna sostiene in MSD concorrono diverse variabili, e

la loro dinamica influenza il rischio di incorrere in costi lontani dal loro valore atteso. Questo rischio per Terna va controllato, perché porta a definire la quantità di denaro da predisporre per l'acquisto di riserva nel giorno successivo o entro un mese.

La complessità delle relazioni tra le variabili è stata modellata in modo additivo non parametrico, approccio scelto per la flessibilità che lo caratterizza. La misura del rischio è stata fornita da modelli di CaR (Costo-a-Rischio), un'applicazione del VaR (Valore-a-Rischio) alla variabile dei costi. Tale misura di rischio è stata applicata sui residui di un modello GAM sulla media condizionata, scelto per spiegare l'effetto congiunto delle variabili sui costi medi giornalieri. Sono stati messi a confronto diversi modelli di CaR, basati sia su distribuzione marginale, che condizionata. Nel primo caso, con l'assunzione di omoschedasticità dei residui, è stato stimato il modello GAM-K, basato sulla distribuzione *kernel*. Il CaR basato su distribuzione condizionata è, invece, in grado di cogliere una varianza dei dati non costante nel tempo. Sono stati stimati un modello GAM-GARCH, basato su un GARCH(1, 1), un modello GAM-QR, basato sulla regressione quantilica, e una sua specificazione introdotta da Engle e Manganelli, il CAViaR, con un GAM-CAViaR.
Tutti questi modelli vanno di fatto a concentrarsi sulla modellazione del quantile di livello 10% della distribuzione dei residui, applicandovi la misura di rischio. I residui sono quelli del modello GAM sulla media; di fatto sono ciò che rappresenta un rischio per Terna: la deviazione dal valore atteso. L'interesse si è concentrando su quel valore della loro distribuzione che corrisponde al quantile al 10%, calcolando di fatto il massimo costo potenziale che Terna deve sostenere con probabilità del 10% entro un giorno ed entro un mese.

Pertanto, oltre a questi modelli di misura del rischio, si è considerato di modellare direttamente il quantile condizionato della distribuzione dei costi con un QGAM. In particolare è stato scelto di adottare un QGAM che fornisse una copertura vicina a quella nominale del 10%, nominato QGAM_{cov}. I risultati sono stati confrontati con procedure di *backtesting*: alcuni modelli si sono mostrati troppo prudenti e conservativi nell'insieme di stima, venendo ritenuti adeguati dai test di validazione, per poi nell'insieme di verifica, vedersi rifiutati per una copertura troppo bassa. Questo risultato si deve soprattutto al fatto che i modelli sono così flessibili che si adattano molto alla variabilità dei dati nell'insieme di stima, per poi essere però confutati dai dati dell'insieme di verifica.
Il modello migliore per la sua adattabilità e capacità predittiva del rischio a 1 giorno è il GAM-GARCH, che per la sua struttura più rigida rispetto a modelli come la regressione quantilica, riesce a contenere meglio il rischio di sovradattamento.

La procedura per il calcolo del rischio a 1 giorno ha portato a risultati tutto sommato

soddisfacenti. Per il rischio a 30 giorni è stato quindi utilizzato un GAM-GARCH, il modello scelto per il rischio a 1 giorno, estendendo a un orizzonte di 30 giorni la modellazione su base giornaliera, con il metodo della simulazione storica filtrata di Barone Adesi. I risultati ottenuti sono abbastanza buoni: la copertura del modello di CaR è pari a quella attesa, ma nell'osservazione della dinamica degli sforamenti, si rivela la presenza di un accumulo in una finestra di più o meno 45 giorni. Tale fenomeno potrebbe essere dovuto alla presenza di cambiamenti strutturali del mercato. Tale ipotesi è ritenuta plausibile a causa dell'influenza decisiva su questi mercati, come mostrato, di normative e regolamentazioni ferree, in grado di modificare la struttura dei fenomeni interni.

Bibliografia

Barone Adesi, G. (2015). *Simulating security returns: A filtered historical simulation approach*, Palgrave Macmillan.

Bellman, R. E. (1961). *Curse of dimensionality. Adaptive control processes: a guided tour*, Princeton University Press.

Biau, G. and Patra, B. (2010). Sequential quantile prediction of time series, *IEEE Transactions on Information Theory* **57**(3): 1664–1674.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity, *Journal of Econometrics* **31**(3): 307–327.

Buchinsky, M. and Hahn, J. (1998). An alternative estimator for the censored quantile regression model, *Econometrica* **66**(3): 653–672.

Caprabianca, M., Falvo, M. C., Papi, L., Promutico, L., Rossetti, V. and Quaglia, F. (2020). Replacement reserve for the italian power system and electricity market, *Energies* **13**(11).

Christoffersen, P. F. (1998). Evaluating interval forecasts, *International Economic Review* **39**(4): 841–862.

Cleveland, W. S., Grosse, E. and Shyu, W. M. (1992). *Statistical Models in S (1st Edition)*, Routledge.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation, *Econometrica* **50**(4): 987–1007.

Engle, R. F. and Manganelli, S. (2004). Caviar: Conditional autoregressive value at risk by regression quantiles, *Journal of Business & Economic Statistics* **22**(4): 367–381.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R. and Goude, Y. (2021a). Fast calibrated additive quantile regression, *Journal of the American Statistical Association* **116**(535): 1402–1412.

Fasiolo, M., Wood, S. N., Zaffran, M., Nedellec, R. and Goude, Y. (2021b). qgam: Bayesian nonparametric quantile regression modeling in r, *Journal of Statistical Software* .

Glachant, J.-M., Joskow, P. L. and Pollitt, M. G. (2021). *Handbook on Electricity Markets*, Edward Elgar Publishing.

Graf, C., Quaglia, F. and Wolak, F. A. (2020). Simplified electricity market models with significant intermittent renewable capacity: Evidence from italy, *Working Paper 27262*, National Bureau of Economic Research.

Hastie, T. (2015). gam: Generalized additive models. r package version 1.12.

Hastie, T. J., Tibshirani and Buja, A. (2021). *Flexible Discriminant Analysis by OptimalScoring*, Vol. 102.

Hastie, T. J., Tibshirani, R. and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd Edition*, Springer New York, NY.

Hastie, T. J., Tibshirani, R., James, G. and Witten, D. (2021). *An Introduction to Statistical Learning (2nd Edition) with Applications in R*, Vol. 102.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models, *Statistical Science* **1**: 297–318.

Koenker, R. (2004a). Quantile regression for longitudinal data, *Journal of Multivariate Analysis* **91**(1): 74–89.

Koenker, R. (2004b). quantreg: An r package for quantile regression and related methods.

Koenker, R. (2005). *Quantile Regression*, Cambridge University Press.

Koenker, R. and Bassett, G. (1978). Regression quantiles, *Econometrica* **46**(1): 33–50.

Koenker, R., Chernozhukov, V., He, X. and Peng, L. (2017). *Handbook of Quantile Regression (1st Edition)*, Chapman and Hall/CRC.

Kupiec, P. H. (1995). Techniques for verifying the accuracy of risk measurement models, *The Journal of Derivatives* **3**.

Pace, L. and Salvan, A. (1996). *Introduzione alla statistica: Inferenza, verosimiglianza, modelli.-2001.-xvi, 422 p*, Cedam.

Pace, L. and Salvan, A. (2001). *Introduzione alla statistica 2: inferenza, verosimiglianza, modelli*, CEDAM, Padova.

Parsons, L., Haque, E. and Liu, H. (2004). Subspace clustering for high dimensional data: A review, *SIGKDD Explor. Newsl.* **6**(1): 90–105.

Terna (2020). Mercato elettrici: Struttura attuale, evoluzione attesa e sfide da affrontare. *Slides*.

Wood, S. N. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **62**(2): 413–428.

Wood, S. N. (2015). mgcv: Mixed gam computation vehicle with automatic smoothness estimation. r package version 1.8-40.

Wood, S. N. (2017). *Generalized additive models: An introduction with R (2nd Edition)*.

Wood, S. N. and Augustin, N. H. (2002). Gams with integrated model selection using penalized regression splines and applications to environmental modelling, *Ecological Modelling* **157**(2): 157–177.

Sitografia

Arera (2021). Relazione annuale 2020: Stato dei servizi, *Technical Report 1*, Arera. https://www.arera.it/allegati/relaz_ann/21/RA21_volume_1.pdf, Accessed: 20 giugno 2022.

Arera (2022). I primi venti importatori in italia nel 2020. https://www.arera.it/it/dati/gm51.htm, Accessed: 4 giugno 2022.

Bongioanni, M. (2022). Da eolico e solare il 10% dell'energia mondiale. ma cresce il carbone", *Lifegate* . https://www.lifegate.it/eolico-solare-rinnovabili-global-electricity-review.

Commission Regulation (EU) No 543/2013 of 14 June 2013 (2013). (Article: 2, 19). https://lexparency.org/eu/32013R0543/, Accessed: 20 giugno 2022.

Dataenergia (2014). Le variabili che influenzano il prezzo dell'energia sul mercato elettrico. https://dataenergia.altervista.org/portale/?q=variabili_che_influenzano_prezzo_energia_mercato_elettrico, Accessed: 4 giugno 2022.

De Rooij, D. (2022). Base load and peak load: understanding both concepts, *Technical report*, Sinovoltaics. https://sinovoltaics.com/learning-center/basics/base-load-peak-load/, Accessed: 10 giugno 2022.

Decreto Legislativo 16 marzo 1999, n. 79 (1999). *Attuazione della direttiva 96/92/CE recante norme comuni per il mercato interno dell'energia elettrica.* . https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:1999-03-16;79, Accessed: 10 giugno 2022.

Decreto Legislativo 29 dicembre 2003, n. 387 (2003). *Attuazione della direttiva 2001/77/CE relativa alla promozione dell'energia elettrica prodotta da fonti energe-*

tiche rinnovabili nel mercato interno dell'elettricità. . https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:decreto.legislativo:2003-12-29;387, Accessed: 10 giugno 2022.

Delibera 08 giugno 2017 n. 419/2017/R/eel (2017). *Valorizzazione transitoria degli sbilanciamenti effettivi nelle more della definizione della disciplina di regime basata su prezzi nodali*, DMEA. https://www.arera.it/allegati/docs/17/419-17.pdf, Accessed: 20 giugno 2022.

Delibera 19 marzo 2019 n. 103/2019/R/eel (2017). *Ulteriori disposizioni in merito alla suddivisione della rete rilevante in zone, in esito al processo di revisione svolto ai sensi del regolamento (ue) 2015/1222 (cacm)*, REU. https://www.arera.it/allegati/docs/19/103-19.pdf, Accessed: 20 giugno 2022.

Direttiva 96/92/CE del Parlamento Europeo e del Consiglio del 19 dicembre 2016 (2016). *Ministero della Transizione Ecologica* . https://www.mite.gov.it/sites/default/files/Direttiva_96_92_CE.pdf, Accessed: 15 giugno 2022.

Disposizione tecnica di funzionamento n. 01 rev. 05 MTE (2016). *Periodo di negoziazione e modalità di regolazione dei contratti nel mte*, GME. https://www.mercatoelettrico.org/it/MenuBiblioteca/Documenti/20160908DTF1rev5.pdf, Accessed: 20 giugno 2022.

Documento per la Consultazione n. 325/2021/R/eel (2021). *Orientamenti per la definizione di un sistema di incentivazione ai fini della riduzione dei costi di dispacciamento*, Arera. https://www.arera.it/allegati/docs/21/325-21.pdf, Accessed: 20 giugno 2022.

EnelGreenPower (2021). Centrale solare. https://www.enelgreenpower.com/it/learning-hub/energie-rinnovabili/energia-solare/centrale-solare, Accessed: 4 giugno 2022.

Entso-E (2022). Terre project. https://www.entsoe.eu/network_codes/eb/terre/, Accessed: 4 giugno 2022.

E.ONEnergia (2020). L'energia geotermica: cos'è, vantaggi e svantaggi. https://www.eon-energia.com/magazine/energia-domestica/che-cose-lenergia-geotermica.html, Accessed: 7 giugno 2022.

EPEXSPOT (2021). European market coupling. https://www.epexspot.com/en/marketcoupling, Accessed: 4 giugno 2022.

Ferraino, G. (2022). Dalle centrali a carbone viene il 4,9% dell'energia elettrica italiana. ecco dove sono e quali sono le sfide, *Corriere della Sera* . https://www.corriere.it/economia/consumi/22_gennaio_21/dalle-centrali-carbone-viene-49percento-dell-energia-elettrica-italiana-ecco-dove-sono-quali-sono-sfide-986e82b6-7a84-11ec-bb07-072210d17db2.shtml, Accessed: 27 giugno 2022.

GME (2022). Mercato elettrico a pronti (mpe) - mgp, mi, mpeg, msd. , Accessed: 4 giugno 2022.

HUPX (2022). Core flow-based market coupling project: announcement on the progress of joint integration testing and new go-live date. https://hupx.hu/en/articles/core-flow-based-market-coupling-project-announcement-on-the-progress-of-joint-integration-testing-and-new-go-live-date/175, Accessed: 4 giugno 2022.

Legge del 23 agosto 2004, n.239 (2004). *Riordino del settore energetico, nonchè delega al Governo per il riassetto delle disposizioni vigenti in materia di energia* . https://www.normattiva.it/uri-res/N2Ls?urn:nir:stato:legge:2004-08-23;239, Accessed: 15 giugno 2022.

Marchisio, L., Genoese, F. and Raffo, F. (2022). L'apertura delle risorse distribuite al mercato dei servizi: quale bilancio?, *Technical report*, Arera. https://download.terna.it/terna/0000/1224/89.PDF, Accessed: 20 giugno 2022.

NEMOCommittee (2019). Euphemia public description: Single price coupling algorithm, *Technical report*, All NEMO Committee. https://www.mercatoelettrico.org/it/MenuBiblioteca/Documenti/20181212EuphemiaPublicDescription.pdf.

NEMOCommittee (2022). Sdac. https://www.nemo-committee.eu/sdac, Accessed: 4 giugno 2022.

SelectraItalia (2022). Filiera dell'energia elettrica: dalla produzione al contatore. https://luce-gas.it/guida/mercato/filiera-elettrica, Accessed: 10 giugno 2022.

Terna (2020). Nuova configurazione dei perimetri di aggregazione. https://www.terna.it/it/sistema-elettrico/pubblicazioni/news-operatori/dettaglio/perimetri-UVAM-2021, Accessed: 4 giugno 2022.

Terna (2021a). Come funziona il sistema elettrico. https://www.terna.it/it/sistema-elettrico/ruolo-terna/come-funziona-sistema-elettrico, Accessed: 4 giugno 2022.

Terna (2021b). Piano di sviluppo 2021, *Technical report*, Terna SpA. https://download.terna.it/terna/Piano_Sviluppo_2021_8d94126f94dc233.pdf, Accessed: 20 giugno 2022.

Terna (2021c). Preparazione del pds e consultazioni. https://www.terna.it/it/sistema-elettrico/rete/piano-sviluppo-rete/preparazione-pds-consultazioni, Accessed: 10 giugno 2022.

Terna (2021d). Rapporto adeguatezza italia 2021, *Technical report*, Terna SpA. https://download.terna.it/terna/Terna_Rapporto_Adeguatezza_Italia_2021_8d9a51d27ad741c.pdf, Accessed: 10 giugno 2022.

Terna (2021e). Relazione finanziaria annuale - rapporto integrato 2021, *Technical report*, Terna SpA. https://download.terna.it/terna/Terna_Rapporto_Integrato_2021_8da18aae2568772.pdf, Accessed: 7 giugno 2022.

Terna (2021f). Sa.pe.i il cavo dei record. https://www.terna.it/it/progetti-territorio/sapei, Accessed: 4 giugno 2022.

Terna (2022a). Le nuove zone del mercato elettrico: quello che c'è da sapere. https://lightbox.terna.it/it/riorganizzazione-zone-mercato-elettrico, Accessed: 10 giugno 2022.

Terna (2022b). M come mercato elettrico. https://lightbox.terna.it/it/m-mercato-elettrico, Accessed: 4 giugno 2022.

Terna (2022c). Metering. https://www.terna.it/it/sistema-elettrico/dispacciamento/metering, Accessed: 4 giugno 2022.

Terna (2022d). Nel 2021 deciso recupero dei consumi elettrici +5,6% rispetto al 2020, tornati sui valori del 2019. https://www.terna.it/it/media/comunicati-stampa/dettaglio/consumi-elettrici-2021, Accessed: 15 giugno 2022.

Terna (2022e). Single intraday coupling. https://lightbox.terna.it/it/dispacciamento-single-intraday-coupling, Accessed: 4 giugno 2022.

Terna (2022f). Tyrrhenian link: il doppio collegamento sottomarino tra sicilia, sardegna e penisola. https://www.terna.it/it/progetti-territorio/progetti-incontri-territorio/Tyrrhenian-link, Accessed: 30 giugno 2022.

Testo Integrato della Disciplina del Mercato Elettrico - Capo 1 BIS (2021). *Technical report*, Gestore dei Mercati Energetici. https://www.mercatoelettrico.org/it/MenuBiblioteca/Documenti/20210921_Testo_Integrato_ME_completo.pdf, Accessed: 7 giugno 2022.

WGAS, E.-E. (2021). Survey on ancillary services procurement, balancing market design 2021. https://eepublicdownloads.azureedge.net/clean-documents/mc-documents/balancing_ancillary/2021/AS_Survey_2020_Results_Updated.pdf, Accessed: 4 giugno 2022.