Università Degli Studi di Padova
Dipartimento di Scienze Statistiche

Laboratorio di Statistica con le aziende
Report finale

Lorenzo Cifelli, Gabriele Massaro,
Alessio Piraccini, Lorenzo Zilioli

9 settembre 2022

Università degli Studi di Padova

Indice

Sommario

La *challenge* richiede di analizzare il mercato del carburante in Italia, facendo riferimento ai dati del 2021. Si richiede nello specifico di definire una metrica basata sul prezzo, che sia depurata da andamenti macroeconomici e tendenziali. Su questa va impostata un'analisi che permetta di identificare le caratteristiche fisse che determinano gli scostamenti dei prezzi dei singoli distributori dalla media nazionale. Per rispondere alle richieste, viene impostata una modellazione su due stadi: prima si modella il *trend* medio dei prezzi sulla base di serie storiche di indici macroeconomici di interesse; in seguito i residui, che risultano depurati dall'andamento nazionale, vengono usati come variabile risposta nella seconda fase di modellazione, volta allo studio dei fattori fissi degli impianti. In base ai risultati emersi dai modelli e a ricerche concernenti lo stato del mercato del carburante in Italia, si riesce a fornire una *price strategy* per un nuovo *player* che desidera entrare nel mercato, come da domanda di ricerca.

1 Introduzione

1.1 La *challenge*

La *challenge* proposta da *Deloitte* concerne l'analisi dei dati forniti dal Ministero dello Sviluppo Economico, riguardanti i prezzi del carburante nelle varie stazioni di rifornimento sul territorio italiano nell'anno 2021.
La finalità del lavoro è quella di analizzare l'andamento del prezzo, al netto di fattori macroeconomici di fondo, cercando di identificare e caratterizzare, se presenti, quei fattori che contribuiscono maggiormente alla sua definizione. Si vuole in sostanza studiare il mercato di riferimento, individuando le caratteristiche intrinseche dei singoli distributori che permettono una maggior flessibilità nella definizione del prezzo, con la doppia finalità di studiare il comportamento dei *competitors* e di fornire indicazioni utili allo sviluppo di una *price strategy* per l'entrata nel mercato.

1.2 Approccio di analisi

L'approccio adottato per l'analisi prevede una modellazione su due stadi. Al primo stadio si modella globalmente l'andamento del prezzo nazionale nel tempo, utilizzando serie storiche relative a fattori macroeconomici ritenuti rilevanti. Successivamente si utilizzano i residui di questa prima modellazione come metrica di interesse, che così facendo risulta depurata dall'andamento globale del prezzo pur mantenendo l'informazione relativa all'eterogeneità dei distributori. Il secondo stadio di modellazione si focalizza quindi sulla previsione di tale metrica sulla base di variabili attinenti alle caratteristiche fisse dei distributori, per individuare, come richiesto, i fattori determinanti nella definizione del prezzo al netto dell'andamento nazionale.

2 I Dati

2.1 Dati di riferimento per la *challenge*

I dati vengono forniti dal Ministero dello Sviluppo Economico e possono essere reperiti al seguente indirizzo: `https://www.mise.gov.it/index.php/it/open-data/elenco-dataset/2036944-carburanti-archivio-prezzi`. Gli archivi sono composti da due *datasets* differenti, uno riguardante le caratteristiche anagrafiche degli impianti e l'altro riguardante le rilevazioni giornaliere del prezzo del carburante. Vengono scaricati i dati relativi a tutto il 2021.
In particolare le variabili contenute nel *dataset anagrafica* sono:

- ID impianto: codice di identificazione;
- Gestore: ragione sociale dell'impresa;
- Bandiera: insegna del distributore;
- Tipo impianto: tipologia di strada su cui si trova l'impianto;
- Nome impianto;
- Indirizzo;
- Comune;
- Provincia;

- Latitudine: coordinate espresse in gradi decimali;
- Longitudine: coordinate espresse in gradi decimali.

Il *dataset prezzo* contiene invece:

- ID impianto: codice di identificazione;
- Desccarburante: tipologia di carburante (benzina, gasolio, GPL . . .);
- Prezzo: prezzo euro/litro;
- Isself: modalità di servizio (1 = self service, 0 = servito);
- Dtcomu: data e ora.

2.2 Dati esterni

Si individuano dei *datasets* esterni da utilizzare per integrare le informazioni presenti nei dati a disposizione.

Partendo dalle nozioni sul comune dove sono situati gli impianti, si aggiungono l'informazione relativa alla regione e diverse misurazioni quali popolazione residente, zona altimetrica, grado di urbanizzazione e se il comune è isolano o costiero. I dati vengono reperiti da *Istat* nei seguenti archivi:

- Elenco comuni:
  - link archivio : `https://www.istat.it/it/archivio/6789\`;
  - link *dataset*: `https://www.istat.it/storage/codici-unita-amministrative/Elenco-comuni-italiani.xls\`.
- Statistiche comuni:
  - link archivio : `https://www.istat.it/it/archivio/156224\`;
  - link *dataset*: `https://www.istat.it/storage/codici-unita-amministrative/Classificazioni-statistiche-Anni$_$2017-2022.zip\`.

Vengono anche aggiunte informazioni per caratterizzare le regioni italiane, come la percentuale di utilizzo di mezzi pubblici, la percentuale di utilizzo di mezzi a motore per spostamenti lavorativi e la percentuale di famiglie che ritengono l'inquinamento un problema della zona in cui vivono. I dati sono relativi al 2021 e vengono messi a disposizione dalla banca dati *Istat* (`http://dati.istat.it/`).

Si considerano le serie storiche settimanali di quattro variabili macroeconomiche ritenute rilevanti, che verranno utilizzate nella creazione della metrica di interesse per spiegare l'andamento globale dei prezzi. Le serie inserite sono il cambio euro/dollaro ed euro/sterlina, per tenere conto della variazione di potere d'acquisto della moneta utilizzata dai fornitori per il reperimento della materia prima, e due ETF relativi a petrolio e gas naturale quotati alla borsa di Milano (WTI Crud Oil e WisdomTree Natural Gas). Gli ETF (*Exchange Traded Funds*) sono strumenti finanziari che replicano fedelmente l'andamento e il rendimento degli indici di cui sono composti: in questo caso l'andamento dei due indici segue l'andamento di tutto il mercato del petrolio grezzo e del gas naturale. Le serie storiche sono scaricate da *Yahoo Finance* ai seguenti indirizzi:

- ETF petrolio grezzo (CRUD.MI): https://https://it.finance.yahoo.com/quote/CRUD.MI?p=CRUD.MI;
- ETF gas naturale (NGAS.MI): https://it.finance.yahoo.com/quote/NGAS.MI/history?p=NGAS.MI\euro/usd;
- tasso di cambio euro/dollaro (EUR/USD): https://it.finance.yahoo.com/quote/EURS-USD/history?p=EURS-USD\eur/gbp;
- tasso di cambio euro/sterlina (EUR/GBP): https://it.finance.yahoo.com/quote/EURGBP%3DX/history?p=EURGBP%3DX.

Per effettuare un'analisi macroeconomica temporalmente più estesa rispetto ai dati forniti, vengono integrate le serie storiche settimanali per il periodo che va dall'anno 2005 al 2021 delle seguenti quantità di interesse:

- prezzi del carburante in Italia: https://dgsaie.mise.gov.it/open-data;
- prezzo del petrolio grezzo in dollari al barile: https://fred.stlouisfed.org/series/DCOILBRENTEU#0;
- tasso di cambio euro-dollaro https://fred.stlouisfed.org/series/DEXUSEU.

2.3 Materiali utilizzati

Tutte le analisi vengono svolte su macchine locali dei componenti del gruppo. A livello indicativo, si prenda come riferimento una macchina con installato Windows 10 a 64 bit, 8 Gb di RAM e un processore Intel® Core™ i5.

Si utilizza il software statistico *R* per tutte le fasi del progetto, di seguito si riportano le principali librerie impiegate per l'analisi:

- *tidyverse* per la manipolazione dei dati (*dplyr*) e le visualizzazioni (*ggplot2*);
- *mgcv* per i modelli additivi;
- *lme4* per i modelli ad effetti casuali;
- *lightgbm* per il *gradient boosting*.

3 Analisi

3.1 Strategia di analisi

L'obiettivo di questo lavoro è lo studio del mercato del carburante in Italia, commissionato da un ipotetico nuovo *player* che vuole entrare nel mercato. Viene richiesto di partire dal prezzo registrato nelle varie stazioni di riferimento e di creare una metrica basata su di esso ma depurata da andamenti tendenziali e macroeconomici. In questo modo, si possono identificare i fattori e le caratteristiche che forniscono ai diversi distributori una maggiore flessibilità nella definizione del prezzo, con il risultato di comprendere le caratteristiche dei vari *competitors* e di fornire indicazioni in merito alla definizione di una *price strategy*.

Come riportato nelle sezioni precedenti, i dati di partenza sono relativi al 2021 e forniti dal Governo Italiano, viene inoltre incentivato l'utilizzo di *datasets* pubblici esterni per aumentare l'informazione disponibile. L'analisi si compone di diverse fasi che vengono sintetizzate di seguito.

- Operazioni preliminari:
  - Si effettuano delle manipolazioni iniziali sui *datasets* di partenza, che successivamente vengono uniti in un unico insieme di dati, utilizzato per le analisi successive.
  - Si inseriscono nuove covariate che si ritiene possano essere potenzialmente interessanti, partendo da fonti esterne o creandole manualmente.
  - Si procede ad un'imputazione dei valori mancanti per completare le operazioni di preparazione del *dataset*.
- Analisi esplorativa e macroeconomica:
  - Si imposta una prima analisi esplorativa basata sui dati del 2021, per osservare l'andamento temporale del fenomeno e la sua distribuzione spaziale.
  - Si opera un'analisi longitudinale più estesa mettendo a confronto le serie settimanali dal 2005 al 2021 del prezzo del gasolio al netto delle tasse e di quello del petrolio grezzo.
- Definizione della metrica di interesse:
  - In base alle indicazioni emerse dall'analisi longitudinale, viene modellato l'andamento medio del prezzo del carburante utilizzando serie storiche attinenti a fattori macroeconomici di interesse.
  - I residui di tale modellazione risultano depurati da andamenti macroeconomici e tendenziali e vanno a definire la metrica utilizzata nella successiva fase di modellazione.
- Modellazione:
  - Sulla base della metrica definita si imposta una modellazione progressiva, che abbia come punti centrali l'interpretabilità dei risultati e la gestione degli effetti a livello spaziale.
  - Partendo da una semplice modellazione lineare, si passa a modelli ad effetti casuali ed infine a modelli additivi. Allo scopo di confrontare la capacità predittiva dei modelli viene anche adattato un *gradient boosting*.

3.2 Operazioni preliminari

3.2.1 Manipolazioni iniziali e unione dei *dataset*

In primo luogo, si lavora distintamente sui due *dataset*, per poi in seguito unirli nell'insieme di dati definitivo utilizzato per l'analisi.

Per il *dataset anagrafica*, che inizialmente contiene 7 702 024 osservazioni relative a 10 variabili, vengono rimosse le variabili gestore, indirizzo e nome impianto, poiché non ritenute utili ai fini dell'analisi, e successivamente vengono rimossi i valori duplicati. In seguito a queste operazioni l'insieme di dati si compone di 26 394 osservazioni relative a 7 variabili.

Per quanto concerne il *dataset prezzo*, composto inizialmente da 32 182 691 osservazioni relative a 8 variabili, la prima scelta è stata quella di focalizzare l'analisi esclusivamente su un unico tipo di

carburante, ossia il Diesel, avendo notato come fosse la tipologia più presente nell'insieme di dati a disposizione. Inoltre, si decide di mantenere solo le osservazioni relativi alle stazioni *self-service*. Questo perché si considera che, dalla prospettiva di un nuovo *player* che vuole entrare nel mercato, la scelta di inserire il servito dipenda da fattori non di diretto interesse in merito alla definizione di una *price strategy* generale (costo della manodopera, disponibilità economiche, possibilità logistiche, . . .). Inoltre, si valuta che tendenzialmente la differenza di prezzo tra *self service* e servito sia pressapoco costante e non dipenda da caratteristiche fisse del distributore, ma piuttosto dai fattori sopracitati.
Successivamente, si opera una compressione del dato, rilevato per ogni impianto a diverse ore ogni giorno, ottenendo la media settimanale dei prezzi. Questa scelta ha una motivazione prettamente analitica, ma porta a delle conseguenze in termini tecnici e computazionali che la giustificano ulteriormente. Relativamente agli obiettivi dell'analisi, si ritiene infatti che non siano di interesse piccole fluttuazioni delle serie per ogni impianto: nell'ottica di voler studiare il fenomeno nel complesso e al netto del suo andamento nazionale, risulta sensato operare un primo lisciamento delle serie ottenendo le medie settimanali dei prezzi. Tale operazione comporta un vantaggio importante in termini computazionali: la riduzione della dimensione dei dati risultante da questa compressione è tale da permettere di svolgere agilmente su macchine locali tutte le seguenti procedure di manipolazione, visualizzazione e modellazione. In seguito a tali operazioni il *dataset prezzo* risulta composto da 822 483 osservazioni di 4 variabili (idImpianto, settimana, mese e prezzo).

Dopo la preparazione si uniscono i due insiemi di dati sulla base dell'identificativo dell'impianto, ottenendo il *dataset* completo con le informazioni relative all'andamento settimanale dei prezzi per ogni impianto e tutte le caratteristiche anagrafiche degli stessi. L'insieme finale contiene 796 233 osservazioni relative a 10 variabili.

3.2.2 Inserimento covariate di interesse

Dopo aver ottenuto l'insieme di dati completo, si passa ad integrarlo mediante l'aggiunta di diverse covariate, create manualmente o provenienti da fonti esterne, con il fine di inserire un maggior numero di fattori che potrebbero emergere come rilevanti sulla base dell'analisi. Inoltre si manipolano alcune variabili per renderle più interpretabili.

- Si aggiungono le variabili regione e zona (Nord-Est, Nord-Ovest, Centro, Sud, Isole), che verranno usate nell'analisi spaziale;
- vengono create variabili dicotomiche indicanti se la regione è a statuto speciale e se il comune ha più di 200 000 abitanti;
- è stata creata una variabile riguardante il numero di impianti nel raggio di 5km e una dicotomica relativa alla presenza di impianti non autostradali entro 500m, con l'idea di identificare la presenza di diretti concorrenti nelle vicinanze;
- per caratterizzare maggiormente le regioni, si inseriscono tramite fonti esterne informazioni relative al PIL del 2020, alla percentuale di famiglie che valuta l'inquinamento come un problema e alle percentuali di utilizzo di mezzi pubblici e di mezzi a motore per spostamenti lavorativi.
- attraverso fonti esterne sono stati aggiunti dati riguardanti la popolazione residente al 31/12/2020, zona altimetrica, comune litoraneo, comune isolano, zona costiera e grado di urbanizzazione per tutti i comuni presenti;
- si divide la variabile relativa alla popolazione residente in quattro classi e si accorpano i livelli della variabile bandiera, mantenendo i 6 più frequenti insieme alla nuova modalità "altro";

- vengono rimosse alcune unità mal registrate.

In seguito a queste operazioni, il *dataset* risulta essere composto da 784 900 osservazioni relative a 25 variabili.

3.2.3 Imputazione valori mancanti

I dati iniziali relativi al prezzo vengono registrati in momenti e giorni diversi per impianti differenti, di conseguenza dopo l'ottenimento delle medie settimanali si hanno alcuni impianti con osservazioni mancanti relativamente a certe settimane. Pertanto, per completare l'operazione di pulizia e preparazione del *dataset* bisogna attuare un'imputazione di tali valori mancanti.

Si decide di utilizzare un approccio semplice, basandosi sull'informazione relativa ai prezzi settimanali di impianti che possano essere intuitivamente considerati simili a quelli che presentano valori mancanti. Avendo deciso di procedere in questo modo, lavorando direttamente sui livelli si perderebbe la continuità delle serie dei singoli impianti. Si valuta dunque di imputare i valori mancanti alle serie delle differenze prime: un valore mancante viene così calcolato a partire dal livello noto della sua serie di riferimento e dalla differenza prima associata. Nello specifico si procede come segue. A partire dalle serie dei prezzi, si ottengono le serie delle differenze prime. A questo punto si calcolano le mediane per ogni settimana, stratificando l'operazione per regione e bandiera per basarsi su impianti simili, e si utilizzano per imputare i valori mancanti. Successivamente si ritrasformano le serie passando dalle differenze prime ai livelli originali.

Dopo l'operazione di pulizia il *dataset* non contiene valori mancanti e si compone di 1 176 032 osservazioni relative a 25 variabili.

3.3 Analisi esplorativa e macroeconomica

Ottenuto il *dataset* finale, si imposta una breve analisi esplorativa che permetta di indagare l'andamento del prezzo del Diesel a livello nazionale e di fornire indicazioni iniziali in merito a differenze regionali. Successivamente si passa ad un'analisi longitudinale più estesa, che mette a confronto le serie settimanali dal 2005 al 2021 del prezzo del gasolio al netto di tasse e del prezzo del petrolio grezzo al barile. Questo permette di caratterizzare la dipendenza del livello medio del fenomeno da fattori macroeconomici e fornisce indicazioni utili in merito alla successiva creazione della metrica basata sul prezzo e depurata da fattori tendenziali.

3.3.1 Dati del 2021

Esplorando i dati di riferimento per la *challenge*, si parte da una visualizzazione che permetta di farsi una prima idea della distribuzione del fenomeno a livello spaziale.

In Figura 1 si riportano le mappe con i valori medi del prezzo del gasolio nel 2021 stratificato per regione e provincia. Dal grafico relativo al prezzo medio regionale è possibile notare come il valore medio più alto si sia registrato in Trentino Alto-Adige, seguito da Val d'Aosta e Liguria, mentre è evidente come quello più basso sia relativo alla Campania. In generale si osserva come i prezzi relativi al Diesel siano più alti al Nord rispetto che al Sud, dove i valori registrati sono tendenzialmente più bassi della media nazionale. Non vi è omogeneità relativa al prezzo per quanto concerne le regioni a statuto speciale, ossia Sicilia, Sardegna, Val d'Aosta, Trentino Alto-Adige e Friuli Venezia Giulia. Per caratterizzare l'informazione, individuando province che registrano prezzi anomali al punto da condizionare il valore regionale, si è deciso di fornire un ulteriore grafico relativo al prezzo medio annuale

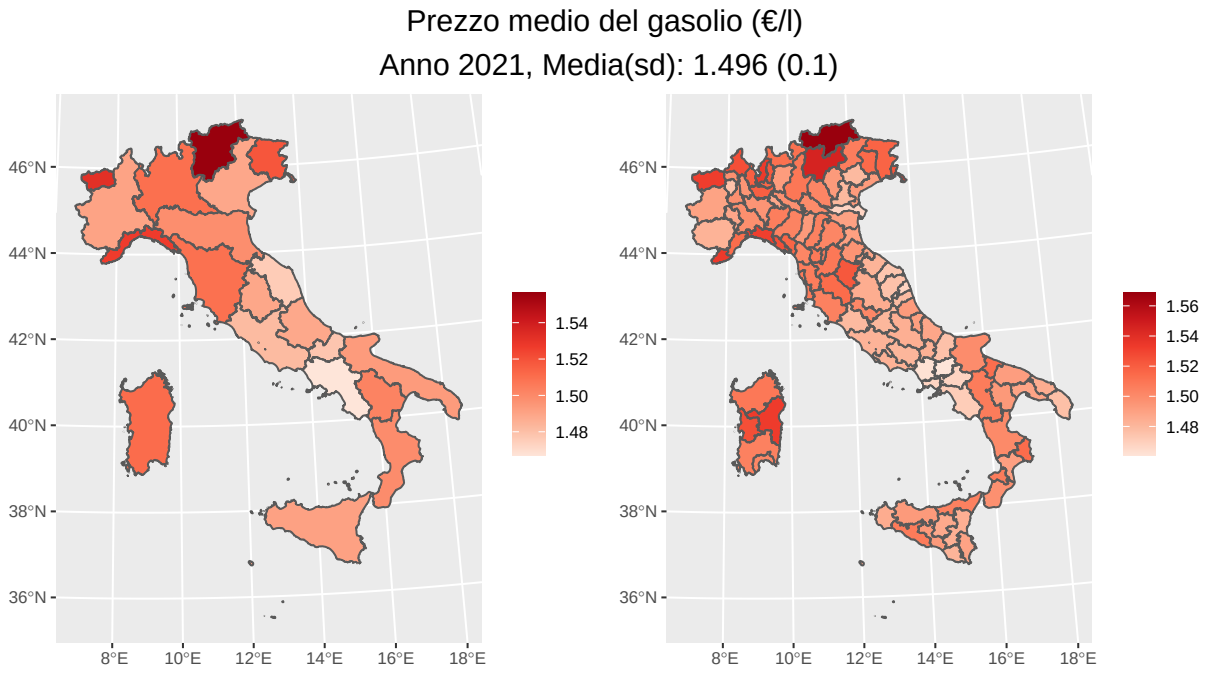

Figura 1: Prezzo medio del carburante per province e regioni.

del Diesel a livello provinciale. Dalla mappa è possibile notare come sia a causa della provincia di Bolzano, associata al valore massimo tra tutte le provincie, che la regione del Trentino Alto-Adige è quella dove si registra un livello medio annuale dei prezzi più alto. Per quanto riguarda le altre regioni, non si nota eccessiva eterogeneità a livello provinciale.

Si passa dunque ad analizzare l'andamento temporale del prezzo. In Figura 2 si riportano le stime di densità del prezzo del gasolio per ogni mese. Si noti come il livello medio del prezzo aumenti col passare dei mesi, eccezion fatta per il periodo compreso tra novembre e dicembre dove la distribuzione dei prezzi si sposta verso sinistra, sottolineando quindi una diminuzione dei valori registrati.

Il grafico in Figura 3, riportante l'andamento del prezzo del Diesel a livello settimanale, porta alle stesse considerazioni. In particolare, si nota un andamento crescente del prezzo al passare delle settimane, con una crescita che appare più marcata nel mese di ottobre per poi arrestarsi a novembre. Interessante il dato relativo alle ultime quattro settimane dove si registra un calo dei valori medi settimanali, in contrasto a quanto successo per tutto il periodo precedente, avvalorando quanto emerso in Figura 2 rispetto al livello medio di dicembre. Inoltre, in quel periodo si registra la maggiore variabilità dei dati.

3.3.2 Analisi longitudinale estesa

In questa sezione si studiano i fattori macroeconomici che influenzano il prezzo del gasolio. Per avere una visione storica più complessiva, come già precedentemente detto, si è presa in esame la serie in un intervallo di tempo più esteso che va dall'inizio del 2005 fino alla fine del 2021. In prima istanza si è valutata la relazione con il prezzo del petrolio (prezzo Brent, relativo al mercato europeo). Dal primo grafico in Figura 4 è possibile notare come i due processi sottendano a uno stesso andamento non stazionario. Tuttavia, si può notare come la serie dei prezzi del gasolio sia meno irregolare.

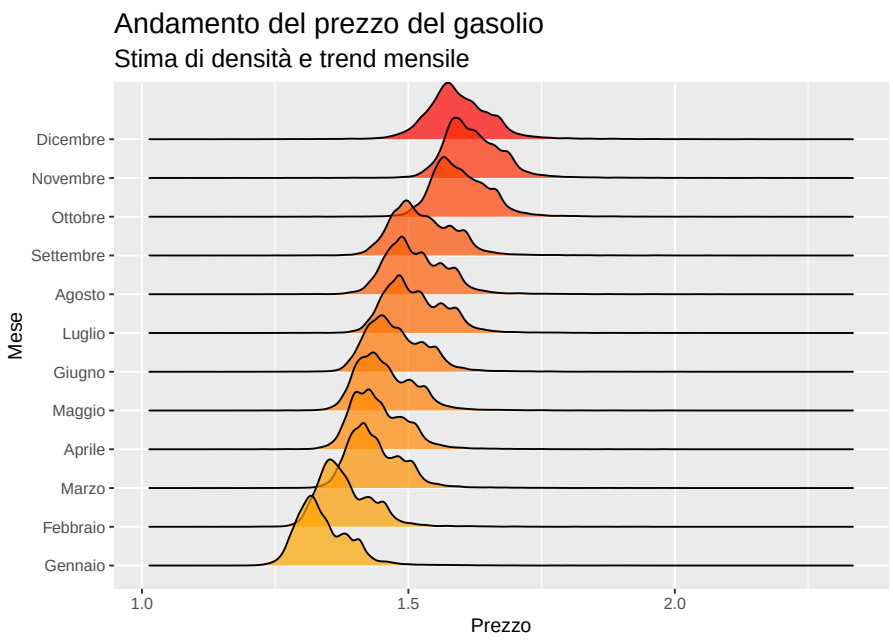

Figura 2: Andamento mensile del prezzo.

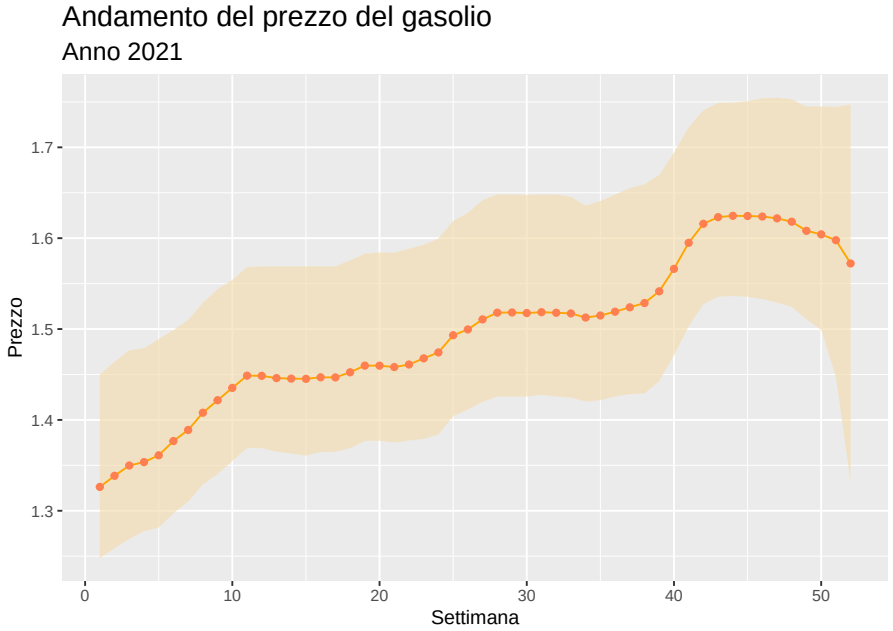

Figura 3: Andamento settimanale del prezzo.

Per dare una validità statistica a queste considerazioni, si è proceduto con un'analisi di cointegrazione, che ha confermato la presenza della relazione. I residui, mostrati nel secondo grafico, mostrano un andamento stocastico ma stazionario, ad eccezion fatta per il periodo conseguente alla crisi del 2008 e quello conseguente allo scoppio della pandemia del Coronavirus all'inizio del 2020. Il comportamento dei residui potrebbe derivare da fattori intrinsechi che intervengono nella determinazione dei prezzi, come per esempio dinamiche economiche e geopolitiche più complesse o interventi statali.

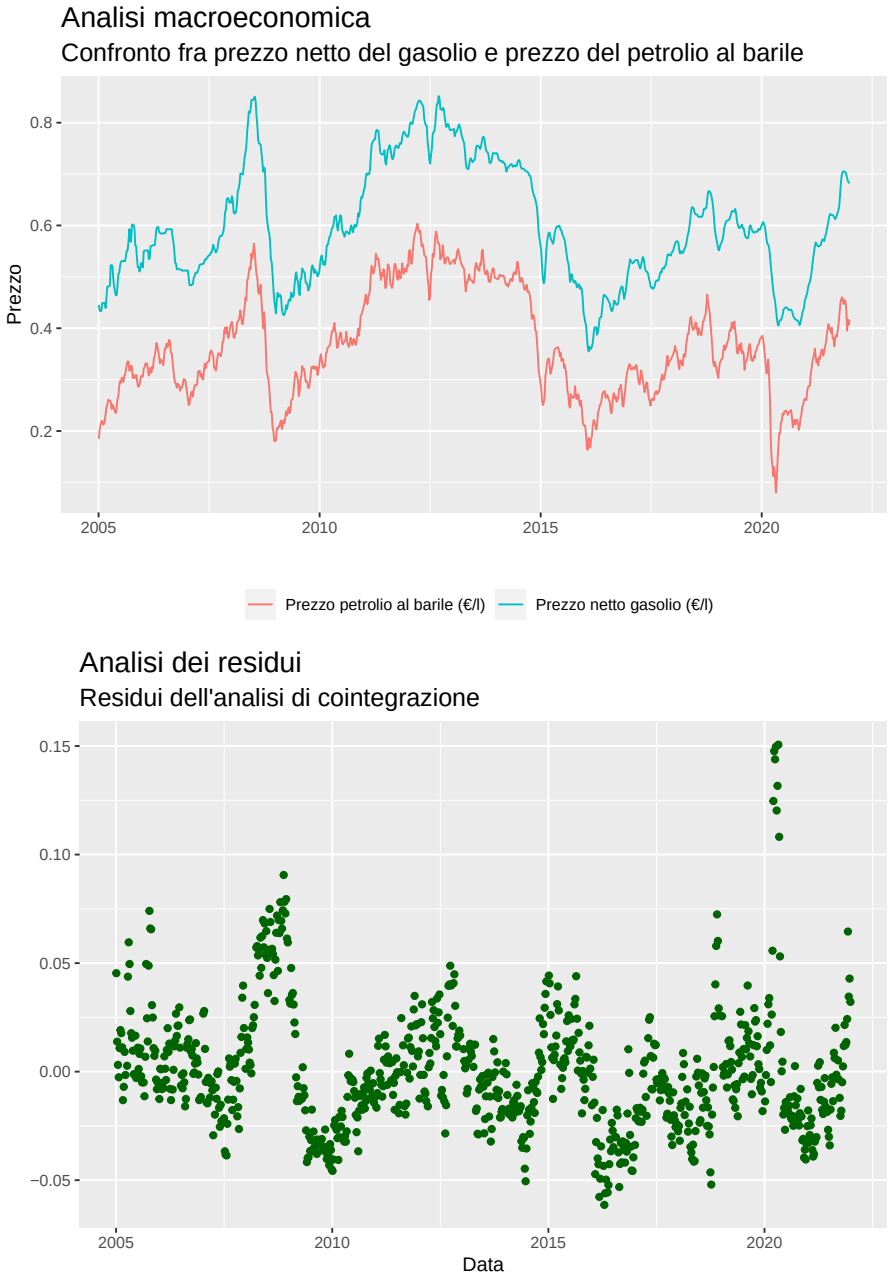

Figura 4: Sopra: andamento settimanale del prezzo del gasolio e del prezzo del petrolio Brent. Sotto: residui dell'analisi di cointegrazione.

3.4 Metrica di interesse basata sul prezzo

L'analisi longitudinale appena mostrata, con la rilevazione di una relazione di cointegrazione tra il prezzo netto del gasolio e il prezzo del petrolio al barile, porta a congetturare che l'andamento medio

del prezzo del carburante di interesse rifletta fedelmente l'evoluzione economica di fondo. Questo significa che tutti gli andamenti tendenziali e gli *shock* del mercato hanno un'influenza diretta sull'andamento del prezzo del gasolio, e in virtù di quanto detto il livello medio potrebbe essere modellato sulla base della serie storica delle materie prime specifiche, o delle serie di indicatori relativi a fattori macroeconomici più generici.

Alla luce di quanto è emerso, al fine di creare una metrica basata sul prezzo del carburante che sia depurata dall'andamento tendenziale del fenomeno, si imposta una modellazione sul prezzo settimanale per ogni impianto, utilizzando come regressori le serie storiche settimanali di quattro indicatori macroeconomici ritenuti rilevanti. Le serie utilizzate sono il tasso di cambio euro/dollaro ed euro/-sterlina, per modellare la variazione di potere d'acquisto, e due indici finanziari (WTI Crud Oil e WisdomTree Natural Gas) il cui andamento rispecchia quello dei mercati di petrolio e gas naturale, rispettivamente. Si noti come, sulla base della relazione di cointegrazione rilevata precedentemente, non sia necessario in questa fase inserire anche l'informazione relativa alla settimana come covariata.

Vengono confrontati tre approcci, valutando i residui delle serie per i singoli impianti tramite la media delle funzioni di autocorrelazione totale e parziale. Si prova in prima istanza a prendere semplicemente le deviazioni della media settimanale, tuttavia questo approccio risulta troppo approssimativo e mantiene una forte autocorrelazione residua tra le serie. Si impostano allora due modelli più complessi e flessibili, che riescano a cogliere meglio l'andamento medio anche grazie alle covariate sopra citate. Vengono stimati un modello additivo e un *gradient boosting*, il quale viene preferito considerando che questa fase di modellazione non richiede un *focus* interpretativo, ma solamente una buona capacità predittiva.

I residui di tale modellazione risultano depurati dall'andamento tendenziale del fenomeno e possono essere usati come metrica di interesse nella successiva fase di analisi, volta ad identificare caratteristiche fisse dei distributori che siano rilevanti nella definizione del prezzo. Va fatto notare, come si può vedere dal grafico in Figura 5, che permane ancora una struttura di autocorrelazione nelle serie dei residui per ogni impianto. Questo avviene perché si sta modellando solo il livello medio, e l'autocorrelazione delle serie origina sia dalla media, che ha un chiaro *trend* crescente, sia dal fatto che osservazioni consecutive per un impianto saranno autodipendenti, anche depurando i valori dal livello medio. Se si volesse arrivare ad avere residui perfettamente incorrelati, bisognerebbe impostare una modellazione autoregressiva che utilizzi tra le esplicative anche il valore del prezzo alla settimana precedente. In questo modo, tuttavia, i residui perderebbero anche tutta la variabilità ascrivibile alle caratteristiche del singolo impianto, e sarebbe quindi impossibile perseguire l'obiettivo principale dell'analisi.

Date le considerazioni appena esposte, si decide di procedere utilizzando i residui di questa prima analisi come variabile risposta nella successiva fase di modellazione. L'obiettivo è quello di riuscire a spiegare la variabilità residua associata ai singoli impianti, modellandola sulla base delle caratteristiche fisse dei distributori.

3.5 Modellazione principale

Il fine ultimo di questa analisi è quello di definire i fattori che determinano il prezzo del gasolio, al netto di andamenti macroeconomici di fondo. Per questo motivo, per spiegare il fenomeno di interesse si sono scelti dei modelli statistici che permettessero una facile interpretazione delle relazioni che intercorrono tra la metrica di riferimento, definita nella sezione precedente, e le diverse variabili

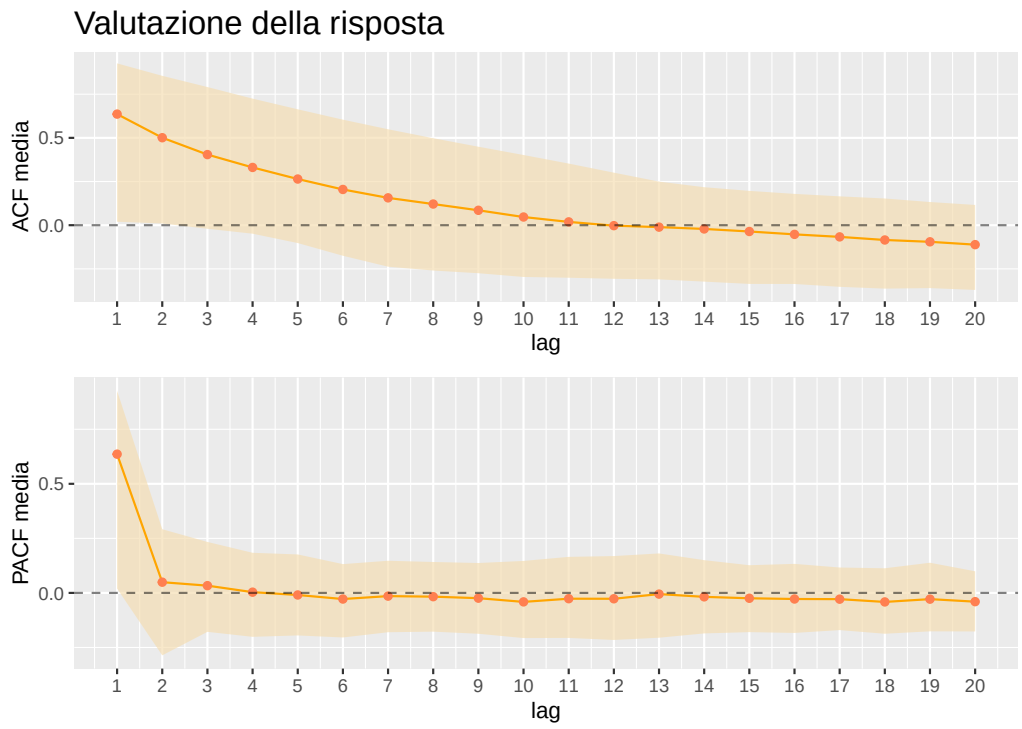

Figura 5: Valutazione della metrica di interesse

esplicative prese in considerazione. Questo ci permette di rispondere in modo chiaro e preciso alla domanda di ricerca. La modellazione è avvenuta seguendo un processo costruttivo, partendo dal modello più semplice e integrando sequenzialmente componenti di complessità.

3.5.1 Covariate utilizzate

Le variabili esplicative utilizzate in tutti i modelli per caratterizzare gli impianti sono elencate di seguito.

- *bandiera*, qualitativa con modalità: Agip.Eni, Api.Ip, Esso, Q8, Tamoil, pompe bianche, altro;
- *tipo impianto*, qualitativa con modalità: autostradale, strada statale, altro;
- *popolazione*, qualitativa dopo il raggruppamento in classi, relativa alla popolazione residente nel comune dell'impianto, con modalità: <2 000 abitanti, 2 000-10 000 ab., 10 000-50 000 ab., 50 000+ ab.;
- *grande comune*, dicotomica indicante l'appartenenza ad un comune con più di 200 000 abitanti;
- *urbanizzazione*, qualitativa con modalità: bassa, media, alta;
- *altitudine*, qualitativa con modalità: collina, pianura, montagna;
- *costa*, dicotomica indicante l'appartenenza a un comune costiero;
- *isoletta*, dicotomica indicante l'appartenenza ad una piccola isola;
- *vicini*, quantitativa relativa al numero di impianti in un raggio di 5 km;

- *molto vicino*, dicotomica relativa alla presenza di impianti nel raggio di 500m.

L'informazione relativa alla regione di appartenenza è stata gestita inserendola direttamente nel modello come variabile esplicativa, oppure attraverso variabili diverse che la potessero caratterizzare, qui riportate.

- *statuto speciale*, dicotomica riportante se la regione è a statuto speciale;
- *zona*, qualitativa con modalità: nord-est, nord-ovest, centro, sud, isole;
- *perc. inquinamento regione*, quantitativa indicante la percentuale di famiglie che percepiscono l'inquinamento come un problema della regione in cui vivono;
- *perc. mezzi lavoro regione*, quantitativa relativa alla percentuale di utilizzo di mezzi a motore per spostamenti lavorativi nella regione;
- *perc. mezzi regione*, quantitativa indicante la percentuale di utilizzo di mezzi pubblici;
- *log PIL Regione*, quantitativa relativa al logaritmo naturale del PIL regionale del 2020.

3.5.2 Modelli stimati

In prima istanza, è stato impostato un modello di regressione lineare. Questo ci permette di identificare in modo preciso le diverse componenti che determinano la variabile risposta. Particolare attenzione è stata data alla componente di variabilità attribuibile alle regioni. Si sono stimati due modelli differenti, secondo le due modalità di caratterizzare la regione di appartenenza dell'impianto sopra riportate.

Con l'idea di lavorare in una scala spaziale più fine rispetto ai modelli precedenti, si passa successivamente ad una modellazione ad effetti casuali. Nel nostro caso questo approccio permette di caratterizzare la variabilità di impianti appartenenti a province diverse, senza sovraparametrizzare i modelli. Vengono impostate due formulazioni differenti: il primo modello utilizza la regione come effetto fisso e aggiunge un effetto casuale per la provincia, mentre per il secondo si è adottata una struttura gerarchica in cui l'effetto della provincia è stimato condizionatamente alla regione di appartenenza.

Infine, si è passati ad un modello additivo generalizzato (*GAM*), che permette di stimare i contributi delle variabili numeriche attraverso metodi non parametrici. In particolare, ciò è stato utile per modellare le differenze a livello geospaziale secondo un approccio diverso rispetto agli altri modelli, introducendo una *spline* bivariata per l'effetto congiunto di longitudine e latitudine. Anche qui si stimano due modelli, uno con le covariate comuni a tutti i modelli e uno che utilizza anche la regione come esplicativa. Se l'effetto spaziale modellato in maniera liscia riuscisse a cogliere le differenze fra località in maniera soddisfacente, l'aggiunta dell'esplicativa relativa alla regione non dovrebbe migliorare il modello.

Per avere un riscontro anche a livello di capacità previsiva, i modelli precedentemente decritti sono stati confrontati con un *gradient boosting*. Le variabili utilizzate per la stima sono quelle comuni a tutti i modelli, con l'aggiunta della covariata relativa alla regione di appartenenza. Questo modello, pur non essendo direttamente interpretabile, fornisce informazioni riguardanti le variabili esplicative più importanti a livello predittivo e può quindi essere utilizzato per integrare le indicazioni emerse dai modelli precedenti.

Il confronto tra modelli viene effettuato mediante la valutazione di tre metriche di errore: il *mean absolute error* (MAE), il *root mean squared error* (RMSE) e l'indice di bontà di adattamento R^2. In questo contesto l'insieme di dati viene diviso in un insieme di stima e uno di verifica (con proporzioni uguali vista l'elevata numerosità). I modelli vengono adattati nell'insieme di stima, successivamente si ottengono le previsioni sull'insieme di verifica e da queste le metriche di errore. Per l'interpretazione finale i modelli vengono stimati nuovamente su tutti i dati.

4 Risultati

4.1 Metriche di errore

Si riportano di seguito le metriche di errore associate ai modelli descritti nella sezione precedente.

	MAE	RMSE	R2
Lineare1	0.0305	0.0496	0.3783
Lineare2	0.0320	0.0504	0.3576
Random1	0.0303	0.0491	0.3894
Random2	0.0303	0.0491	0.3894
Additivo1	0.0299	0.0480	0.4176
Additivo2	0.0299	0.0479	0.4188
Boosting	0.0253	0.0403	0.5896

Tabella 1: Metriche di errore per i modelli stimati.

La prima indicazione emersa è che la modellazione costruttiva adottata permette di ottenere un adattamento via via migliore. Discorso a parte va fatto per il *gradient boosting* che, come si può vedere ad esempio dal valore R^2, ottiene risultati nettamente superiori a livello predittivo rispetto al migliore dei restanti modelli.

Relativamente alla modellazione interpretabile, si noti dai risultati per i modelli lineari come la variabilità regionale venga modellata meglio con un singolo effetto fisso piuttosto che mediante le diverse covariate disponibili in alternativa. Si può quindi affermare come tali variabili non riescano a caratterizzare a sufficienza le differenze regionali.

Sulla base dei risultati dei modelli con effetti casuali si può intuire come l'utilizzo dell'informazione relativa alla provincia di appartenenza porti ad un lieve miglioramento in termini di prestazioni. I due modelli risultano equivalenti a livello predittivo, si preferisce quindi il primo che permette di stimare un effetto fisso per *regione*. Il coefficiente di correlazione intraclasse stimato, che misura il grado di omogeneità entro le province, presenta un valore basso (pari a circa 0.03), ma nonostante ciò con tale formulazione si migliora leggermente l'adattamento del modello ai dati osservati.

L'approccio alternativo alla gestione dell'informazione geospaziale, permesso dai due modelli additivi, porta ad un ulteriore miglioramento in termini previsivi e generalmente alle prestazioni migliori, fatta eccezione per il *gradient boosting*. Nello specifico si noti come l'aggiunta dell'effetto relativo alla regione nel secondo modello non porti a variazioni sostanziali in termini di adattamento, a indicare che le caratteristiche spaziali vengono colte in maniera soddisfacente dall'effetto liscio su longitudine e latitudine.

4.2 Interpretazione delle covariate

In questa sezione si cerca di interpretare il contributo dei diversi fattori relativi alle caratteristiche fisse dei distributori nella definizione del prezzo, sulla base dei modelli stimati. I commenti vengono basati principalmente sui risultati del primo modello additivo, valutato come il migliore in termini predittivi. Poiché in quest'ultimo gli effetti geo-spaziali vengono considerati modellando congiuntamente longitudine e latitudine, per valutare l'effetto delle diverse regioni ci si basa sulle stime del miglior modello ad effetti misti. Questo utilizza la variabile *regione* come effetto fisso e aggiunge un effetto casuale sulla provincia di appartenenza; inoltre, ha stime comparabili a quelle del miglior modello lineare, avendo la stessa formulazione per la parte fissa. Di seguito si riportano solo le stime dei coefficienti relativi alla variabile *regione*, dove il livello di riferimento è rappresentato dal Lazio. Questa scelta è stata fatta sulla base dell'analisi esplorativa, per prendere come riferimento una regione associata ad un prezzo medio annuale intermedio.

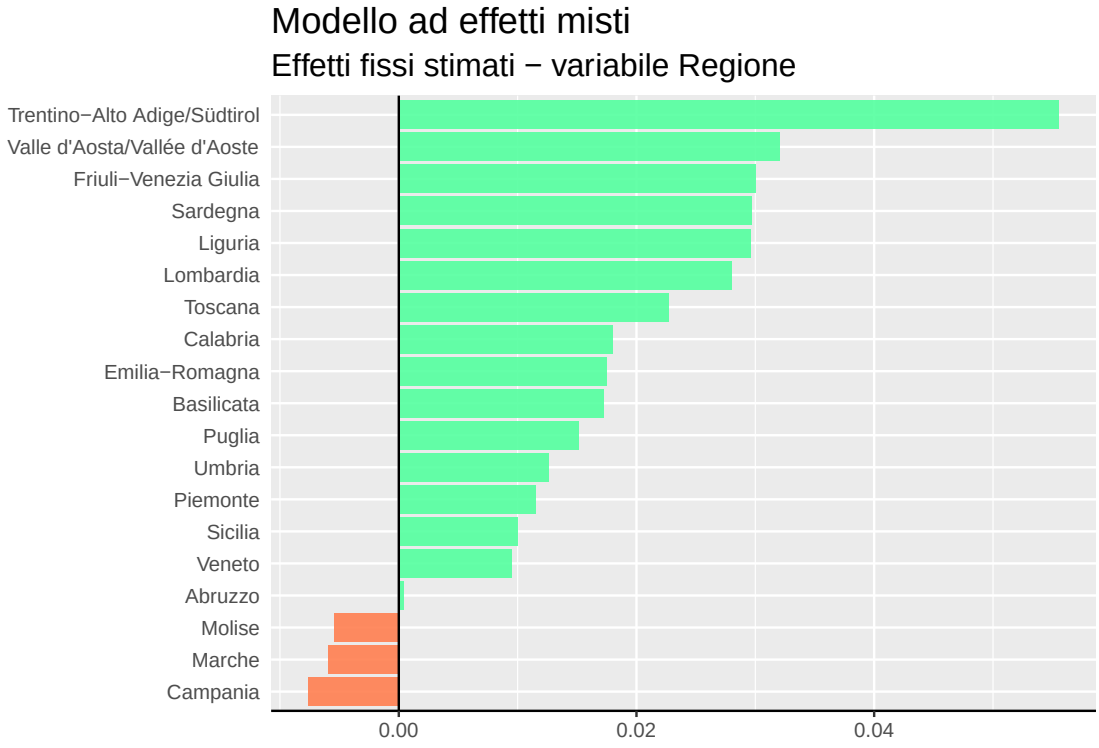

Figura 6: Coefficienti relativi alla variabile *regione* del miglior modello con effetti casuali.

Per quanto riguarda le stime, salta all'occhio il fatto che le prime quattro regioni a cui è associato un effetto positivo sui prezzi siano tutte a statuto speciale (Trentino Alto-Adige, Val d'Aosta, Friuli Venezia-Giulia e Sardegna). Analogamente si nota come le regioni a cui sono associate i tre coefficienti più bassi siano Campania, Marche e Molise. Si riportano ora in Figura 7 le stime degli effetti fissi per il primo modello additivo. Avendo infatti valutato in precedenza come questo approccio di modellazione delle caratteristiche geo-spaziali permetta di ottenere l'adattamento migliore fra i modelli interpretabili, ne consegue che a questo modello sono associate le stime più affidabili degli effetti relativi alle caratteristiche fisse dei singoli impianti. Non si riportano visualizzazioni tridimensionali per l'effetto liscio associato all'interazione tra longitudine e latitudine, in quanto non visibilmente informative.

Gli effetti stimati con un maggior impatto in positivo sulla definizione del prezzo sono associati all'appartenenza ad un comune isolano e agli impianti autostradali. Relativamente alle piccole isole, è

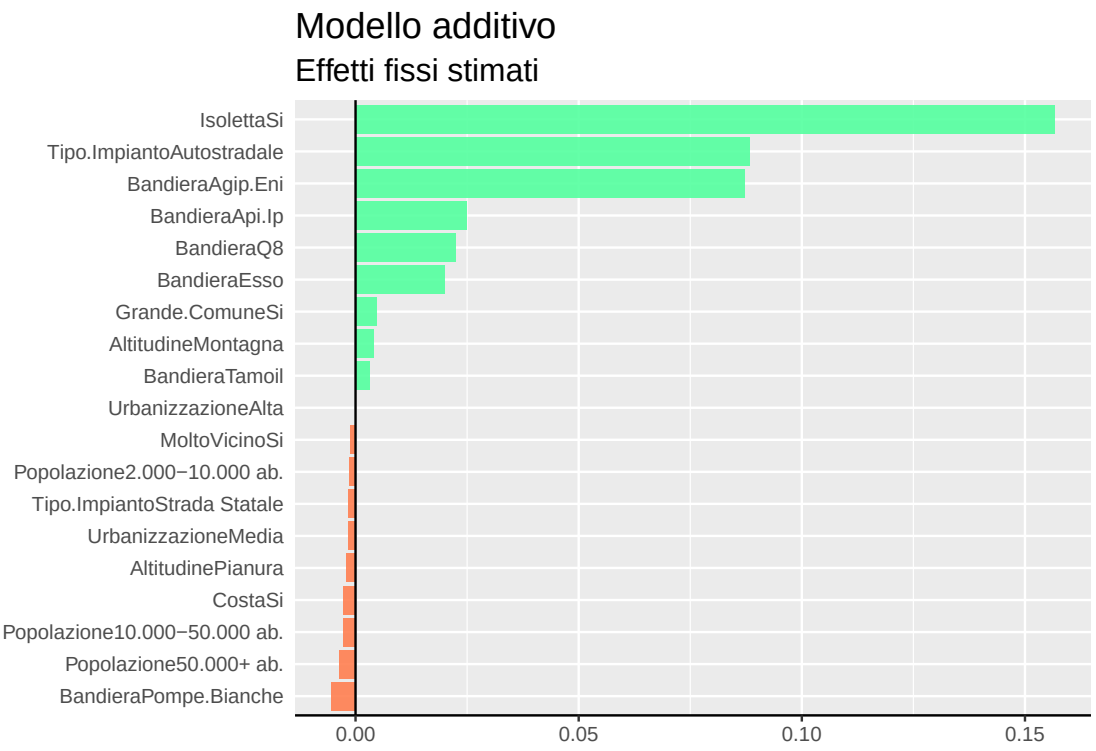

Figura 7: Coefficienti del miglior modello additivo.

facile immaginare come questo fenomeno sia attribuibile ai maggiori costi di trasporto del carburante; per quanto riguarda agli impianti autostradali i prezzi maggiori sono imputabili a caratteristiche del mercato di riferimento che verranno prese in analisi successivamente.

Per la variabile *bandiera*, che sintetizza la capacità dei *competitors* di definire prezzi maggiori o minori della media nazionale, i coefficienti permettono di confrontare gli attori maggiormente presenti sul mercato, prendendo a riferimento il livello di tutti i *players* in minoranza. Il maggior effetto positivo sulla definizione del prezzo viene stimato per Agip.Eni, seguita da Api.Ip, Q8, Esso e Tamoil. Di particolare interesse è il caso delle pompe bianche, ossia distributori di carburante indipendenti non vincolati ad una delle *major* petrolifere, alle quali sono associati prezzi più bassi sia secondo il modello, sia secondo le ricerche effettuate.

Per quanto riguarda la popolazione del comune di riferimento sembra che i prezzi siano maggiori in zone meno popolose, con l'eccezione dei grandi comuni. Inoltre, emerge come a zone montane e collinari siano associate stime più alte dei prezzi, così come per gli impianti in comuni costieri. Per quanto riguarda l'effetto del numero degli impianti vicini e della presenza di impianti nelle immediate vicinanze, nonostante la piccola magnitudine dei coefficienti, sembra di poter affermare che la presenza di concorrenti diretti nelle vicinanze dell'impianto di riferimento porti ad una diminuzione dei prezzi del carburante, il che potrebbe essere attribuito a dinamiche di tipo concorrenziale.

Viene ora riportato in Figura 8 il grafico di importanza relativa delle variabili per il *gradient boosting*. Nonostante questi valori non permettano di caratterizzare la direzione e la magnitudine degli effetti associati alle varie modalità dei fattori, in virtù della superiore capacità predittiva di questo modello si valuta integrare le considerazioni emerse fino a questo punto con le indicazioni del grafico.

Secondo il modello l'importanza maggiore a livello previsivo viene attribuita alla *bandiera*, seguita

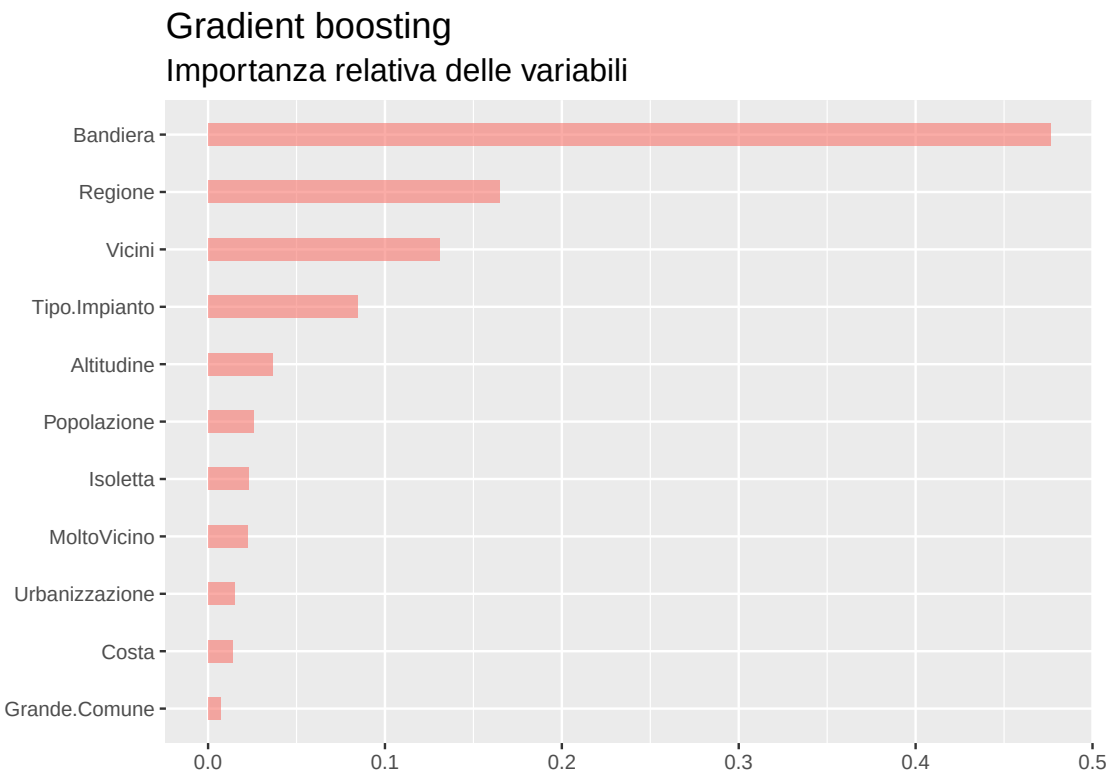

Figura 8: Grafico di importanza delle variabili per il *gradient boosting*.

da *regione*. Questo sembra suggerire che i principali *competitor* presentano caratteristiche uniche, tali da determinare variazioni dei prezzi negli impianti associati a diverse bandiere maggiori rispetto rispetto a quelle imputabili alle differenze geografiche. Interessante notare l'importanza della variabile *vicini* associata al numero di impianti concorrenti in un raggio di 5 km, che corrobora le considerazioni fatte precedentemente.

4.3 Prezzi ricostruiti

Per concludere l'analisi dei risultati, si visualizzano i risultati finali della modellazione in due stadi adottata. Si va dunque a ricostruire il valore del prezzo sommando i valori adattati del primo *gradient boosting*, utilizzato nella definizione della metrica di interesse, e quelli del primo modello additivo, relativo alla fase di modellazione principale. Si riportano nelle mappe in Figura 9 i prezzi medi annuali previsti, stratificati per regione e provincia. Si può vedere come i grafici risultino praticamente non distinguibili da quelli presentati in Figura 1 e basati sui prezzi osservati. A conferma del buon adattamento raggiunto, si noti come in Figura 10 il valore medio settimanale su tutti gli impianti previsto (linea spessa arancione) sia sovrapposto al valore osservato (linea tratteggiata nera).

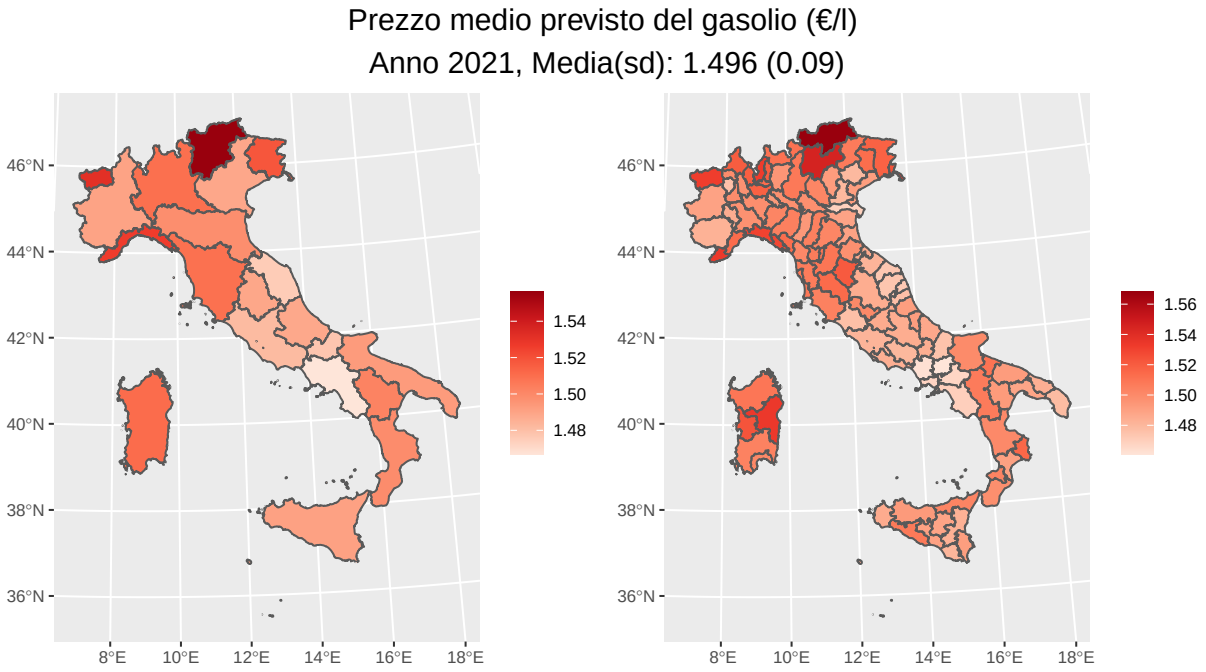

Figura 9: Prezzo medio previsto del carburante per province e regioni.

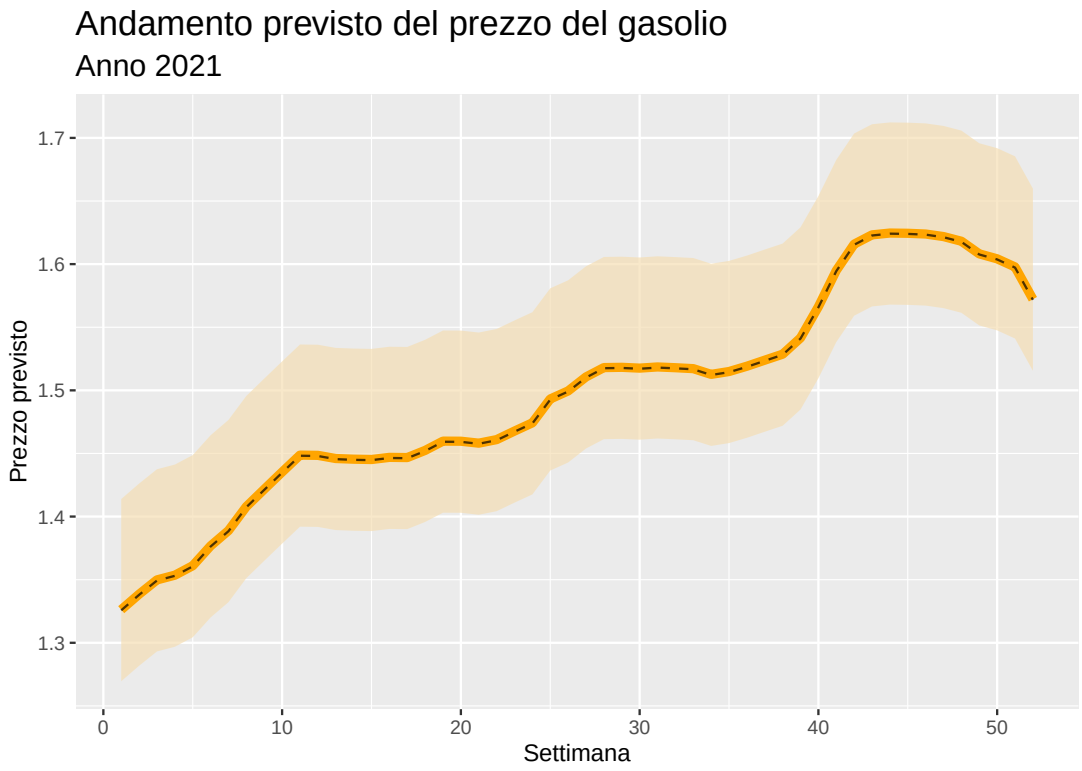

Figura 10: Andamento settimanale previsto.

5 Strategia di business

Per poter definire al meglio la strategia di business, è necessario prima soffermarsi su alcuni aspetti che caratterizzano il mercato italiano del carburante.

5.1 Caratteristiche del mercato italiano

5.1.1 Mercato saturo

In primo luogo, come evidenziato dal grafico riportato in Figura 11, si è notato come il numero di distributori in Italia (21 750) risulti essere nettamente maggiore rispetto agli altri principali paesi europei come Francia (11 160), Germania (14 459) e Spagna (11 650).

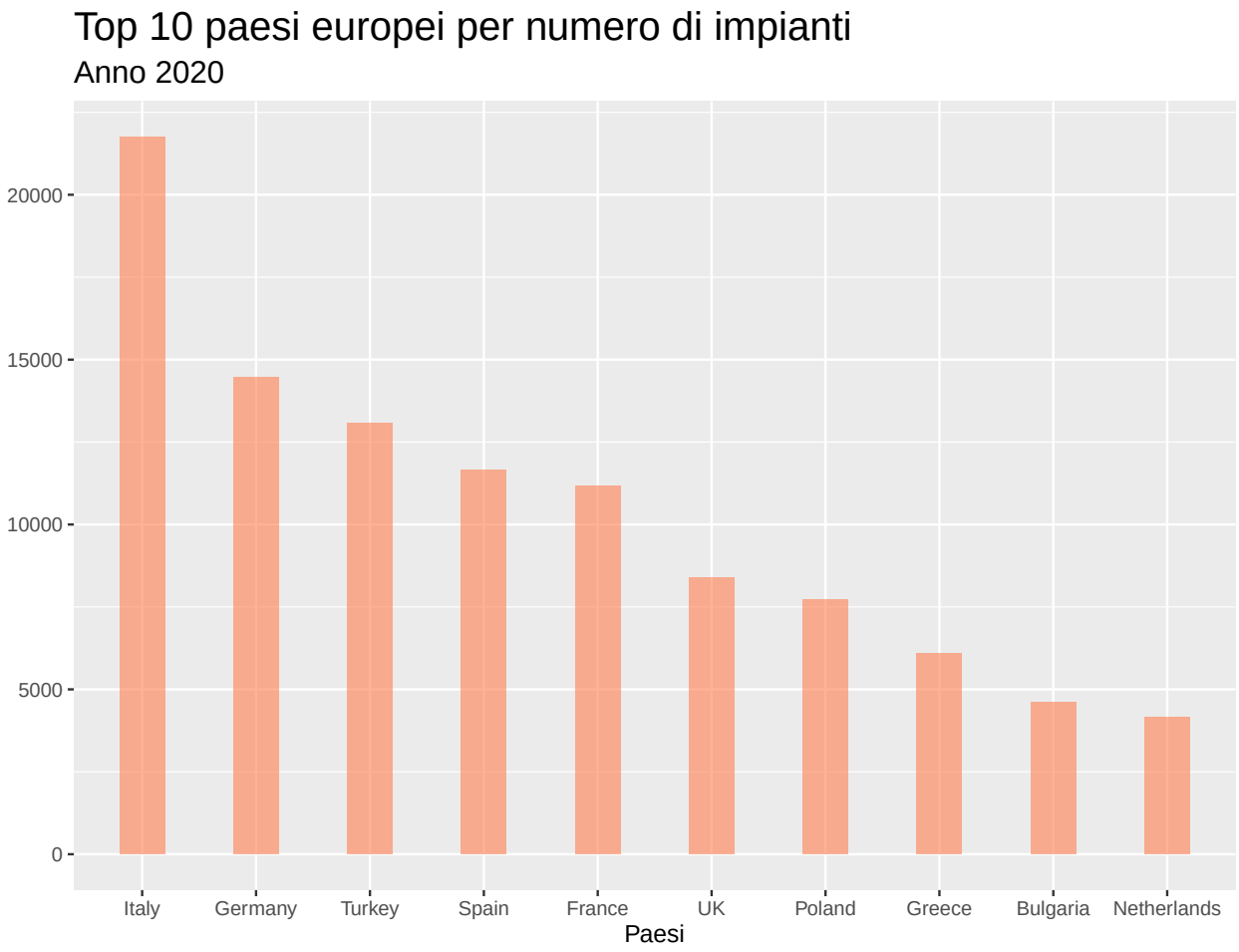

Figura 11: Numero di impianti per paese.

La capillarità del sistema distributivo rappresenta evidentemente un'agevolazione per il consumatore, ma d'altra parte risulta essere un grosso svantaggio per un nuovo *player* intenzionato ad immettersi sul mercato.
Per far fronte a questa situazione il governo italiano, a partire dai primi anni 2000 con il D.lgs n.32/98, ha iniziato ad osteggiare e in alcuni casi addirittura vietare (art. 3 D.lgs n.32/98) l'apertura di nuovi impianti, imponendo condizioni stringenti. Questa politica ha portato a un'iniziale diminuzione del numero di distributori, fino alla stabilizzazione intorno alle 21 000 unità.

Altro aspetto importante riguardante l'apertura di nuovi impianti e la riqualificazione di quelli già esistenti emerge dal D. lgs 257/2016, che regola la diffusione di combustibili alternativi in Italia.
In particolare, nell'art.18 è scritto che *"le regioni, nel caso di autorizzazione alla realizzazione di nuovi impianti di distribuzione carburanti e di ristrutturazione totale degli impianti di distribuzione carburanti esistenti, prevedono l'obbligo di dotarsi di infrastrutture di ricarica elettrica . . . , nonché di rifornimento di GNC o GNL".* Citando sempre l'art.18 *"Non sono soggetti a tale obbligo gli impianti di distribuzione carburanti localizzati nelle aree svantaggiate".*
È bene evidenziare come la dotazione di impianti per il rifornimento di GNC e GNL (gas naturale compresso e gas naturale liquido, alternative al gas metano) comporti costi stimabili intorno agli 800 000 euro. Risulta quindi evidente che una strategia di business basata sull'apertura di nuovi impianti debba essere accompagnata da grossi investimenti iniziali e debba tenere conto della situazione di mercato "saturo" presente in Italia.
L'unica strada percorribile per tagliare notevolmente i costi iniziali, riprendendo quando detto dall'art.18 del D. lgs 257/2016, sarebbe quella di considerare l'apertura di nuovi impianti nelle aree svantaggiate, ovvero nelle zone geografiche meno raggiungibili (comuni di montagna o isolani) e meno urbanizzate.

5.1.2 Autostrade

Un discorso a parte va fatto per gli impianti di distribuzione autostradali. Le aree di servizio vengono affidate in sub-concessione attraverso bandi di gara tenuti dai concessionari autostradali (ASPI, ANAS, SAT. . .). Questi bandi, oltre all'affidamento dei servizi di distribuzione carburanti, garantiscono tipicamente la concessione dei servizi di ristoro e *market* collocati lungo la rete autostradale nazionale. Risulta quindi evidente come non si possa basare una *price strategy* unicamente su questa tipologia di impianti, per la forte dipendenza dalle gare di appalto per le concessioni, dall'esito incerto. D'altro canto, pur non rappresentando il cuore della strategia di business, qualora il nostro *player* si occupasse anche di servizi quali ristoro e *market*, la concessione di alcuni impianti autostradali potrebbe portare a maggiori margini di guadagno ed essere una buona appendice per la società.

5.1.3 Pompe Bianche

Altro aspetto interessante è quello delle pompe bianche, ovvero quei distributori di carburante indipendenti non vincolati ad una delle *major* petrolifere. Le pompe bianche possono essere legate ad aziende di distribuzione più o meno estese che fungono da grossisti, oppure, in casi più rari, essere totalmente indipendenti e rifornirsi all'ingrosso.
Di particolare interesse è il trend crescente del numero di pompe bianche, che dalle circa 1 800 unità del 2010 è arrivato ad oltre 4 000, in totale controtendenza rispetto all'andamento del numero di impianti in Italia.
L'aumento di questo tipo di impianti risulta essere giustificato da alcuni vantaggi. Non esponendo nessun logo, le pompe bianche non sono tenute a pagare *royalties* alle *major* petrolifere. Inoltre, il non essere legate a vincoli contrattuali permette loro di rifornirsi all'ingrosso, scegliendo il prezzo più vantaggioso. Tutto questo permette di stabilire prezzi più bassi, mantenendo comunque un buon margine di guadagno.

5.2 Approccio suggerito

Alla luce di quanto detto finora riguardo al mercato del carburante in Italia, si suggerisce al committente di adottare una strategia che tenga conto della situazione di mercato semi-saturo e della diversa condizione dei distributori indipendenti. Nello specifico, un piano di entrata nel mercato potrebbe

concretizzarsi con la rilevazione di un numero medio-alto di impianti, opposta invece all'apertura di nuove stazioni. Secondo quanto emerso sulla base dell'art.18 del D.lgs 257/2016, questo porterebbe ad una riduzione dei costi associati. Per quanto concerne gli impianti rilevati, a seconda delle caratteristiche del *player* e della natura dell'investimento programmato, emergono due strade. La prima, nel caso si stia pensando all'ipotetico committente come ad una *major* petrolifera, sarebbe chiaramente quella di riconvertire le stazioni alla propria bandiera, con i conseguenti vantaggi in termini di visibilità, *royalties* sul servito, ecc... In alternativa, pensando ad un generico investimento, il suggerimento sarebbe quello di convertire gli impianti a pompe bianche, non legandosi ad una *major petrolifera* e rifornendosi di carburante all'ingrosso. In sostanza il *player* avrebbe dunque il controllo di tali stazioni, senza che figuri una bandiera sulle stesse. A prescindere dalle due strade, l'indicazione comune riguarda la rilevazione di impianti esistenti, da realizzare in modo capillare sul territorio. La scelta dei luoghi dovrebbe essere fatta ricercando le zone con un mercato associato a prezzi maggiori della media, per massimizzare il margine di guadagno.

A questo punto, per dare indicazioni più precise a livello spaziale, ci si appoggia sui risultati emersi dalla modellazione sviluppata. Una prima indicazione generica riguarda le regioni più appetibili per tale investimento, ovvero Trentino Alto-Adige, Valle d'Aosta, Friuli Venezia-Giulia, e Sardegna. Queste quattro regioni, tutte a statuto speciale, sono quelle risultate associate a prezzi mediamente maggiori. Altro aspetto evidenziato dai modelli è quello relativo alle zone altimetriche e all'urbanizzazione. In particolare, come già detto in precedenza, località collinari e montuose sono legate a prezzi maggiori, e lo stesso discorso vale per zone con un minor livello di urbanizzazione. La strategia suggerita è dunque quella di concentrare gli investimenti in queste zone, tenendo sempre in considerazione il bacino di traffico e l'utenza media in modo da rendere sostenibile la spesa. Questa proposta, oltre allo svincolarsi dagli obblighi previsti dalle ultime normative, permette anche di avere meno *competitors* sul territorio. La presenza di impianti nelle immediate vicinanze, per quanto emerso dall'analisi, comporta una diminuzione dei prezzi del carburante, legata verosimilmente a logiche concorrenziali. Seguendo dunque le indicazioni proposte, gli impianti rilevati sarebbero per la maggior parte in regioni con un mercato associato a prezzi più alti. Inoltre, le caratteristiche delle zone individuate (zone meno urbanizzate e con pochi vicini) sono tali da permettere una maggiore flessibilità nella definizione di una *price strategy*.