Università degli studi di Milano–Bicocca
Scuola di Economia e Statistica

Corso di Laurea in
Scienze Statistiche ed Economiche

Sondaggi Elettorali: Studio sull'Affidabilità Statistica degli Istituti Demoscopici e delle loro Distorsioni Sistematiche

Relatore: Dott. Tommaso Rigon
Correlatore: Dott. Paolo Maranzano

Tesi di laurea di:
Tommaso Menghini
Matricola N. 864946

Anno Accademico 2025/2026

Indice

Introduzione

In questa tesi lo studio è stato condotto lungo due binari paralleli, che danno luogo a due analisi principali. La prima ha l'obiettivo di comprendere se esistano distorsioni sistematiche nelle stime delle intenzioni di voto basate su sondaggi pre-elettorali. La seconda è invece volta a verificare la presenza di differenze strutturali nell'affidabilità delle principali agenzie demoscopiche operanti in Italia, interrogandosi sulla possibilità che alcune risultino, strutturalmente, più precise di altre nel prevedere gli esiti elettorali. La tesi nasce da un interesse personale per l'imponente apparato statistico che accompagna le elezioni presidenziali statunitensi. Nella copertura da parte dei media dell'evento elettorale, seguito a livello globale, la previsione del risultato ricopre un largo spazio. In questo contesto, il ruolo dello statistico assume una rilevanza centrale. Lo statistico esce dalla dimensione più riservata del lavoro tecnico, svolto lontano dall'attenzione pubblica, e diventa una figura esposta nel dibattito collettivo. Le metodologie statistiche impiegate per la previsione elettorale diventano esse stesse parte integrante dell'evento, oggetto di attenzione, discussione e spesso di critica. Questo scenario rende naturale interrogarsi non solo sull'accuratezza delle previsioni, ma anche sui limiti intrinsechi degli strumenti utilizzati e sulle modalità con cui l'incertezza viene trattata e comunicata. Al tempo stesso, rappresenta un banco di prova utile per comprendere il ruolo dello statistico nella società e nel rapporto con la società stessa.

Nel capitolo 1 si discute il ruolo ricoperto dai sondaggi politici nelle democrazie contemporanee e le principali sfide legate alla loro produzione, interpretazione e affidabilità. I sondaggi elettorali rappresentano uno degli strumenti più efficienti per tentare di misurare le opinioni di un'intera popolazione; allo stesso tempo, costituiscono un elemento delicato e spesso controverso del processo democratico. Essi hanno infatti il potere di influenzare la percezione pubblica, la comunicazione dei media e, in alcuni casi, anche il comportamento degli elettori. Il capitolo si apre con un inquadramento storico, richiamando alcune delle principali polemiche sorte attorno alla credibilità dei sondaggi, in particolare il referendum sulla Brexit

e le elezioni presidenziali statunitensi del 2016. Successivamente, senza entrare nei dettagli matematici formali, viene illustrato cosa significhi costruire un sondaggio elettorale: dalle diverse tipologie di errore che possono compromettere la qualità delle stime, alle strategie di correzione adottate dai sondaggisti per mitigare tali distorsioni. Viene poi approfondito il modo in cui la presenza di distorsioni sistematiche e l'utilizzo di metodi di aggregazione influenzino le previsioni elettorali. Infine, si introduce il fenomeno dell'*herding*, ovvero la tendenza dei sondaggi a convergere tra loro più di quanto sarebbe lecito attendersi se gli istituti demoscopici operassero in modo indipendente. L'insieme di questi temi fornisce il quadro concettuale necessario per comprendere le potenzialità e le criticità dei sondaggi politici.

Il capitolo 2 è dedicato alla descrizione dei dati utilizzati nella tesi. Essi sono stati messi a disposizione dall'azienda YouTrend, che li impiega per la costruzione della *Supermedia settimanale*, un aggregatore di sondaggi politici provenienti da diversi istituti demoscopici, che aggiornano le proprie rilevazioni con cadenza settimanale. Nel complesso, sono disponibili rilevazioni condotte dall'inizio del 2018 fino alla fine del 2024. Per ciascun sondaggio si conoscono la data di fine rilevazione, la numerosità campionaria e le stime delle intenzioni di voto per un insieme di partiti politici. Il capitolo definisce il perimetro dell'analisi, illustrando i criteri adottati per la selezione dei partiti di interesse - limitati alle principali forze politiche presenti stabilmente nel periodo considerato - e degli istituti demoscopici, includendo esclusivamente quelli che dispongono di un numero sufficiente di rilevazioni pre-elettorali per tutte le elezioni analizzate. Per le motivazioni appena illustrate, la discrepanza tra previsioni basate sui sondaggi pre-elettorali e risultati ufficiali delle urne assume un ruolo cruciale nella definizione della variabile risposta. Quest'ultima deve essere in grado di catturare l'errore di previsione in modo coerente e comparabile tra partiti caratterizzati da livelli di consenso molto differenti. Viene quindi introdotto l'errore relativo, calcolato sull'ultimo sondaggio disponibile per ciascun istituto, come indice di misura dell'affidabilità di un'agenzia demoscopica.

Il processo metodologico utilizzato per costruire l'evoluzione giornaliera delle intenzioni di voto per ciascun partito, a partire dai sondaggi pubblicati dalle agenzie demoscopiche selezionate, è trattato nel Capitolo 3. Gli strumenti centrali sono il modello *state space* e il filtro di Kalman. Il primo consente di modellare le serie storiche come somma di componenti distinte, offrendo un approccio flessibile e adatto ad affrontare un'ampia varietà di problemi; il secondo rappresenta l'algoritmo che regola la stima coerente delle intenzioni di voto di un determinato partito in un dato periodo. La prima parte del capitolo è la più tecnica dell'intera tesi ed è dedicata

all'introduzione del quadro teorico di riferimento. Nella seconda parte viene invece illustrata l'applicazione del modello ai dati e l'ottenimento delle stime necessarie alla costruzione di una variabile risposta alternativa, basata sull'interpolazione delle intenzioni di voto fino al giorno dell'elezione.

L'ultimo capitolo della tesi è dedicato alla descrizione delle analisi condotte sui due dataset ottenuti rispettivamente al termine dei capitoli 2 e 3. I dataset in questione condividono la medesima struttura informativa e le stesse unità statistiche. In entrambi i casi, le variabili esplicative considerate sono `Istituto`, `Partito` e `Elezione`. La differenza tra i due riguarda esclusivamente la definizione della variabile risposta, che viene costruita in modo distinto. Nel primo caso, la previsione elettorale è rappresentata dall'ultimo sondaggio pubblicato da ciascuna agenzia prima dell'elezione; nel secondo, essa è invece ottenuta attraverso l'interpolazione delle intenzioni di voto stimata nel giorno dell'elezione mediante il filtro di Kalman. In questo senso, più che di due dataset distinti, si tratta di due diverse specificazioni della variabile risposta applicate a una medesima base dati. Le due analisi vengono portate avanti parallelamente: da un lato si studia l'eventuale presenza di distorsioni sistematiche nelle previsioni delle intenzioni di voto per ciascun partito e turno elettorale; dall'altro si analizza l'eterogeneità nella precisione delle previsioni fornite dai diversi istituti demoscopici.

Al testo principale sono affiancate diverse appendici, che raccolgono materiale aggiuntivo relativo a grafici, calcoli, approfondimenti metodologici e dimostrazioni, con l'obiettivo di rendere il corpo centrale della tesi più fluido senza rinunciare alla completezza.

Capitolo 1

Il ruolo e le sfide legate ai sondaggi politici

In questo capitolo si affrontano il ruolo ricoperto dai sondaggi politici nelle democrazie contemporanee e le principali sfide legate alla loro produzione, interpretazione e affidabilità. I sondaggi elettorali rappresentano infatti uno degli strumenti più utili per tentare di misurare e comprendere le opinioni dei cittadini. Allo stesso tempo, essi costituiscono anche un elemento delicato e discusso del processo democratico, poiché hanno il potere di influenzare la percezione pubblica, la comunicazione dei media e talvolta anche il comportamento degli elettori.

Dopo una rapida introduzione sulle ragioni della crescenti polemiche sorte attorno alla credibilità dei sondaggi - facendo anche esempi ormai storici - il capitolo presenta ciò che significa costruire un sondaggio: dalle diverse tipologie di errore che possono compromettere la qualità delle stime, alle strategie di correzione adottate dai sondaggisti per mitigare tali distorsioni. Poi viene approfondito come la presenza di queste distorsioni sistematiche e l'uso di metodi di aggregazione influiscano sulle previsioni elettorali. Infine si introduce il fenomeno dell'herding, ovvero la tendenza dei sondaggi a convergere tra loro più di quanto sarebbe dovuto se gli istituti demoscopici agissero in modo indipendente.

Questo insieme di temi fornisce il quadro concettuale necessario per comprendere le potenzialità e le criticità dei sondaggi politici.

1.1 Sondaggi politici e democrazia

I sondaggi elettorali sono una delle manifestazioni più comuni della statistica nella vita di tutti i giorni; nonostante ciò, nell'ultimo decennio, l'affidabilità legata a questi strumenti è stata sempre più messa in discussione, specialmente alla luce di previsioni per eventi elettorali di alto profilo rivelatesi poi incorrette. Alcuni eventi politici chiave sono stati al centro di forti polemiche - di tipo politico, ma anche

scientifico - legate a discrepanze tra le previsioni basate sui sondaggi e l'effettivo risultato. Tali discrepanze si sono rivelate tanto importanti che sono passate alla storia, con il referendum della Brexit e le elezioni presidenziali americane del 2016 che si distinguono come gli esempi più celebri. Come discusso in Gelman (2021), alla vigilia del referendum britannico sull'uscita del Regno Unito dall'Unione Europea, i principali sondaggi indicavano un lieve vantaggio per il *Remain*, con percentuali intorno al 48% contro il 46% attribuito al *Leave*. L'esito finale del referendum fu invece opposto, con il 52% dei voti a favore del *Leave* e il 48% per il *Remain*. In tal senso il risultato delle elezioni presidenziali degli Stati Uniti del 2016 è più di tutti riconosciuto come avvenimento scioccante (si legga Gelman & Azari (2017)): Hillary Clinton era data in testa nei sondaggi nazionali e soprattuto in vari *swing states*, ma poi perse nella corsa alla presidenza, in quanto ottenne meno grandi elettori del suo avversario. Quest'ultimo, forse per l'importanza storica legata a quello che la presidenza degli Stati Uniti comporta, rappresenta il vero spartiacque nel modo in cui l'opinione pubblica - e la comunità scientifica - percepisce l'affidabilità dei sondaggi politici e delle previsioni elettorali.

In realtà i sondaggi politici non si riducono solamente ad essere strumenti per prevedere elezioni: i sondaggisti contribuiscono alla comprensione di oggetti di altra natura, come trend di opinioni o preferenze verso certe politiche. Nonostante ciò, i sondaggi elettorali sostanzialmente monopolizzano l'attenzione del pubblico. Questo è dovuto certamente al contesto che si viene a creare quando la previsione basata su quest'ultimi viene confrontata con i veri risultati dell'elezione. Un momento tanto intrattenente quanto delicato.

Le indagini campionarie pre-elettorali sono un aspetto centrale in una democrazia moderna. Quel tipo di errori menzionati precedentemente, come fatto notare in Bunker (2025), non solo fanno emergere dubbi sulla capacità dei metodi tradizionali di fare sondaggi di catturare in modo accurato le tendenze degli elettori, ma in una certa misura rischiano di minare anche la fiducia pubblica che si ha nel processo democratico.

Non solo, è ragionevole pensare che la maggior parte delle persone che vada a votare si affidi ad informazioni di varia natura per guidare le loro decisioni in cabina elettorale. Tra questo insieme di informazioni ci sono anche i sondaggi pre-elettorali. Dunque non si tratta solo di essere uno strumento descrittivo; i sondaggi hanno anche la possibilità di influenzare il comportamento di un elettore. É chiaro come entrando in contatto con sondaggi elettorali si possa essere più esposti a fenomeni come l'effetto *bandwagon* ("salire sul carro del vincitore") o l'effetto dell'*underdog*

("empatia con lo sfavorito"): impressioni entrambe connesse alla conoscenza della popolarità di un determinato candidato. Si immagini un'elezione con una soglia di sbarramento del 4%: gli elettori che supportano un partito che è al di sotto di quel livello nei sondaggi potrebbero decidere di evitare di sprecare il loro voto trasferendo il loro supporto altrove. Questa possibilità dei sondaggi di influenzare, che va oltre al semplice misurare, comporta una vulnerabilità che rende ogni errore una minaccia alla propria credibilità.

1.2 La credibilità dei sondaggi

Le indagini demoscopiche sono uno degli strumenti attraverso il quale una democrazia può guardarsi allo specchio, e malgrado i loro celebri fallimenti, rimangono il mezzo più efficiente per stimare l'opinione di una intera popolazione. Infatti decisori politici, giornalisti, ma anche cittadini comuni si affidano quotidianamente a sondaggi perchè quest'ultimi permettono di migliorare il processo decisionale, rispetto al basarsi solamente a pure intuizioni personali.

Dopotutto i sondaggi non sono così lontani da quello che è la realtà. Un errore di 2 o 3 punti percentuali è un problema se si deve prevedere il risultato di un'elezione davvero combattuta, ma altrimenti non rappresenta uno sbaglio con grosse conseguenze. Nelle elezioni presidenziali del 2016, si stimava che Hillary Clinton godesse di una percentuale di voto pari al 53%. A votazione conclusa la vera percentuale di voto fu pari al 51%, non così lontano da come si era previsto (Gelman, 2021). La ragione per cui i sondaggi hanno questo livello di precisione è dovuta al fatto che si conoscono molte, se non la maggior parte, delle variabili chiave che predicono il voto - come ad esempio l'età, il genere, il livello d'educazione, l'etnia, cosa si è votato in elezioni precedenti - e il sondaggista sa come aggiustare il proprio campione per questi fattori.

Esistono elezioni combattute, caratterizzate da contesti incerti, tant'è che i fallimenti citati sono veri, ma è altrettanto vero che le performance dei sondaggi non sono così povere come spesso le si vuol far passare. Non solo, in quanto, per altri scopi, che non sono prevedere elezioni combattute, andrebbe anche bene stimare opinioni con una percentuale di incertezza di entità simili.

Per un sondaggista professionista, pubblicare una previsione di un evento unico, come un'elezione, è rischioso per la propria reputazione. Da una parte c'è l'intenzione di assicurare che la previsione fornita sia informativa, dunque non può essere vaga senza motivo valido; dall'altra parte, una previsione precisa nella direzione opposta

risulterebbe essere un errore imbarazzante. Questo *trade-off* si inserisce in un contesto più ampio, in cui gli accademici sono più avversi al rischio di fare un errore e risultano essere più cauti del dovuto nelle loro previsioni, mentre media tradizionali e moderni premiano audacia e contenuti che vanno dritti al punto - spesso troppo velocemente per trasmettere le sfumature di quello che comporta prevedere un'elezione (Gelman & Azari, 2017). Quando i sondaggi e le previsioni ad essi associati sono interpretati con l'adeguata consapevolezza del contesto, allora possono essere uno strumento che incide.

La crisi nella credibilità dei sondaggi allora ha un significato più profondo, e per una certa parte è anche originata da una crisi nel rapporto che intercorre tra pubblico, media ed esperti. Il grande pubblico finisce per scambiare incertezza per incompetenza, mentre gli esperti fanno fatica a tradurre quella che è letteratura statistica in contenuto comunicabile ai più (Gelman, 2021). Questa problematica si acutizza durante lo svolgimento di eventi politici chiave, per i quali la posta in gioco è alta ed un fallimento - o la percezione di esso - può facilmente alimentare scetticismo nei confronti di sondaggi e sondaggisti.

1.3 Le tipologie di errore nei sondaggi elettorali

Costruire un sondaggio d'opinione comporta contattare un certo numero di persone e porre una serie di domande, tra le quali anche chi si intende votare nella prossima elezione. Ci sono diverse ragioni per cui questo processo possa andare storto. I sondaggi elettorali soffrono di una larga varietà di errori di campionamento e non; collettivamente ci si riferisce a questi errori come *total survey error*. Tra queste tipologie di errore, quello di campionamento è il più gestibile: se 1000 persone sono selezionate a caso, allora è inverosimile che le opinioni politiche di questo campione possano corrispondere esattamente all'intera popolazione. Di maggiore importanza, invece, sono i *nonsampling errors*. Anche se, con ogni elezione, i sondaggisti sviluppano tecniche e modelli di previsione sempre più complesse, questo tipo di errori non può essere eliminato, perché le condizioni cambiano di elezione in elezione. Diversamente dall'errore di campionamento, i *nonsampling errors* non possono essere ridotti semplicemente conducendo più sondaggi, oppure contattando più persone.

Esistono almeno quattro componenti di errore all'interno del *nonsampling error* (Shirani-Mehr et al., 2018): di struttura, di non risposta, di misurazione e di specificazione.

L'errore di struttura si verifica quando c'è una discrepanza tra la struttura alla base del processo di campionamento e la popolazione target. Ad esempio, un sondaggio condotto solamente attraverso le interviste telefoniche non riuscirà mai a contattare quelle persone che non possiedono un telefono. Questo problema diventa specialmente rilevante quando si tratta di sondaggi elettorali in quanto la popolazione che quest'ultimi hanno come target - i votanti - è indefinita al momento del campionamento. Infatti le persone che rispondono alle interviste non sono un campione casuale della popolazione dei votanti, in quanto non tutti quelli intervistati e che hanno risposto alla domanda "Chi voterai alle prossime elezioni" poi effettivamente voteranno. Dunque la struttura di campionamento coinvolge una parte di adulti che verosimilmente non andranno ad esprimere il loro voto e pertanto è necessario che i sondaggisti riconoscano questo fenomeno e aggiustino il proprio campione verso la migliore possibile rappresentazione di quello che possa essere la potenziale popolazione di votanti: questo si chiama *turnout modelling*.

L'errore di non risposta si verifica quando i valori mancanti sono sistematicamente legati alla risposta, ad esempio quando un intervistato, sostenitore di un partito che sta andando male nei sondaggi preferisce non rispondere. Questo meccanismo è alla base di quello che in Gelman (2021) e in Gelman et al. (2016) viene identificato come uno dei principali responsabili delle inesattezza dei sondaggi, cioè la non risposta differenziale: i sostenitori di un partito che sta ottenendo buoni risultati sono più propensi a rispondere, mentre quelli di un partito in calo mostrano una minore disponibilità a partecipare.

Non solo, la crescente sfiducia verso i sondaggi, comune a molte delle più grandi democrazie del pianeta, è anch'essa correlata al declino del tasso di risposta - e quindi alla generazione di maggior errore di non risposta (Vandenplas et al., 2018) -, il quale sta diventando un problema sempre più preoccupante anche dal punto di vista economico. É evidente che nel caso in cui si vogliano costruire delle stime di intenzioni di voto su un campione di 1000 unità sarà più caro farlo se si dovranno contattare $10,000$ persone (con un tasso di non risposta del 90%) al posto di sole 3000 (con un tasso di non risposta del 67%).

L'errore di misurazione si verifica quando lo strumento di intervista influisce sulla risposta, per esempio la formulazione della domanda.

L'errore di specificazione avviene quando l'interpretazione di un intervistato ad una domanda differisce differisce da ciò che il ricercatore intende comunicare.

Sia l'errore di misurazione che quello di specificazione contribuiscono al *total survey error*, ma i sondaggi elettorali risultano generalmente meno esposti a queste due fonti di errore (Shirani-Mehr et al., 2018).

Insieme a queste tipologie di errore, si deve tenere conto anche del fatto che un sondaggio è una misura in un punto nel tempo, non una previsione. Se l'intenzione è tipicamente quella di valutare cosa i votanti faranno il giorno delle elezioni, i sondaggi elettorali possono solo misurare direttamente le opinioni correnti e non possono garantire che le persone stiano rispondendo in modo non sincero o che non possano cambiare opinione prima del giorno delle elezioni.

1.4 Aggiustare i dati

Le agenzie demoscopiche tipicamente costruiscono un campione di potenziali elettori, ed in seguito generalizzano le proprie stime campionarie al livello dell'intera popolazione. Come illustrato nella sezione 1.3 esistono diverse componenti di errore da tenere in conto. In particolare gli istituti demoscopici devono salvaguardarsi dal *nonsampling error*, che rappresenta una distorsione per cui stime di quantità riferite alla popolazione sarebbero comunque in media inaccurate anche se il sondaggio fosse ripetuto un grande numero di volte, in quanto l'errore è sistematico e risiede nella tecnica di campionamento.

I sondaggisti fanno del loro meglio per minimizzare questo errore aggiustando i loro dati per differenze note tra il loro campione grezzo e la popolazione. Per fare questo utilizzano tecniche di post-stratificazione. L'approccio più semplice prevede che ad ogni intervistato venga associato un peso numerico, definito in modo tale che le distribuzioni pesate di diverse variabili demografiche (e.g. età, genere, etnia, e redditto) delle unità del campione grezzo corrispondano alle distribuzioni marginali riferite alla popolazione target (Voss et al., 1995). Infatti non tutte le persone hanno la stessa probabilità di essere intervistati. Per esempio, la verosimiglianza della selezione potrebbe variare tra famiglie se una loro caratteristica è correlata con una variabile del sondaggio. Se si fallisce nel correggere per queste diverse probabilità d'inclusione si rischia di distorcere pesantemente le stime legate al campione ottenuto. Di conseguenza, alle persone che fanno parte di gruppi che hanno meno probabilità di essere inclusi nell'insieme degli intervistati, viene assegnato un peso maggiore, e il campione ponderato rappresenta approssimativamente la popolazione.

Pesare le unità del campione del sondaggio è utile in quanto determinate risposte a certe domande dimostrano una correlazione con ampie categorie demografiche.

Quando il campione di un sondaggio d'opinione è sottorappresentativo di una di queste categorie, allora il sondaggista è allertato del fatto che il campione grezzo potrebbe essere inaccurato nello stimare opinioni e preferenze e pertanto è necessario un intervento per aggiustare la situazione.

Un altro problema cruciale è la stima dell'affluenza per sotto-gruppo. Nei sondaggi elettorali l'obiettivo è quello di stimare l'opinione dei votanti, non di tutti i cittadini abilitati al voto. Come detto in Voss et al. (1995) la proporzione di persone che dichiara che andrà a votare è mediamente più grande della proporzione di persone che effettivamente lo faranno. Per questo motivo è necessario determinare la probabilità che un intervistato vada effettivamente a votare.

Ci sono metodi alternativi, oltre all'assegnare pesi, per aggiustare le interviste d'opinione. La *multilevel regression and post-stratification* (MRP), che ora è il metodo standard utilizzato da molte agenzie, è ben spiegata in Gelman (2021).

Tuttavia, esiste un problema pratico: qualsiasi metodologia di adeguamento dei sondaggi - sia esso basato sulla ponderazione o sull'MRP - richiede dati, provenienti sia dal sondaggio che dalla popolazione, non disponibili pubblicamente. Nella maggior parte dei casi infatti vengono pubblicati solo i risultati principali: statistiche riassuntive del sondaggio, senza fornire dati grezzi e spesso con informazioni incomplete su come tali stime sono state prodotte. Questo è il caso del presente studio, in cui l'analisi è limitata a lavorare con le intenzioni di voto dei pariti resi pubblici dalle organizzazioni che hanno condotto il sondaggio.

1.5 Distorsione sistematica e metodi di aggregazione

Fino ad ora ci si è concentrati nello spiegare le possibili fonti di errore legate al condurre un sondaggio elettorale (in sezione 1.3) e nel brevemente citare come i sondaggisti possano lavorare, aggiustando i dati, per mitigarle (in sezione 1.4).

Detto ciò, il sondaggio è un entità singola che però fa parte di un contesto più ampio: esistono diverse agenzie che pubblicano indagini d'opinione (e quindi anche elettorali) e, chiaramente, queste stesse agenzie rilasciano sondaggi con frequenze eterogenee. Quando un'elezione politica si avvicina, allora, si può sfruttare il fatto di poter avere a disposizione più sondaggi elettorali da più istituti demoscopici.

In Shirani-Mehr et al. (2018) si sono concentrati sullo studio di una quantità notevole di sondaggi elettorali riferiti ad altrettanto numerose elezioni avvenute negli Stati Uniti. In particolare, il risultato ufficiale di ciascuna elezione viene assunto come riferimento empirico per valutare la capacità dei sondaggi di anticipare l'esito della

competizione elettorale. L'errore dei sondaggi viene quindi misurato confrontando le stime pre-elettorali con il risultato finale della rispettiva tornata elettorale.

In seguito, decomponendo l'errore in termini di *bias* e varianza, si è riscontrato che i sondaggi per una data elezione condividono fra loro una comune componente di *bias*: una distorsione sistematica. Le ragioni che sottendono questo risultato verosimilmente originano dal fatto che la maggior parte delle indagini di opinione sull'intenzione di voto, anche se condotte da diversi istituti demoscopici, affronta problemi simili e si affida su strutture e metodi affini per risolverli. In primo luogo i sondaggi per una data elezione spesso hanno strutture di campionamento simili e per questo può esistere la difficoltà ad arrivare a contattare gli stessi sotto-gruppi della popolazione trasversalmente a molte agenzie. Non solo, infatti potrebbero affidarsi, nella fase di correzione del proprio campione, a metodi e modelli simili, come quelli utilizzati per aggiustare per la probabilità di voto di un dato intervistato o per l'errore di non risposta.

É comunque da citare il fatto che questo *bias* sistematico non colpisce tutte le quantità che si vuole stimare in modo eguale. Infatti è comune il voler tenere traccia nello svolgersi della campagna elettorale - o semplicemente di una qualsivoglia finestra temporale - dello spostamento dell'umore politico verso un determinato partito, candidato o una certa questione pubblica. In questo caso allora, non si sarebbe tanto interessati al suo supporto in un momento preciso, ma alla sua differenza tra due istanti diversi e pertanto la componente della distorsione si cancellerebbe - assumendo che questa componente sia costante nel tempo.

Come accennato precedentemente un sondaggio elettorale può essere visto come un'individualità, ma anche come parte di un contesto più ampio. Esistono varie metodologie che sfruttano la possibilità di avere a disposizione più sondaggi per una certa finestra temporale. Uno dei metodi più popolari, che deve la sua fortuna alla sua semplicità ed immediatezza è l'aggregazione, che permette di combinare le stime di diversi sondaggi al fine di produrne una singola. Questa strategia è anche chiamata comunemente dai media *Poll of Polls*.

Combinare stime provenienti da più campioni in una certa misura può essere considerato come collezionare un campione più ampio. Allora, ad esempio, si potrebbe pensare semplicemente di fare una media di diversi sondaggi, considerando un certo intervallo di tempo, senza nessun sistema di ponderazione.

Il vantaggio dell'aggregazione risiede nel fatto che la media dei risultati provenienti da diversi istituti tenderebbe a restituire una stima più affidabile e una prospettiva più ampia sull'andamento della competizione elettorale, evitando di fondarsi sul

dato – potenzialmente instabile – di un singolo sondaggio. In particolare, data la variabilità che caratterizza le diverse rilevazioni, ricorrere alla media di più sondaggi permetterebbe di ottenere un quadro della situazione più completo e soprattutto più stabile.

Ma alla luce dei risultati in Shirani-Mehr et al. (2018), è necessaria cautela nell'applicare, interpretare e fidarsi di strumenti come quello appena descritto. Infatti aggregare risultati di diversi sondaggi elettorali non comporta l'eliminazione della componente condivisa di distorsione, che rimane inalterata, anche quando si media su un grande numero di sondaggi. Questo può portare a risultati fuorvianti. Un modello che aggrega è della stessa qualità dei sondaggi di cui è composto. In conclusione è necessario essere circospetti nell'applicare aggregazioni in modo naïve e tenere bene a mente come errori correlati nei sondaggi possono portare a previsioni errate.

In questo senso esistono altri metodi per combinare non solo risultati di più indagini elettorali, ma anche informazioni del contesto in cui si sviluppa l'elezione (e.g. stato dell'economia, come si è votato in precedenti elezioni) che permettono di arrivare sicuramente a risultati migliori. Questi modelli sono di certo strumenti superiori ed infatti in Heidemanns et al. (2020) viene descritto quello sviluppato dall' *Economist* per la previsione delle elezioni presidenziali degli Stati Uniti svoltesi nel 2020.

1.6 Herding

Nella sezione 1.5 si è discusso di come la scomposizione dell'errore in *bias* e varianza mostra che i sondaggi relativi a una stessa elezione tendano a condividere una quota comune di distorsione. Le ragioni di questo fenomeno sono legate al fatto che pur provenendo da istituti demoscopici differenti, le indagini di opinione sull'intenzione di voto affrontano sfide analoghe e si basano su procedure metodologiche affini.

In generale, esiste ed è studiata, però anche una tendenza delle varie agenzie a produrre risultati che si assomigliano, specialmente alla fine di una campagna elettorale e a ridosso delle elezioni. Questo comportamento è denominato *herding* ed è definito dalla American Association for Public Opinion Research come:

> "[...] la possibilità che le agenzie usino i risultati di sondaggi esistenti per aggiustare la presentazione dei loro nuovi risultati. Le strategie di *herding* possono variare dal fare aggiustamenti di carattere statistico per assicurarsi che i risultati rilasciati appaiano simili a quelli dei sondaggi esistenti al decidere

> se pubblicare oppure no un sondaggio in base ai suoi risultati comparati con quelli di indagini già esistenti."

Questa definizione di *herding* comporta che un istituto demoscopico, in una certa misura, si lasci guidare anche dai risultati delle altre agenzie. Tuttavia in Sturgis et al. (2016) si critica questa posizione, in quanto in realtà molti istituti potrebbero assumere questi comportamenti in modo inconsapevole o comunque potrebbero ridurre la varianza dei loro sondaggi in altri modi, senza per forza far riferimento a risultati di altre agenzie. Un esempio può essere il non pubblicare un sondaggio che mostra risultati anomali rispetto a proprie indagini passate, indipendentemente da che risultati riportano gli altri istituti.

Pertanto l'*herding* non implica per forza che gli istituti demoscopici deliberatamente facciano scelte che spingano i propri risultati verso il consenso generale, ma, allo stesso tempo, questa è un'opzione che non si può scartare.

Infatti bisogna sottolineare che i sondaggisti hanno un forte interesse ad essere il più accurati possibile in quanto la loro credibilità è il loro capitale reputazionale.

Pertanto non assumono questo tipo di comportamenti al fine di deliberatamente distorcere i risultati, ma sono spinti dalla paura di un errore nel loro processo metodologico che possa portare ad una sostanziale minore affidabilità percepita rispetto ai diretti concorrenti al momento della verità delle elezioni.

Nondimeno questo processo conduce alla perdita dei vantaggi legati alla "saggezza della folla", in quanto quest'ultimi prescindono dal fatto che le persone agiscano in modo indipendente. Un'agenzia che si limita a pubblicare medie di sondaggi passati non offre alcuna informazione indipendente.

In Sturgis et al. (2016) si illustrano due possibili ipotesi di convergenza verso le quali i sondaggisti potrebbero plausibilmente tendere alla fine del periodo della campagna elettorale. Le agenzie potrebbero convergere verso la media dei risultati dei propri sondaggi passati. Oppure le agenzie potrebbero convergere alla media che coinvolge i sondaggi pubblicati più recentemente. In particolare allinearsi alla media dei sondaggi più recenti sarebbe allettante per un sondaggista che desidera non sbagliare più dei suoi diretti concorrenti.

In entrambi i casi, l'*herding* non è sufficiente a spiegare un' eventuale distorsione sistematica. Nell'ipotesi che i sondaggi non siano distorti, in presenza di *herding* quest'ultimi si allineerebbero verso risultati per l'appunto non distorti. Ma se l'ipotesi non dovesse essere vera, allora questo fenomeno spiegherebbe il perché tutti i sondaggisti finirebbero col dare la stessa risposta sbagliata.

Per concludere sul tema dell'*herding*, è difficile dimostrarne l'esistenza, in quanto è possibile che l'opinione pubblica sia effettivamente stabile quanto suggeriscono i dati. Whiteley (2016) e Silver (2014) suggeriscono che un modo per identificarlo possa essere guardare alla differenza di variabilità tra sondaggi pubblicati ad inizio e alla fine della campagna elettorale per una determinata elezione. L'eterogeneità dei risultati tra le diverse società di sondaggi dovrebbe mantenersi approssimativamente costante per l'intera campagna; tuttavia, se questa si riducesse in modo significativo man mano che la campagna volge al termine, ciò potrebbe costituire un sintomo di *herding*.

Capitolo 2

I dati

In questo capitolo si descrive l'origine dei dati utilizzati nello studio. Essi sono stati messi a disposizione dall'azienda YouTrend, che li utilizza per costruire la loro *Supermedia settimanale*, un aggregatore di sondaggi politici provenienti da vari istituti demoscopici che aggiornano settimanalmente. Il dataset comprende rilevazioni condotte dall'inizio del 2018 fino alla fine del 2024. A queste si aggiungono i sondaggi realizzati da un'ulteriore agenzia, Termometro Politico, circa all'interno della stessa finestra temporale. Per ogni sondaggio si ha disponibile la data di fine rilevazione, la numerosità campionaria e le intenzioni di voto stimate per una pluralità - non costante nel tempo - di partiti politici.

La natura frammentata del panorama politico italiano e la forte eterogeneità nella frequenza di pubblicazione dei sondaggi impongono una serie di scelte soggettive. Per mantenere coerenza comparativa lungo tutto il periodo osservato, il capitolo illustra i criteri adottati per selezionare sia i partiti di interesse — limitati alle principali forze politiche presenti stabilmente dal 2018 al 2024 — sia gli istituti demoscopici, includendo soltanto quelli che dispongono di un numero sufficiente di rilevazioni pre-elettorali per tutte le elezioni considerate.

Una volta definito il perimetro dell'analisi, il capitolo descrive come i dati siano stati organizzati e filtrati per permettere la costruzione di un indice di errore confrontabile tra istituti. Poiché l'obiettivo è quantificare l'affidabilità di una serie di diversi istituti demoscopici, la distorsione dei sondaggi pubblicati rispetto ai risultati ufficiali delle urne assume particolare rilevanza nella definizione della variabile risposta. Quest'ultima deve catturare l'errore di previsione in maniera coerente, comparabile tra partiti con livelli di consenso molto diversi.

Per questa ragione il capitolo introduce l'errore relativo basato sull'ultimo sondaggio disponibile per ogni istituto come indice di misura dell'affidabilità di un'agenzia.

2.1 Origine dei dati

L'analisi di questo studio si è basata principalmente su dati resi disponibili dall'azienda di comunicazione politica YouTrend. Il dataset in questione è quello che utilizzano per costruire la loro *Supermedia settimanale*, un aggregatore di sondaggi che aggiornano e pubblicano sui loro canali settimanalmente. Al suo interno vi si trovano sondaggi condotti nell'intervallo temporale che inizia il 07/01/2018 e finisce il 23/12/2024. Per ogni sondaggio si ha il nome dell'istituto demoscopico che lo ha realizzato, la data - che si riferisce all'ultimo giorno in cui è stato svolto, in quanto i sondaggi tipicamente sono condotti nel corso di più giorni - ed infine la numerosità campionaria.

Ogni sondaggio si compone quindi della stima delle intenzioni di voto per una serie di partiti. Va tuttavia evidenziato che la lista dei partiti oggetto di sondaggio varia nel tempo. Nel periodo 2018–2024 si registrano infatti nuove formazioni politiche, scioglimenti e partiti che compaiono solo per una parte limitata dell'arco temporale.

In questa finestra temporale si sono tenute quattro elezioni: le elezioni politiche del 04/03/2018, le elezioni europee del 26/05/2019, le elezioni politiche del 25/09/2022 e infine le elezioni europee del 09/06/2024. I risultati associati a queste elezioni non sono trattati come osservazioni nel dataset, bensì come verità di fondo delle opinioni della popolazione. L'idea dello studio infatti è quella di valutare l'affidabilità associata ad ogni istituto demoscopico in base al confronto delle proprie previsioni con i risultati delle elezioni corrispondenti. In questo senso, in seguito sarà utile concentrarsi sulla parte di sondaggi pubblicati a ridosso delle elezioni, in particolare sull'ultimo sondaggio disponibile per istituto demoscopico. Prima di questo, però, è necessario fare un controllo su quali sondaggi è possibile sfruttare per l'analisi. Infatti, come previsto dall'art. 7 del Regolamento AGCOM sui sondaggi elettorali — che richiama l'art. 8, comma 1, della legge 22 febbraio 2000, n. 28 — è vietata la diffusione dei sondaggi politici nei quindici giorni precedenti la data delle votazioni. Pertanto si scartano tutti quei sondaggi che non rispettano questa condizione[1].

[1]Per le Politiche del 04/03/2018 si hanno disponibili sondaggi fino allo 02/03/2018. La finestra temporale di sondaggi non pubblicabili parte dal 18/02/2018 e pertanto è necessario scartare ben sedici sondaggi riservati. Mentre per le Politiche del 25/09/2022 si hanno disponibili sondaggi fino al 22/09/2022. La finestra temporale in questo caso parte dall'11/09/2022 e pertanto è necessario scartare ben nove sondaggi riservati.

Oltre al dataset principale fornito da YouTrend, si considerano anche sondaggi condotti nella finestra temporale che inizia il 29/05/2018 e finisce il 25/04/2024 dall'agenzia Termometro Politico. Queste indagini elettorali sono disponibili su sondaggipoliticoelettorali.it.

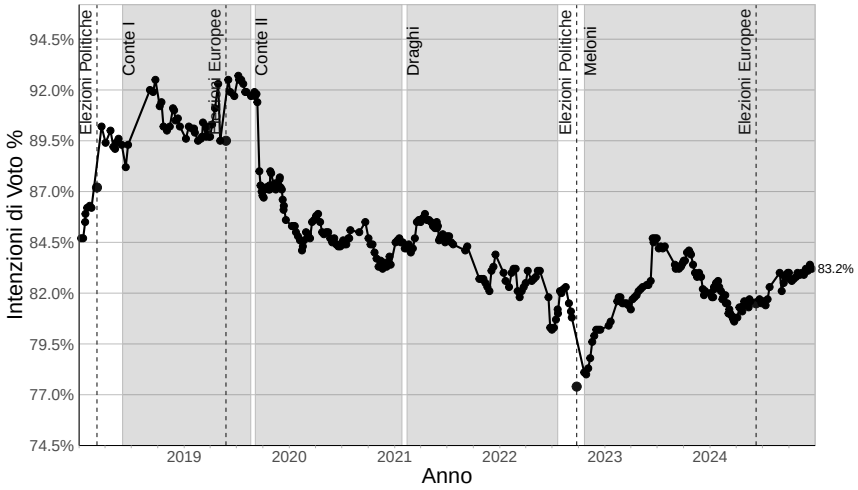
2.2 Partiti politici considerati

Come spiegato nella sezione 2.1 un sondaggio restituisce le percentuali di consenso per i diversi partiti politici e quest'ultimi cambiano nel tempo. É un fatto noto che la realtà politica italiana sia piuttosto frammentata - basti solo pensare a quella americana in cui sostanzialmente sono presenti due soli partiti. Ne deriva un panorama in cui nuove formazioni politiche emergono, altre scompaiono e alcune hanno cicli di vita particolarmente brevi e nella finestra temporale considerata diversi partiti sono nati e dissolti, oppure hanno subito fusioni, scissioni o cambiamenti di denominazione. Questa volatilità rende difficile mantenere un insieme omogeneo di soggetti politici lungo tutto il periodo d'analisi.

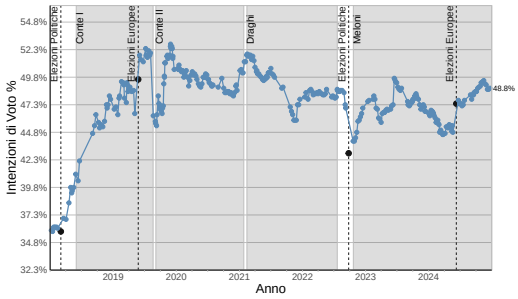
Alla frammentazione si aggiunge il ruolo delle coalizioni, spesso costituite in modo variabile a seconda della specifica elezione. In alcuni casi i sondaggi rilevano l'intenzione di voto per liste aggregate o coalizzate, mentre in altre occasioni gli stessi partiti vengono rilevati singolarmente. Questa discontinuità introduce ulteriori difficoltà nel confrontare serie storiche in modo coerente.

Per evitare incoerenze e ridurre la complessità introdotta da partiti poco importanti o da rilevazioni non uniformi, si decide quindi di restringere l'attenzione ai partiti che hanno assunto un ruolo politico significativo e continuo nel periodo considerato. Nello specifico si considerano solo i cinque partiti maggiori nella scena politica italiana dal 2018 al 2024 cioè Fratelli d'Italia, Forza Italia, Lega, Movimento 5 Stelle e Partito Democratico. Questa scelta consente di lavorare su un insieme stabile e comparabile di soggetti politici, migliorando la qualità e l'interpretabilità delle analisi.

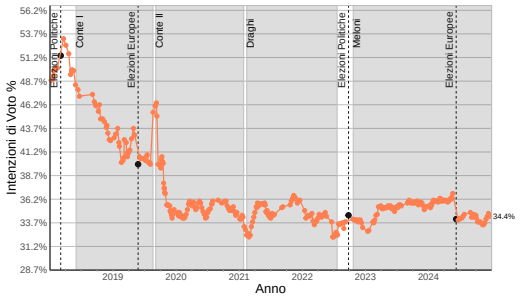
In media, i cinque partiti citati rappresentano l'$83,6\%$ del totale dei consensi rilevati tra il 2018 e il 2024 tenendo in considerazione i sondaggi di tutte le agenzie disponibili. Il valore minimo registrato è pari al 74%, osservato dall'istituto Noto il 31/07/2022, mentre il massimo raggiunge il 94%, rilevato sempre da Noto il 25/06/2019. Questi estremi mostrano come, pur nella forte frammentazione del sistema politico italiano, una parte molto consistente dell'elettorato, e di sicuro maggioritaria, si concentra stabilmente su un gruppo ristretto di forze politiche.



(a) Serie storica del consenso dei 5 partiti



(b) Serie storica del consenso del centro-destra (FDI, FI e LEGA)



(c) Serie storica del consenso del campo largo (PD e M5S)

Figura 2.1: Serie storiche delle intenzioni di voto rilevate dall'agenzia Tecnè. Il pannello **(a)** rappresenta il consenso complessivo dei cinque partiti analizzati. I pannelli **(b)** e **(c)** riportano le corrispondenti serie ottenute aggregando i partiti per area politica: centrodestra (Fratelli d'Italia, Forza Italia e Lega) e campo largo (Partito Democratico e Movimento 5 Stelle).

In Figura 2.1 si visualizza ciò che è stato appena spiegato solo considerando i sondaggi pubblicati dall'istituto demoscopico Tecnè, cioè che i cinque partiti scelti hanno rappresentato per tutta la durata della finestra temporale a disposizione per lo studio la maggior parte del consenso elettorale in Italia.

2.3 Istituti demoscopici considerati

L'idea alla base dell'analisi è cercare di attribuire una valutazione all'affidabilità associata a ciascun istituto demoscopico sulla base della qualità dei suoi sondaggi, definita come la capacità di prevedere correttamente le intenzioni di voto della popolazione in occasione di eventi elettorali. In altre parole, si intende confrontare le

stime prodotte dagli istituti con i risultati reali ottenuti dalle forze politiche nelle varie competizioni elettorali.

Non tutte le agenzie, però, dispongono di una serie storica sufficientemente estesa e regolare. Alcune hanno iniziato a pubblicare sondaggi solo in anni recenti, altre esattamente il contrario, mentre altre ancora hanno serie molto discontinue. Questo comporta che non tutti gli istituti possano essere valutati in relazione a tutte le elezioni, poiché per alcune competizioni semplicemente non esistono rilevazioni antecedenti su cui basare il confronto.

Sono presenti anche istituti che, pur coprendo formalmente l'intera finestra temporale considerata, hanno pubblicato sondaggi raramente. In questo caso la scarsità di osservazioni rende poco informativa la stima della loro accuratezza pre-elettorale: un istituto che pubblica uno o due sondaggi all'anno non può fornire una base solida per una valutazione affidabile.

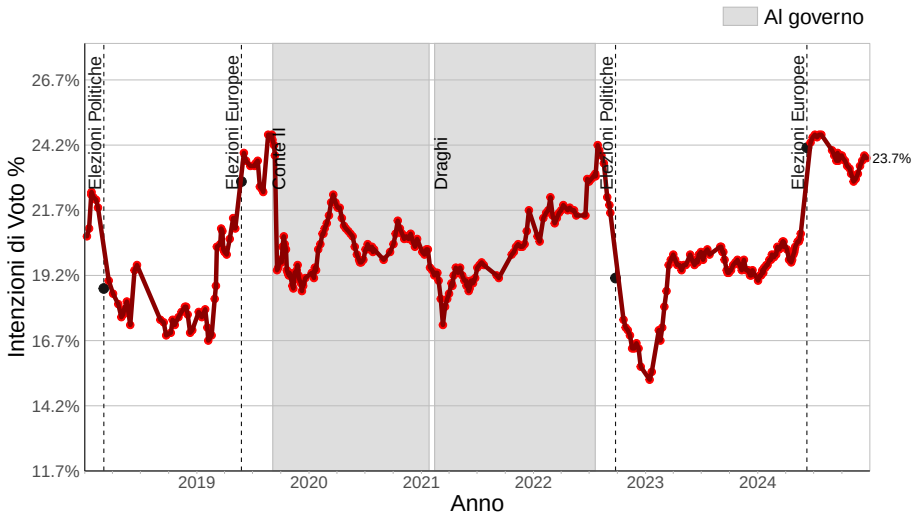


Figura 2.2: Serie Storica Consenso PD sull'intero arco temporale. La parte di sfondo colorato in grigio indica quando il partito è stato al governo.

Intervallo	**Numero rilevazioni**
Politiche 18 – Europee 19	41
Europee 19 – Politiche 22	171
Politiche 22 – Europee 24	93

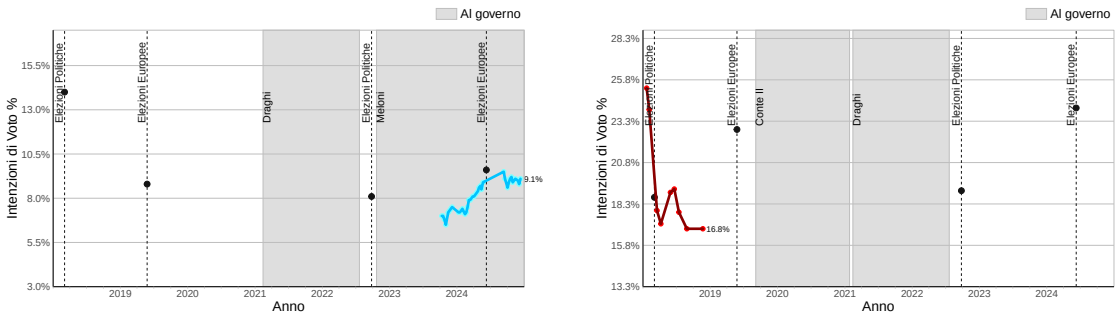
Tabella 2.1: Numero di rilevazioni per Tecnè per ciascun intervallo fra elezioni

Per procedere con uno studio coerente, è necessario costruire un confronto bilanciato, in cui vi sia lo stesso numero di combinazioni elezione-partito per ogni agenzia. Per questa ragione s'individuano come appropriate quegli istituti demoscopici

che presentano almeno una rilevazione precedente alle elezioni Europee del 2019, alle elezioni Politiche del 2022 e alle elezioni Europee del 2024. Non solo, questi istituti devono garantire una copertura temporale sufficiente per effettuare una valutazione comparabile della loro capacità predittiva nel corso delle varie competizioni.

In Figura 2.2 è riportata la serie storica delle intenzioni di voto stimate da Tecnè per il Partito Democratico sull'intero intervallo temporale considerato. Si osserva chiaramente la presenza di rilevazioni antecedenti a ciascuna delle tre elezioni analizzate. La tabella 2.1 riporta inoltre il numero di sondaggi pubblicati nei singoli intervalli tra un'elezione e la successiva: complessivamente, tra le elezioni del 2019 e quelle del 2024 risultano disponibili 305 rilevazioni.

Tecnè rappresenta dunque un esempio di istituto virtuoso che soddisfa pienamente i criteri individuati per l'inclusione nell'analisi. Applicando la medesima procedura di verifica a tutti gli altri istituti - per i quali si rimanda ai grafici riportati in appendice A.1 - si procede all'esclusione dei seguenti: Cise, Demos, Eumetra e Lorien.



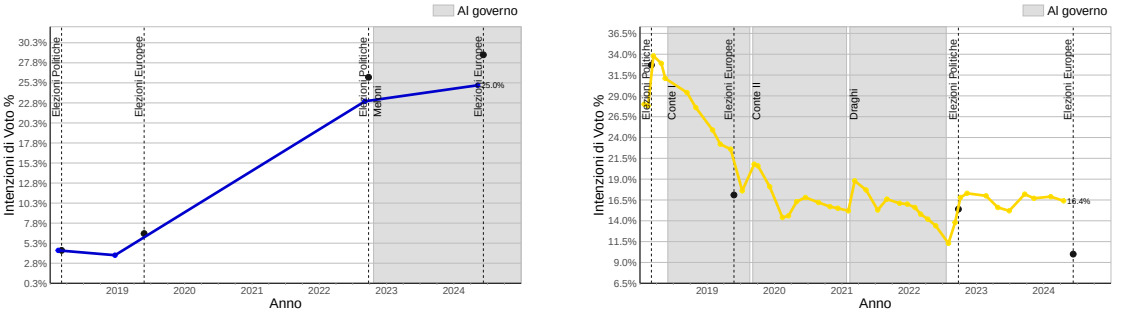
(a) Serie Storica Consenso FI per Eumetra **(b)** Serie Storica Consenso PD per Lorien

Figura 2.3: Serie Storica delle intenzioni di voto rese pubbliche dal 2018 alla fine del 2024 da **(a)** Eumetra per Forza Italia e **(b)** Lorien per il Partito Democratico

Istituto Demoscopico	Numero rilevazioni
Cise	3
Demos	41

Tabella 2.2: Numero di rilevazioni per le agenzie Cise e Demos nell'intervallo di tempo che va dall'elezione politica del 2018 fino all'elezione europea del 2024

Per quanto riguarda gli istituti demoscopici Eumetra e Lorien, in Figura 2.3, si hanno rilevazioni solo su un periodo di tempo ridotto rispetto all'intera finestra disponibile. La prima rilevazione pubblicata da Eumetra risale al 11/10/2023 e l'ultima al 11/12/2024. Mentre per Lorien la prima risulta in data 21/01/2018 e l'ultima in data 22/11/2018. In entrambi i casi non si hanno rilevazioni antecedenti ad



(a) Serie Storica Consenso FDI per Cise **(b)** Serie Storica Consenso M5S per Demos

Figura 2.4: Serie Storica delle intenzioni di voto rese pubbliche dal 2018 alla fine del 2024 da **(a)** Cise per Fratelli d'Italia e **(b)** Demos per il Movimento 5 Stelle

alcune delle elezioni incluse nell'analisi, pertanto è necessario rimuoverle dall'insieme delle agenzie partecipanti allo studio.

In relazione agli istituti Cise e Demos in figura 2.4, invece, si ha almeno un'osservazione prima di ognuna delle tre elezioni facenti parte dell'analisi. Nonostante ciò si è deciso di escluderle dallo studio in quanto, come mostrato in tabella 2.2, si è giudicato insufficiente il numero di sondaggi resi pubblici nell'arco temporale d'interesse.

In tabella 2.3 si mostra il nome degli istituti demoscopici inclusi nello studio associati al numero di sondaggi pubblicati nella finestra temporale dell'analisi. In totale lo studio si basa su 1676 sondaggi completati da undici agenzie diverse a partire dal 04/03/2018 fino al 24/05/2024.

Istituto	n
demopolis	81
emg	189
euromedia	143
ipsos	88
ixè	89
noto	102
piepoli	108
quorum	77
swg	280
tecnè	305
termometro politico	210

Tabella 2.3: Numero dei sondaggi pubblicati per Istituto Demoscopico dall'elezione politica del 04/03/2018 fino all'elezione europea del 24/05/2024

2.4 Descrizione dei dati

Questa tesi si propone di attribuire una valutazione all'affidabilità, definita come capacità di prevedere correttamente le intenzioni di voto della popolazione, associata ad ogni istituto demoscopico incluso nell'analisi. In questo modo s'intende anche esaminare l'eterogeneità di questa affidabilità, ovvero se gli istituti abbiano un livello di credibilità simile fra loro oppure no.

Come spiegato nelle sezioni 2.2 e 2.3 si hanno a disposizione i risultati di sondaggi per Fratelli d'Italia, Forza Italia, Lega, Movimento 5 Stelle e Partito Democratico in un arco temporale che inizia con l'elezione politica del 2018 e finisce con l'elezione europea del 2024 per undici istituti demoscopici. Per ogni sondaggio condotto, insieme all'intenzione di voto stimate, si hanno la data e la numerosità campionaria. Quest'ultima è, nello studio in questione, ignorata. Le ragioni di questa scelta sono da ricercarsi nel capitolo 1. É vero che fra due sondaggi, identici nelle metodologie utilizzate per la loro realizzazione, quello che gode di una numerosità campionaria maggiore possiederà un errore di campionamento minore. Nonostante ciò il *total survey error* comprende anche la componente di *nonsampling error* - in media tanto grande quanto l'errore di campionamento (Gelman, 2021) che non può essere ridotta semplicemente conducendo interviste a più persone.

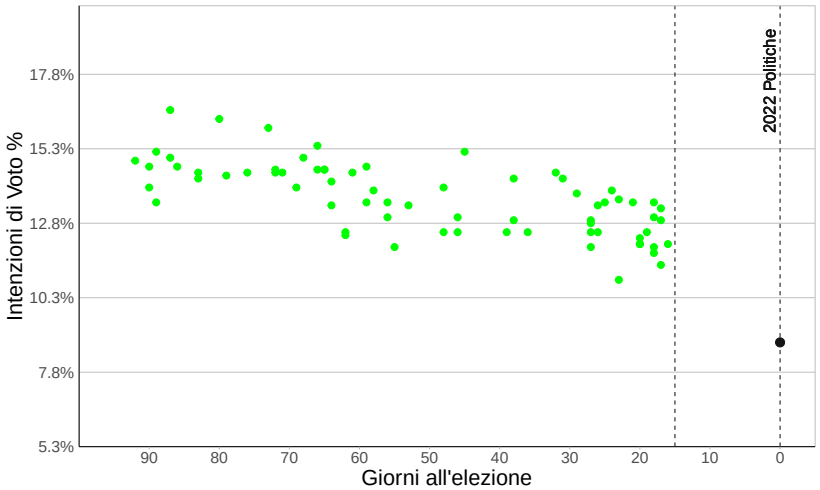


Figura 2.5: Intenzioni di Voto stimate per la Lega dalle varie agenzie nel corso della campagna elettorale per l'elezione politica del 25/09/2022. La prima linea verticale tratteggiata indica la soglia entro cui poi non è più possibile pubblicare indagini politiche, mentre la seconda indica il giorno dell'elezione.

Con l'analisi si valuta empiricamente e sistematicamente la capacità di un'agenzia nella previsione delle intenzioni di voto di una popolazione confrontando i sondaggi

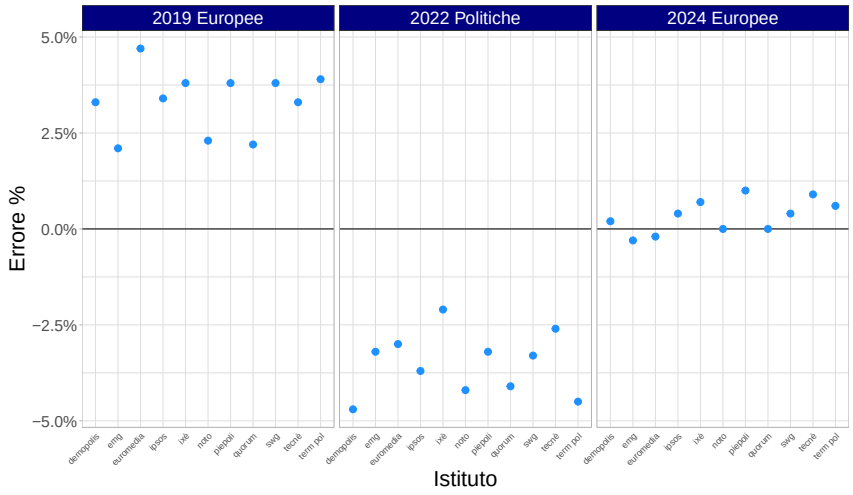


Figura 2.6: Differenza tra il risultato delle elezioni corrispondenti e le intenzioni di voto stimate dall'ultimo sondaggio disponibile per agenzia per la Lega.

pubblicati a ridosso delle elezioni - Europee del 2019, Politiche del 2022 e Europee del 2024 - con l'esito delle stesse che viene assunto come verità nei riguardi delle preferenze politiche degli elettori. Per questo, in realtà, sarebbe necessario focalizzare l'attenzione solo su un intervallo di tempo relativamente vicino alle elezioni imminenti. In figura 2.5 sono riportati tutti i sondaggi pubblicati dalle undici agenzie selezionate per la Lega negli ultimi tre mesi prima della votazione alle Politiche del 2022. Si ricordi che per legge in Italia non è possibile rendere pubblici sondaggi condotti nei quindici giorni precedenti alle elezioni per questo specificatamente a ridosso dell'elezione non si trovano osservazioni. Dalla figura 2.5 emerge chiaramente come le varie agenzie demoscopiche tendano a proporre stime molto simili tra loro, evidenziando anche una specifica dinamica temporale nell'intenzioni di voto della Lega. Questi aspetti suggeriscono in una qualche misura una convergenza o quantomeno una lettura simile dello scenario politico a ridosso dell'elezione.

Dal grafico in figura 2.6 è ancora più immediato cogliere queste intuizioni. Infatti prendendo in considerazione l'ultimo sondaggio disponibile per le agenzie prima dell'elezione successiva, si nota come quest'ultime non si distinguano in modo netto per errore commesso - calcolato in questo caso come differenza tra risultato delle elezioni e stima delle intenzioni di voto - nel prevedere il partito Lega [2].

La lettura del grafico evidenzia una dinamica chiara ed uniforme. Per la Lega si osserva che la differenza tra il risultato delle elezioni europee del 2019 e l'ultimo sondaggio pre-elettorale disponibile sia maggiore di zero per tutte le agenzie. Ciò sta a significare che gli istituti demoscopici hanno sottostimato il partito. Una

[2]Questo vale anche per gli altri partiti, si vedano in i grafici aggiuntivi in appendice A.2

dinamica speculare si osserva per le politiche del 2022, in quanto l'errore è, in questo caso, minore di zero per ogni agenzia. Questo significa che tutti gli istituti hanno sovrastimato il partito circa della stessa intensità. Infine per l'elezione europea del 2024 le differenze sono tutte distribuite attorno allo zero con agenzie che hanno sottostimato leggermente il partito, quando altri, invece, l'hanno sovrastimato in misura altrettanto ridotta.

Complessivamente i grafici in figura 2.5 e figura 2.6 suggeriscono che non esistono differenze significative tra le undici agenzie analizzate in termini di affidabilità. Gli errori tendono a seguire la medesima direzione in ogni elezione. Questa uniformità suggerisce che la capacità predittiva non sia imputabile alla singola agenzia, bensì al contesto politico ed elettorale del momento, oltre che all'individualità del partito. Il comportamento degli istituti appare quindi omogeneo e soprattutto non discriminante in termini di accuratezza.

2.5 La scelta della variabile risposta

Alla luce delle riflessioni compiute nella sezione precedente 2.4 si procede con la valutazione dell'affidabilità dei diversi istituti demoscopici confrontando, per ciascuna elezione, l'ultimo sondaggio pubblicato da ogni agenzia con il risultato effettivo ottenuto dal partito in questione. L'assunzione alla base di questo confronto è che proprio l'ultima rilevazione disponibile per ogni istituto dovrebbe approssimare con maggiore precisione il risultato delle elezioni prossime.

Tuttavia, per rendere l'analisi metodologicamente corretta, è necessario definire adeguatamente la variabile risposta che rappresenta la distorsione, quindi l'errore commesso da un sondaggio. L'uso diretto della differenza tra il risultato elettorale e la percentuale stimata risulta non idonea, in quanto non comparabile tra partiti diversi.

Un errore di identica ampiezza ha un peso completamente diverso a seconda della dimensione del consenso del partito considerato. Ad esempio sbagliare di cinque punti percentuali un partito stimato attorno al 20% nei sondaggi ha un impatto relativo più contenuto rispetto allo sbagliare della stessa quantità un partito che si colloca al 10%. La semplice differenza non tiene conto della proporzione dell'errore rispetto alla dimensione della grandezza che si sta cercando di prevedere.

Questa asimmetria rende indispensabile la scelta di una variabile risposta che esprima non solo quanto lontana sia la previsione dal risultato delle elezioni, ma anche quanto quell'errore pesi in termini relativi. Una variabile costruita in questo modo

consente di confrontare con coerenza la distorsione legati agli istituti nei confronti di partiti anche molto diversi tra loro per dimensioni elettorali.

Sia dunque $i = 1, ..., 11$ l'indice legato all'agenzia, $k = 1, ..., 5$ l'indice per il partito e $j = 1, 2, 3$ l'indice associato al turno elettorale.

Si indica con:

- V_{kj} il valore osservato (risultato ufficiale alle urne) del partito k all'elezione j;
- $\hat{V}_{ikj}$ il valore stimato dall'istituto i per il partito k all'elezione j, ottenuto dall'ultimo sondaggio disponibile prima del voto;

La variabile risposta viene definita come:

$$y_{ikj} = \frac{V_{kj} - \hat{V}_{ikj}}{\hat{V}_{ikj}}. \tag{2.1}$$

La formulazione della distorsione in (2.1), ovvero l'errore relativo, consente un'interpretazione immediata dello scarto tra previsione e risultato elettorale. La normalizzazione rispetto al valore stimato rende la misura indipendente dalla scala del partito considerato e permette quindi di confrontare in modo omogeneo errori riferiti a forze politiche con livelli di consenso molto diversi. Un'ulteriore caratteristica utile di questa definizione è che l'indice così costruito assume valori su tutto $\mathbb{R}$: un valore positivo segnala che l'istituto ha sottostimato il risultato del partito, mentre un valore negativo indica una sovrastima.

A ulteriore conferma del fatto che la scelta della variabile risposta sia in parte soggettiva — e che sarebbero state possibili anche altre formulazioni — si introduce un indice alternativo: l'errore logaritmico, definito come

$$y_{ikj} = \log\left(\frac{V_{kj}}{\hat{V}_{ikj}}\right). \tag{2.2}$$

Questa misura condivide le principali proprietà desiderabili dell'errore relativo precedentemente definito. In primo luogo, non dipende dalla scala del fenomeno osservato: confronta infatti il rapporto tra valore osservato e valore stimato, risultando quindi adatta a comparare errori associati a partiti con dimensioni elettorali molto diverse. Inoltre, come per l'indice precedente, anche l'errore logaritmico è definito su tutto $\mathbb{R}$: valori positivi indicano una sottostima da parte dell'agenzia, mentre valori negativi corrispondono a una sovrastima.

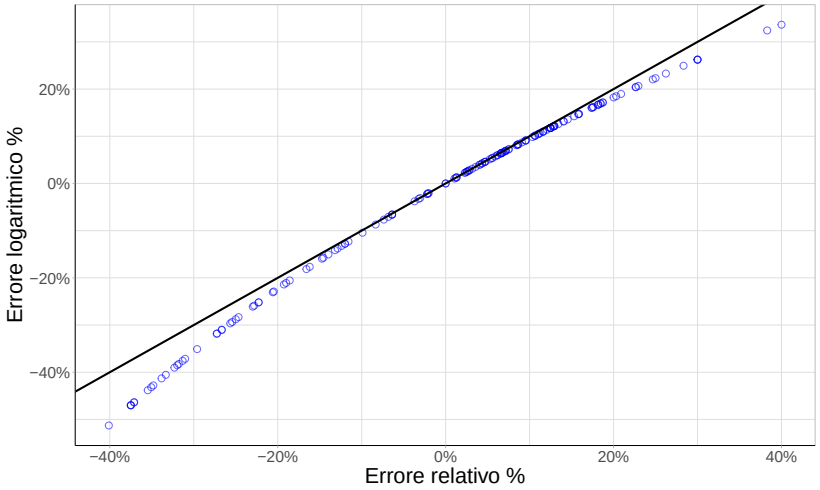


Figura 2.7: Relazione tra Errore Logaritmo e Errore Relativo per tutte le combinazioni di agenzia-partito-elezione. La linea nera rappresenta la bisettrice del primo-terzo quadrante.

Nonostante la diversa formulazione matematica, i due indici risultano empiricamente molto simili[3]. Il grafico in Figura 2.7 mostra la relazione tra i due indici - errore relativo ed errore logaritmico - per ciascuna combinazione di istituto-partito-elezione. I punti blu rappresentano le osservazioni empiriche mentre la linea nera la bisettrice del primo e terzo quadrante.

Si nota chiaramente come la maggior parte delle osservazioni si dispongano vicino alla bisettrice, implicando che nell'intervallo degli errori presenti nei dati, l'errore logaritmico e l'errore relativo forniscono valutazioni simili della distorsione.

Le differenze che emergono sono minime e comunque insufficienti da modificare in modo sostanziale le conclusioni che si possono trarre da un'analisi basata solamente sull'errore relativo.

In conclusione l'errore relativo definito in equazione (2.1) rappresenta quindi una scelta adeguata e sufficiente per lo scopo, pur sapendo che l'errore logaritmico in equazione (2.2) costituirebbe un'alternativa sostanzialmente equivalente.

Il dataset, risultato delle aggregazioni e assunzioni sviluppate, utilizzato per l'analisi sull'affidabilità statistica delle varie agenzie nel prevedere correttamente le performance politiche delle maggiori forze in campo da dopo l'elezione politica del 2018 fino all'elezione europea del 2024 è composto dalle tre variabili categoriche indipendenti `Istituto`, `Partito` ed `Elezione` e dalla variabile risposta definita come in equazione (2.1). In totale si hanno a disposizione 165 osservazioni, corrispondenti alle combinazioni di agenzia-partito-elezione.

[3]Il primo è l'espansione di Taylor del secondo

Tuttavia, l'utilizzo dell'ultimo sondaggio disponibile come unica fonte informativa per costruire il valore previsto presenta dei limiti.

Infatti utilizzare un singolo punto - l'ultimo - della serie storica ignora l'evoluzione temporale delle intenzioni di voto e non tiene conto della dinamica dei sondaggi nel tempo, specie in prossimità delle elezioni. Quest'ultima può fornire informazioni utili per stimare più accuratamente il supporto effettivo dei partiti. Un istituto che rileva un trend discendente o crescente nelle settimane precedenti potrebbe essere penalizzato o favorito artificialmente dal fatto che l'ultimo sondaggio disponibile non rifletta pienamente tale andamento.

Per mitigare questi limiti e ottenere una misura di distorsione maggiormente robusta, si considera quindi un'alternativa nella definizione del valore previsto utilizzata nelle equazioni (2.2) e (2.1). Invece di basarsi sull'ultima rilevazione pubblicata, si utilizza una stima interpolata della serie dei sondaggi fino al giorno dell'elezione, ottenuta tramite il Filtro di Kalman (Kalman, 1960). Questo approccio consente di incorporare in maniera coerente l'informazione contenuta nell'intera traiettoria dei sondaggi. Il capitolo 3 si concentra nello spiegare come si arriva a tale stima.

L'analisi condotta con questa seconda definizione del valore previsto permette dunque di verificare se le conclusioni tratte sull'affidabilità degli istituti dipendano dalla particolare scelta del punto di previsione o se risultino invece solide rispetto a una metodologia più sofisticata e informativamente ricca.

Capitolo 3

Filtro di Kalman

In questo capitolo si spiega il processo metodologico utilizzato per costruire l'evoluzione temporale delle intenzioni di voto di ogni partito considerato a partire dai sondaggi pubblicati dalle agenzie demoscopiche selezionate. Gli strumenti centrali sono due: il modello *state space*, che offre la possibilità di lavorare all'interno di un paradigma estremamente flessibile e permette di affrontare una ampia serie di problemi nell'analisi di serie storiche, e il filtro di Kalman che è l'algoritmo che consente poi di stimare le intenzioni di voto di un determinato partito in un determinato periodo in modo coerente.

Il capitolo si apre con una breve introduzione ai modelli statistici in forma *state space* e ai modelli lineari dinamici Gaussiani, presentando la formulazione generale del modello e un esempio di componente non osservata, il *local linear trend*. Si passa quindi alla derivazione del filtro di Kalman e si spiegano alcune sue peculiarità: il trattamento dei valori mancanti, la previsione di valori futuri, l'inizializzazione diffusa del filtro e la stima di massima verosimiglianza di eventuali parametri ignoti. Successivamente si introducono i modelli *state space* non lineari e non Gaussiani, con particolare attenzione ai casi in cui le osservazioni appartengono alla famiglia esponenziale, come la distribuzione Binomiale, rilevante quando si tratta di sondaggi. In questo contesto si introduce la pratica del filtro approssimato basato sulla stima della moda.

Nella parte finale del capitolo, la teoria sviluppata viene applicata ai dati descritti nel capitolo 2, specificando un modello *state space* binomiale con *local linear trend* per le intenzioni di voto di ciascun partito e istituto demoscopico. Le stime ottenute nei giorni di elezione costituiscono infine i valori previsti utilizzati per definire una misura alternativa di errore di previsione, che andrà a partecipare al dataset analizzato per valutare l'affidabilità statistica degli istituti demoscopici.

3.1 Introduzione all'analisi state space

Rappresentare un modello statistico in forma *state space* offre la possibilità di lavorare all'interno di un paradigma estremamente flessibile che permette di affrontare una ampia serie di problemi nell'analisi di dati sia in formato cross-sezionale che serie storiche. Gli *unobserved component models* (UCM), i modelli ARIMA, la regressione lineare, i modelli ad effetti misti sono solo alcuni esempi di modelli statistici che possono essere espressi in forma *state space*.

In questa tesi si lavora con gli UCM, una classe di modelli che assume che l'evoluzione temporale del sistema studiato sia determinata da una sequenza di componenti non osservate - e non osservabili - $\alpha_1, ..., \alpha_n$, alle quali è associata una sequenza di osservazioni $y_1, ..., y_n$. La relazione tra gli α_t e le y_t è regolata proprio attraverso il modello *state space*.

Il principale obiettivo dell'analisi *state space* è quello di inferire su proprietà d'interesse degli α_t avendo a disposizione solamente la conoscenza delle osservazioni $y_1, ..., y_n$. Ma è possibile anche indirizzare la propria attenzione su altri compiti fondamentali, come la previsione di valori futuri, la ricostruzione di valori mancanti e la stima di parametri tramite metodi come la massima verosimiglianza.

I primi passi dell'impostazione all'analisi *state space* non sono stati fatti nel campo della statistica, bensì in quello dell'ingegneria con l'articolo pioneristico di Kalman (1960). I principi che hanno fatto la fortuna dei modelli *state space* sono due e sono cruciali: un'ampia classe di problemi può essere espressa in forma *state space*; in virtù della natura markoviana del modello, i calcoli necessari per applicazioni pratiche possono essere impostati in forma ricorsiva in maniera molto conveniente per un computer.

3.2 Modelli lineari state space

Una delle classi di modelli *state space* più semplici è quella legata ai modelli lineari generali *state space* Gaussiani (Durbin & Koopman, 2012) - anche conosciuti con il nome di modelli lineari dinamici. Il modello lineare generale *state space* Gaussiano può essere scritto come:

$$\begin{aligned} y_t &= Z_t\alpha_t + \epsilon_t, \qquad \epsilon_t \sim N(0, H_t), \\ \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t, \quad \eta_t \sim N(0, Q_t), \quad t = 1, \ldots, n, \end{aligned} \tag{3.1}$$

con y_t che è un vettore $p \times 1$ che contiene le osservazioni all'istante temporale t, mentre α_t che è un vettore $m \times 1$ che contiene le componenti non osservate della serie al tempo t.

La prima equazione di (3.1) è detta equazione di misurazione, mentre la seconda è chiamata equazione di stato. Il modello si appoggia sull'intuizione che lo sviluppo del sistema nel tempo sia determinato dal vettore α_t in base alla seconda equazione in (3.1), ma, poiché α_t risulta non osservabile direttamente, risulta necessario affidarsi esclusivamente alle osservazioni y_t. Le matrici Z_t, T_t, e R_t, insieme con le matrici di varianza e covarianza H_t e Q_t dipendono dalla particolare definizione del modello e spesso sono costanti nel tempo, cioè non dipendono da t. Tipicamente alcune di queste matrici contengono qualche parametro ignoto da stimare.

Si assume che i termini d'errore ϵ_t e η_t siano serialmente indipendenti e indipendenti l'un l'altro in ogni istante temporale. Inoltre si assume anche che il vettore di stato all'istante $t = 1$ sia distribuito come una $N(a_1, P_1)$ e che sia indipendente da $\epsilon_1, \ldots, \epsilon_n$ e $\eta_1, \ldots, \eta_n$.

La prima equazione in (3.1) ha la struttura di una regressione lineare in cui il vettore degli stati α_t varia nel tempo. La seconda equazione invece rappresenta un modello AR(1), la cui natura markoviana conferisce molte delle proprietà eleganti legate ai modelli *state space*.

3.2.1 Local linear trend

L'approccio moderno allo studio delle serie storiche si basa sull'assumere che una serie storica sia la somma di componenti non direttamente osservabili come trend, stagionalità e ciclo. Queste componenti possono essere viste come versioni stocastiche di funzioni deterministiche del tempo. Dunque uno statistico, quando lavora con una serie storica, deve per prima cosa identificare e stimare un modello stocastico che ben approssimi il meccanismo che genera i dati. Il modello UCM che viene utilizzato in questo studio è caratterizzato da una sola componente non osservabile, quella del trend, modellato come *local linear trend* (LLT). Dunque il seguente è indicato come modello *local linear trend*:

$$\begin{aligned} y_t &= \mu_t + \epsilon_t, & \epsilon_t &\sim N(0, \sigma_\epsilon^2), \\ \mu_{t+1} &= \mu_t + \nu_t + \xi_t, & \xi_t &\sim N(0, \sigma_\xi^2), \\ \nu_{t+1} &= \nu_t + \zeta_t, & \zeta_t &\sim N(0, \sigma_\zeta^2) \end{aligned} \tag{3.2}$$

con μ_t che è il livello della serie al tempo t, mentre ν_t è l'incremento del livello dall'istante t a quello successivo $t+1$. Si può riscrivere il modello (3.2) in forma matriciale come:

$$y_t = \begin{pmatrix} 1 & 0 \end{pmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \epsilon_t,$$
$$\begin{pmatrix} \mu_{t+1} \\ \nu_{t+1} \end{pmatrix} = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \mu_t \\ \nu_t \end{pmatrix} + \begin{pmatrix} \xi_t \\ \zeta_t \end{pmatrix},$$

ed è facile notare che quest'ultimo risulta essere un caso speciale di (3.1).

3.3 Filtro di Kalman

In questa sezione si fornisce una panoramica sul modello lineare *state space* Gaussiano (3.1). Le osservazioni y_t sono trattate come multivariate. In quanto i concetti esposti in questa sezione sono trattati nella loro completezza in Durbin & Koopman (2012), la notazione utilizzata è sostanzialmente identica a quella usata da Durbin e Koopman.

In sotto-sezione 3.3.1 si presenta il Lemma 3.1: si considera una coppia di vettori casuali distribuiti congiuntamente come una Normale e si mostra che la distribuzione condizionata di x dato y risulta ancora essere Normale. Quindi si deriva il suo vettore delle medie e la sua matrice di covarianza. Il Lemma 3.1 è cruciale in quanto rappresenta il fondamento teorico su cui i calcoli del filtro di Kalman poggiano.

Per approfondimenti nel caso in cui non si volesse lavorare in un contesto in cui si assume normalità oppure si preferisca muoversi in un quadro bayesiano piuttosto che classico si legga Durbin & Koopman (2012). Vengono infatti introdotte giustificazioni teoriche che supportano la derivazione del filtro di Kalman anche in quelle circostanze.

Si denoti l'insieme di osservazioni $y_1, \ldots, y_t$ con Y_t. In sotto-sezione 3.3.2 si derivano le equazioni del filtro di Kalman, che è un algoritmo ricorsivo incentrato nel calcolare $a_{t|t} = \mathbb{E}(\alpha_t \mid Y_t)$, $a_{t+1} = \mathbb{E}(\alpha_{t+1} \mid Y_t)$, $P_{t|t} = \mathrm{Var}(\alpha_t \mid Y_t)$ e $P_{t+1} = \mathrm{Var}(\alpha_{t+1} \mid Y_t)$ dati a_t e P_t. Per il contesto di questa tesi la derivazione di questi calcoli necessita solamente dei risultati ottenuti con il Lemma 3.1.

In sotto-sezione 3.3.3 si affronta il problema delle osservazioni mancanti e si mostra come con l'approccio *state space* la questione è semplicemente risolta attraverso piccole modifiche nelle equazioni del filtro di Kalman. Non solo, infatti si mostra anche come la previsione di osservazioni future e di stati futuri può essere semplicemente ottenuta trattando osservazioni future come osservazioni mancanti. Quest'ultimo è un risultato particolarmente d'impatto nel lavoro pratico con le serie storiche. Infine

le ultime due sotto-sezioni, 3.3.4 e 3.3.5, introducono le questioni dell'inizializzazione del filtro e della stima di massima verosimiglianza dei suoi parametri ignoti.

3.3.1 Proprietà condizionali della Normale multivariata

Si supponga che x ed y siano vettori casuali distribuiti congiuntamente come una Normale con

$$\mathbb{E}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \mu_x \\ \mu_y \end{pmatrix}, \qquad \mathrm{Var}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy}^\top & \Sigma_{yy} \end{pmatrix} \tag{3.3}$$

con Σ_{yy} che si assume sia una matrice non singolare.

Lemma 3.1. *La distribuzione condizionata di x dato y è Normale con vettore delle medie*

$$\mathbb{E}\big[x \mid y\big] = \mu_x + \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \tag{3.4}$$

e matrice delle varianze e covarianze.

$$\mathrm{Var}(x \mid y) = \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{xy}^\top. \tag{3.5}$$

La dimostrazione è disponibile in appendice B.1. Si noti che la varianza condizionata definita in (3.5) non dipende da un particolare valore di y al quale x dovrebbe essere condizionato. Questa proprietà è esclusiva della distribuzione Normale ed in generale non si allarga ad altre distribuzioni.

3.3.2 Derivazione del filtro di Kalman

In riferimento al modello definito in (3.1), si denoti con Y_{t-1} l'insieme di osservazioni passate $y_1, \ldots, y_{t-1}$ per $t = 2, 3, \ldots$ e con Y_0 l'osservazione associata all'istante $t = 1$ per cui non ve ne sono altre precedenti. Inoltre si tenga in mente che Y_t è definita come il vettore $(y_1^\top, \ldots, y_t^\top)^\top$. Si derivano quindi le equazioni del filtro di Kalman per il modello (3.1) costruendo in modo ricorsivo le distribuzioni di α_t e y_t. Infatti per la sua natura markoviana vale che

$$\begin{aligned} p(y_t \mid \alpha_1, ..., \alpha_t, Y_{t-1}) &= p(y_t \mid \alpha_t), \\ p(\alpha_{t+1} \mid \alpha_1, ..., \alpha_t, Y_t) &= p(\alpha_{t+1} \mid \alpha_t). \end{aligned}$$

In tabella 3.1 si mostrano le dimensioni dei vettori e delle matrici coinvolte nel modello (3.1). É da sottolineare che per la derivazione del filtro, almeno inizialmente, si assume che lo stato iniziale α_1 sia distribuito come una $N(a_1, P_1)$ con a_1 e P_1 noti.

Vector		**Matrix**	
y_t	$(p \times 1)$	Z_t	$(p \times m)$
α_t	$(m \times 1)$	T_t	$(m \times m)$
ϵ_t	$(p \times 1)$	H_t	$(p \times p)$
η_t	$(r \times 1)$	R_t	$(m \times r)$
		Q_t	$(r \times r)$
α_1	$(m \times 1)$	P_1	$(m \times m)$

Tabella 3.1: Dimensioni dei vettori e delle matrici coinvolte nel modello *state space* (3.1)

L'obiettivo è quello di ottenere le distribuzioni condizionate di α_t e α_{t+1} dato Y_t per $t = 1, \ldots, n$. Sia

$$a_{t|t} = \mathbb{E}(\alpha_t \mid Y_t) \qquad P_{t|t} = \mathrm{Var}(\alpha_t \mid Y_t)$$
$$a_{t+1} = \mathbb{E}(\alpha_{t+1} \mid Y_t) \qquad P_{t+1} = \mathrm{Var}(\alpha_{t+1} \mid Y_t)$$

Poiché le distribuzioni coinvolte in (3.1) sono Normali, applicando il Lemma 3.1, segue che

$$\alpha_t \mid Y_t \sim N(a_{t|t}, P_{t|t})$$
$$\alpha_{t+1} \mid Y_t \sim N(a_{t+1}, P_{t+1}).$$

La derivazione delle equazioni del filtro di Kalman parte da $\alpha_t \mid Y_{t-1} \sim N(a_t, P_t)$, di cui conosciamo a_t and P_t. Infatti attraverso queste quantità si riescono a calcolare ricorsivamente per $t = 1, ..., n$ le seguenti d'interesse $a_{t|t}$, a_{t+1}, $P_{t|t}$ e P_{t+1}.

Si definisce l'errore di previsione a un passo di y_t dato Y_{t-1} come

$$\begin{aligned} v_t &= y_t - \mathbb{E}\big[y_t \mid Y_{t-1}\big] \\ &= y_t - \mathbb{E}\big[Z_t\alpha_t + \epsilon_t \mid Y_{t-1}\big] \\ &= y_t - Z_t a_t \end{aligned} \tag{3.6}$$

Se Y_{t-1} e v_t fossero noti, allora anche Y_t lo sarebbe e viceversa. Dunque si può scrivere $\mathbb{E}[\alpha_t \mid Y_t] = \mathbb{E}[\alpha_t \mid Y_{t-1}, v_t]$. Vale che

$$\begin{aligned} \mathbb{E}\big[v_t \mid Y_{t-1}\big] &= \mathbb{E}\big[y_t - Z_t a_t \mid Y_{t-1}\big] \\ &= \mathbb{E}\big[Z_t\alpha_t + \epsilon_t - Z_t a_t \mid Y_{t-1}\big] = 0 \end{aligned}$$

Conseguentemente

$$\begin{aligned}\mathbb{E}\big[v_t\big] &= 0\\ \mathrm{Cov}\big(y_j, v_t\big) &= \mathbb{E}\big[y_j v_t\big]\\ &= \mathbb{E}\big[\mathbb{E}\big[y_j v_t \mid Y_{t-1}\big]\big]\\ &= \mathbb{E}\big[y_j \mathbb{E}\big[v_t \mid Y_{t-1}\big]\big] = 0\end{aligned}$$

per $j = 1, \ldots, t-1$. Pertanto si può scrivere

$$\begin{aligned}a_{t|t} &= \mathbb{E}\big[\alpha_t \mid Y_t\big] = \mathbb{E}\big[\alpha_t \mid Y_{t-1}, v_t\big]\\ a_{t+1} &= \mathbb{E}\big[\alpha_{t+1} \mid Y_t\big] = \mathbb{E}\big[\alpha_{t+1} \mid Y_{t-1}, v_t\big]\end{aligned}$$

Ora si applichi il Lemma 3.1 alla distribuzione congiunta di α_t e v_t dato Y_{t-1}, considerando rispettivamente come x e y del Lemma 3.1 α_t e v_t. Dunque si arriva a

$$a_{t|t} = \mathbb{E}\big[\alpha_t \mid Y_{t-1}\big] + \mathrm{Cov}(\alpha_t, v_t)[\mathrm{Var}(v_t)]^{-1} v_t \tag{3.7}$$

con Cov e Var che denotano la covarianza e la varianza legata alla distribuzione congiunta di α_t e v_t condizionata a Y_{t-1}. Nell'equazione (3.7) per definizione di a_t è facile vedere che $\mathbb{E}\big[\alpha_t \mid Y_{t-1}\big] = a_t$ e per definizione di P_t

$$\begin{aligned}\mathrm{Cov}(\alpha_t, v_t) &= \mathbb{E}(\alpha_t v_t \mid Y_{t-1})\\ &= \mathbb{E}(\alpha_t (Z_t \alpha_t + \epsilon_t - Z_t a_t)^\top \mid Y_{t-1})\\ &= P_t Z_t^\top\end{aligned}$$

Il calcolo esteso si può trovare in appendice B.2. In seguito si definisca

$$\begin{aligned}F_t &= \mathrm{Var}(v_t \mid Y_{t-1})\\ &= \mathrm{Var}(Z_t \alpha_t + \epsilon_t - Z_t a_t \mid Y_{t-1})\\ &= \mathrm{Var}(Z_t \alpha_t + \epsilon_t \mid Y_{t-1})\\ &= \mathrm{Var}(Z_t \alpha_t \mid Y_{t-1}) + \mathrm{Var}(\epsilon_t \mid Y_{t-1}) + 2\mathrm{Cov}(Z_t \alpha_t, \epsilon_t \mid Y_{t-1})\\ &= Z_t P_t Z_t^\top + H_t\end{aligned} \tag{3.8}$$

quindi si può scrivere

$$a_{t|t} = a_t + P_t Z_t^\top F_t^{-1} v_t \tag{3.9}$$

Per l'equazione (3.5) nel Lemma 3.1 si ottiene che

$$\begin{aligned} P_{t|t} &= \mathrm{Var}(\alpha_t \mid Y_t) \\ &= \mathrm{Var}(\alpha_t \mid Y_{t-1}, v_t) \\ &= \mathrm{Var}(\alpha_t \mid Y_{t-1}) - \mathrm{Cov}(\alpha_t, v_t)[\mathrm{Var}(v_t)]^{-1}\mathrm{Cov}(\alpha_t, v_t)^\top \\ &= P_t - P_t Z_t^\top F_t^{-1} Z_t P_t. \end{aligned} \tag{3.10}$$

Si assume che F_t sia non singolare. Le relazioni descritte in equazione (3.9) e (3.10) sono indicate con il nome di *updating step* del filtro di Kalman.

Ora si sviluppano i passaggi necessari ad arrivare all'ottenimento ricorsivo di a_{t+1} e P_{t+1}. Si ricordi la seconda equazione legata al modello lineare *state space* Gaussiano (3.1). Da questa relazione deriva che

$$\begin{aligned} a_{t+1} &= \mathbb{E}(T_t\alpha_t + R_t\eta_t \mid Y_t) \\ &= T_t\mathbb{E}(\alpha_t \mid Y_t) + R_t\mathbb{E}(\eta_t \mid Y_t) \\ &= T_t\mathbb{E}(\alpha_t \mid Y_t) \end{aligned} \tag{3.11}$$

$$\begin{aligned} P_{t+1} &= \mathrm{Var}(T_t\alpha_t + R_t\eta_t \mid Y_t) \\ &= T_t\mathrm{Var}(\alpha_t \mid Y_t)T_t^\top + R_tQ_tR_t^\top \end{aligned} \tag{3.12}$$

per $t = 1, \ldots, n$.

Sostituendo ciò che è contenuto nell'equazione (3.9) in equazione (3.11) emerge che

$$\begin{aligned} a_{t+1} &= T_t a_{t|t} \\ &= T_t a_t + T_t P_t Z_t^\top F_t^{-1} v_t \\ &= T_t a_t + K_t v_t, \qquad t = 1, \ldots, n \end{aligned} \tag{3.13}$$

con

$$K_t = T_t P_t Z_t^\top F_t^{-1}.$$

Ci si riferisce alla matrice K_t come *Kalman gain*. In equazione (3.13) si osserva come a_{t+1} venga ottenuto come una funzione lineare del valore precedente a_t e v_t, l'errore di previsione a un passo di y_t dato Y_{t-1}. Si può quindi affermare, usando una terminologia tipica del campo del *machine learning*, che il filtro impara dai suoi errori in una misura regolata proprio dalla matrice K_t. Infine, rimpiazzando in (3.12) quanto riportato in (3.10) si ottiene

$$\begin{aligned} P_{t+1} &= T_t P_{t|t} T_t^\top + R_t Q_t R_t^\top \\ &= T_t P_t \left(T_t - K_t Z_t\right)^\top + R_t Q_t R_t^\top \qquad t = 1, \ldots, n. \end{aligned} \tag{3.14}$$

Il calcolo esteso si può trovare in appendice B.2. Le relazioni in (3.13) e (3.14) sono indicate come *prediction step* del filtro di Kalman.

Le equazioni espresse in (3.9), (3.13), (3.10) e in (3.14) costituiscono il filtro di Kalman per il modello descritto in (3.1). Permettono di aggiornare la propria conoscenza del sistema per ogni istante in cui una nuova osservazioni viene rilevata. Inoltre è da sottolineare il fatto che la procedura ricorsiva testè descritta è stata derivata attraverso la semplice applicazione del Lemma 3.1. La caratteristica di maggior rilievo del filtro consiste nel fatto che, per aggiornare la stima al tempo t, non occorre invertire una matrice $(pt \times pt)$, operazione computazionalmente onerosa. È sufficiente, invece, invertire la sola matrice $(p \times p)$ F_t, con p tipicamente molto inferiore a n; nello studio considerato, $p = 1$.

In conclusione in (3.15) si raggruppano tutte le equazioni coinvolte nel filtro di Kalman e in tabella 3.2 si illustrano le dimensioni dei vettori e delle matrici coinvolte.

$$\begin{aligned} v_t &= y_t - Z_t a_t & F_t &= Z_t P_t Z_t^\top + H_t \\ a_{t|t} &= a_t + P_t Z_t^\top F_t^{-1} v_t & P_{t|t} &= P_t - P_t Z_t^\top F_t^{-1} Z_t P_t \\ a_{t+1} &= T_t a_t + K_t v_t & P_{t+1} &= T_t P_t \left(T_t - K_t Z_t\right)^\top + R_t Q_t R_t^\top \end{aligned} \tag{3.15}$$

Vector		**Matrix**	
v_t	$(p \times 1)$	F_t	$(p \times p)$
		K_t	$(m \times p)$
a_t	$(m \times 1)$	P_t	$(m \times m)$
$a_{t\|t}$	$(m \times 1)$	$P_{t\|t}$	$(m \times m)$

Tabella 3.2: Dimensioni dei vettori e delle matrici coinvolte nelle equazioni del filtro di Kalman.

3.3.3 Valori mancanti e previsione

È comune avere a che fare con serie storiche con valori mancanti. Nel caso in cui si adotti il modello lineare state space definito in (3.1) per l'analisi, la gestione delle osservazioni mancanti nella derivazione del filtro di Kalman risulta particolarmente agevole. Si supponga che le osservazioni y_j per $j = \tau, \ldots, \tau^*$ con $1 < \tau < \tau^* < n$ siano assenti. Allora per $t = \tau, \ldots, \tau^* - 1$ si ha che

$$\begin{aligned} a_{t|t} &= \mathbb{E}(\alpha_t \mid Y_t) \\ &= \mathbb{E}(\alpha_t \mid Y_{t-1}) \end{aligned}$$

$$
\begin{aligned}
&= a_t \\
P_{t|t} &= \mathrm{Var}(\alpha_t \mid Y_t) \\
&= \mathrm{Var}(\alpha_t \mid Y_{t-1}) \\
&= P_t \\
a_{t+1} &= \mathbb{E}(\alpha_{t+1} \mid Y_t) \\
&= \mathbb{E}(T_t\alpha_t + R_t\eta_t \mid Y_{t-1}) \\
&= T_t a_t \\
P_{t+1} &= \mathrm{Var}(\alpha_{t+1} \mid Y_t) \\
&= \mathrm{Var}(T_t\alpha_t + R_t\eta_t \mid Y_{t-1}) \\
&= T_t P_t T_t^\top + R_t Q_t R_t^\top
\end{aligned}
$$

Ne segue che il passo del filtro di Kalman nel caso di osservazioni mancanti si ottiene semplicemente ponendo Z_t in (3.15) pari a 0 per $t = \tau, \ldots, \tau^* - 1$.

La previsione di valori futuri del vettore degli stati è una pratica d'interesse piuttosto comune quando si ha a che fare con serie storiche. Si dimostra che l'errore quadratico medio delle previsioni è minimizzato quando i valori futuri di y_t sono trattati come osservazioni mancanti.

Si supponga di avere a disposizione il vettore di osservazioni $y_1, \ldots, y_n$ modellizzate secondo (3.1). L'obiettivo è quello di prevedere y_{n+j} per $j = 1, \ldots, J$. Pertanto si cerca $\hat{y}_{n+j}$ stimatore per y_{n+j}, tale per cui la matrice dell'errore quadratico medio dato Y_n, definita come $\hat{F}_{n+j} = \mathbb{E}[(\hat{y}_{n+j} - y_{n+j})(\hat{y}_{n+j} - y_{n+j})^\top \mid Y_n]$, sia minimizzata[1]. É noto che se x è un vettore random con media μ e matrice delle varianze e covarianze finita, allora il vettore λ che minimizza $\mathbb{E}[(\lambda - x)(\lambda - x)^\top]$ è $\lambda = \mu$. Segue che l'errore quadratico medio della previsione di y_{n+j} dato Y_n è minimizzato da $\bar{y}_{n+j} = \mathbb{E}[Y_{n+j} \mid Y_n]$.

La previsione per $j = 1$ è diretta. Dall'equazione in (3.1) segue che $y_{n+1} = Z_{n+1}\alpha_{n+1} + \epsilon_{n+1}$, pertanto:

$$
\begin{aligned}
\bar{y}_{n+1} &= Z_{n+1}\mathbb{E}\big[\alpha_{n+1} \mid Y_n\big] \\
&= Z_{n+1}a_{n+1},
\end{aligned}
$$

[1]Per ogni altri stimatore $\tilde{y}_{n+j}$ la matrice definita come $\hat{F}_{n+j} - \tilde{F}_{n+j}$ è semidefinita positiva

con a_{n+1} che è la stima di α_{n+1} prodotta dal filtro di Kalman in equazione (3.13). La matrice dell'errore quadratico medio dato Y_n è definita dall'equazione (3.8)

$$\begin{aligned}\bar{F}_{n+1} &= \mathbb{E}\big[(\bar{y}_{n+1} - y_{n+1})(\bar{y}_{n+1} - y_{n+1})^\top \mid Y_n\big] \\ &= Z_{n+1}P_{n+1}Z_{n+1}^\top + H_{n+1}.\end{aligned}$$

Ora si generalizza questo risultato per $j = 2, \ldots, J$. Siano $\bar{a}_{n+j} = \mathbb{E}[\alpha_{n+j} \mid Y_n]$ e $\bar{P}_{n+j} = \mathbb{E}[(\bar{a}_{n+j} - \alpha_{n+j})(\bar{a}_{n+j} - \alpha_{n+j})^\top \mid Y_n]$. Poiché $y_{n+j} = Z_{n+j}\alpha_{n+j} + \epsilon_{n+j}$ si ha che

$$\begin{aligned}\bar{y}_{n+j} &= Z_{n+j}\mathbb{E}(\alpha_{n+j} \mid Y_n) \\ &= Z_{n+j}\bar{a}_{n+j}\end{aligned}$$

con la matrice dell'errore quadratico medio condizionato a Y_n pari a

$$\begin{aligned}\bar{F}_{n+j} &= \mathbb{E}[\{Z_{n+j}(\bar{a}_{n+j} - \alpha_{n+j}) - \epsilon_{n+j}\}\{Z_{n+j}(\bar{a}_{n+j} - \alpha_{n+j}) - \epsilon_{n+j}\}^\top \mid Y_n] \\ &= Z_{n+j}\bar{P}_{n+j}Z_{n+j}^\top + H_{n+j}.\end{aligned}$$

Dunque si derivano ricorsivamente $\bar{a}_{n+j}$ e $\bar{P}_{n+j}$. Si ha che $\alpha_{n+j+1} = T_{n+j}\alpha_{n+j} + R_{n+j}\eta_{n+j}$ pertanto

$$\begin{aligned}\bar{a}_{n+j+1} &= T_{n+j}\mathbb{E}\big[\alpha_{n+j} \mid Y_n\big] \\ &= T_{n+j}\bar{a}_{n+j},\end{aligned}$$

per $j = 1, \ldots, J-1$ e con $\bar{a}_{n+1} = a_{n+1}$. Inoltre,

$$\begin{aligned}\bar{P}_{n+j+1} &= \mathbb{E}\big[(\bar{a}_{n+j+1} - \alpha_{n+j+1})(\bar{a}_{n+j+1} - \alpha_{n+j+1})^\top \mid Y_n\big] \\ &= T_{n+j}\mathbb{E}\big[(\bar{a}_{n+j} - \alpha_{n+j})(\bar{a}_{n+j} - \alpha_{n+j})^\top \mid Y_n\big]T_{n+j}^\top \\ &\quad + R_{n+j}\mathbb{E}\big[\eta_{n+j}\eta_{n+j}^\top\big]R_{n+j}^\top \\ &= T_{n+j}\bar{P}_{n+j}T_{n+j}^\top + R_{n+j}Q_{n+j}R_{n+j}^\top\end{aligned}$$

per $j = 1, \ldots, J-1$. La derivazione estesa di questo calcolo si può trovare in appendice B.2.

Si osserva che le equazioni ricorsive che definiscono $\bar{a}_{n+j}$ e $\bar{P}_{n+j}$ sono le stesse per a_{n+j} and P_{n+j} del filtro di Kalman descritte in (3.15) con Z_{n+j} posto uguale a 0 per $j = 1, \ldots, J-1$. Quest'ultima è la condizione che si era imposta precedentemente quando si era spiegato come gestire le osservazioni mancanti. Si è dunque dimostrato come le previsioni $y_{n+1}, \ldots, y_{n+J}$ e le matrici di varianza dell'errore associate possono essere ottenute in modo ricorsivo nell'analisi di serie storiche in forma *state space*

basate su modelli lineari Gaussiani semplicemente proseguendo il filtro di Kalman aldilà dell'istante $t = n$ con $Z_t = 0$ per $t > n$.

3.3.4 Inizializzazione del filtro

Fino ad ora si è sviluppato il modello (3.1) assumendo che $\alpha_1 \sim N(a_1, P_1)$ con a_1 e P_1 noti. Tuttavia, nella maggior parte delle applicazioni, almeno alcuni elementi di a_1 e P_1 sono ignoti. In queste circostanze è necessario sviluppare dei metodi alternativi per determinare lo stato iniziale del filtro. Questo processo è indicato come inizializzazione. Un modello generale per il vettore degli stati iniziale α_1 è

$$\alpha_1 = a + A\delta + R_0\eta_0, \quad \eta_0 \sim N(0, Q_0) \tag{3.16}$$

con a che è un vettore $m \times 1$ di quantità note, δ un vettore $q \times 1$ di quantità ignote, le matrici A e R_0 sono di selezione, ovvero sono costituite da colonne della matrice identità I_m e le loro dimensioni sono rispettivamente $m \times q$ e $m \times (m - q)$. Sono inoltre costruite in modo che se prese insieme le loro colonne compongono un insieme di g colonne di I_m con $g \leq m$ e $A^\top R_0 = 0$. Si assume che la matrice Q_0 sia definita positiva e sia nota. Nella maggior parte dei casi il vettore a sarà trattato come un vettore di zeri, a meno che alcuni elementi del vettore degli stati iniziale α_1 non siano delle costanti note.

Il vettore δ può essere trattato come un vettore di parametri ignoti fissi oppure come un vettore di variabili casuali normali con varianza infinita. Si approfondisce solamente questo secondo caso, e se si fosse interessati a studiare anche il primo si rimanda a Durbin & Koopman (2012). Dunque δ è casuale e si assume che

$$\delta \sim N(0, \kappa I_q) \tag{3.17}$$

con $\kappa \to \infty$. Il filtro di Kalman è inizializzato alle condizioni seguenti: $a_1 = \mathbb{E}[\alpha_1] = a$ e $P_1 = \mathrm{Var}(\alpha_1)$, con

$$P_1 = \kappa P_\infty + P_*, \tag{3.18}$$

con $P_\infty = AA^\top$ e $P_* = R_0 Q_0 R_0^\top$. Poiché A è composta da colonne di I_m, segue che P_∞ è una matrice diagonale di dimensione $m \times m$, con gli elementi sulla diagonali pari a 1 e i restanti pari a 0. Inoltre, quando un elemento sulla diagonale di P_∞ risulta diverso da zero, allora si pone il rispettivo elemento in a pari a zero. Un vettore δ distribuito come $N(0, \kappa I_q)$ con $\kappa \to \infty$ è indicato come diffuso. L'inizializzazione del

filtro di Kalman quando alcuni elementi di α_1 sono diffusi è detta inizializzazione diffusa del filtro.

Tuttavia il filtro di Kalman, nel caso fosse inizializzato in modo diffuso, necessita di alcune modifiche. Una tecnica di approssimazione naïve sarebbe quella di rimpiazzare κ in (3.18) con un numero arbitrariamente grande. Questo approccio non è consigliato in quanto, generalmente, porta ad ampi errori di approssimazione. Il metodo utilizzato è quello definito come esatto - per un approfondimento si legga Durbin & Koopman (2012).

3.3.5 Stima di massima verosimiglianza dei parametri

Nel modello LLT descritto in sotto-sezione 3.2.1 le varianze σ_ϵ^2, σ_ξ^2 e σ_ζ^2 potrebbero essere non note e quindi dovrebbero essere stimate. Nell'approccio classico, vengono considerate come parametri fissi ed ignoti, ed un modo per trovare la loro stima è quello della massima verosimiglianza. Per il modello lineare Gaussiano (3.1) si dimostra che la verosimiglianza può essere calcolata applicando il filtro di Kalman.

In un primo momento si assume che il vettore inziale degli stati abbia densità $N(a_1, P_1)$ con a_1 e P_1 noti, in seguito si generalizzerà anche al caso in cui l'inizializzazione fosse diffusa. La verosimiglianza è

$$L(Y_n) = p(y_1, ..., y_n) = p(y_1) \prod_{t=2}^{n} p(y_t \mid Y_{t-1})$$

con $Y_t = (y_1, ..., y_t)$. É più semplice, però, lavorare con la log-verosimiglianza,

$$\log L(Y_n) = \sum_{t=1}^{n} \log p(y_t \mid Y_{t-1}), \tag{3.19}$$

con $p(y_1 \mid Y_0) = p(y_1)$. Per il modello (3.1) si ha che $\mathbb{E}[y_t \mid Y_{t-1}] = Z_t a_t$. L'espressione per la log-verosimiglianza si ottiene definendo $v_t = y_t - Z_t a_t$ e $F_t = \mathrm{Var}(y_t \mid Y_{t-1})$ e sostituendo $N(Z_t a_t, F_t)$ per $p(y_t \mid Y_{t-1})$. Si ricava quindi la seguente espressione per (3.19)

$$\log L(Y_n) = -\frac{np}{2} \log 2\pi - \frac{1}{2} \sum_{t=1}^{n} (\log|F_t| + v_t^\top F_t^{-1} v_t), \tag{3.20}$$

Le quantità v_t e F_t sono calcolate in modo iterativo dal filtro di Kalman e di conseguenza anche $\log L(Y_n)$ è facilmente ottenibile.

Ora si considera il caso in cui alcuni elementi in α_1 siano diffusi. Si assume che $\alpha_1 = a + A\delta + R_0\eta_0$ con a che è un vettore di costanti note, $\delta \sim N(0, \kappa I_q)$, $\eta_0 \sim N(0, Q_0)$ e $A^\top R_0 = 0$ da cui risulta che $a_1 \sim N(a_1, P_1)$ con $P_1 = \kappa P_\infty + P_*$

e $\kappa \to \infty$. In questo contesto diffuso la log-verosimiglianza in (3.20) conterrà un termine pari a $-\frac{1}{2}q \log 2\pi\kappa$ tale per cui $\log L(Y_n)$ non converga per $\kappa \to \infty$. Quindi si definisce la log-verosimiglianza diffusa come

$$\log L_d(Y_n) = \lim_{\kappa \to \infty} [\log L(Y_n) + \frac{q}{2} \log \kappa] \tag{3.21}$$

e si lavora con $\log L_d(Y_n)$ al posto di $\log L(Y_n)$ per stimare i parametri ignoti. Sono disponibili diversi algoritmi per la massimizzazione della log-verosimiglianza rispetto a questi parametri. Per ulteriore approfondimento si rimanda a Durbin & Koopman (2012).

3.4 Modelli state space non Gaussiani

Fino ad ora si è presentata la costruzione e l'analisi dei modelli lineari *state space* Gaussiani. I metodi basati su questi modelli sono appropriati per una ampia serie di problemi nell'analisi delle serie storiche. Tuttavia, esistono sistemi dinamici tali per cui il modello lineare Gaussiano fallisce nel produrre una rappresentazione accettabile dei dati. Per esempio, se le osservazioni sono il numero di intervistati che ha risposto di votare un certo partito in un sondaggio, la distribuzione Binomiale produce un modello più appropriato per i dati rispetto alla distribuzione Normale. Dunque si necessita di un modello per lo sviluppo nel tempo di una variabile Binomiale piuttosto che Normale.

Una forma generale per il modello *state space* non lineare e non Gaussiano è la seguente

$$y_t \sim p(y_t \mid \alpha_t), \quad \alpha_{t+1} \sim p(\alpha_{t+1} \mid \alpha_t), \quad \alpha_1 \sim p(\alpha_1)$$

per $t = 1, \ldots, n$. Si assume inoltre che

$$p(Y_n \mid \alpha) = \prod_{t=1}^{n} p(y_t \mid \alpha_t), \quad p(\alpha) = p(\alpha_1) \prod_{t=1}^{n-1} p(\alpha_{t+1} \mid \alpha_t)$$

con $Y_n = (y_1^\top, \ldots, y_n^\top)^\top$ e $\alpha = (\alpha_1^\top, ..., \alpha_n^\top)^\top$. La densità delle osservazioni $p(y_t \mid \alpha_t)$ gestisce la relazione fra il vettore delle osservazioni y_t e il vettore degli stati α_t. La densità $p(\alpha_{t+1} \mid \alpha_t)$ regola invece la relazione fra il vettore degli stati all'istante successivo α_{t+1} con quello all'istante corrente α_t. Nel caso in cui le relazioni descritte in $p(y_t \mid \alpha_t)$ e in $p(\alpha_{t+1} \mid \alpha_t)$ siano lineari allora il modello sarà indicato come *state space* lineare non Gaussiano. Al contrario, se le densità $p(y_t \mid \alpha_t)$ e $p(\alpha_{t+1} \mid \alpha_t)$ sono Gaussiane, ma almeno una relazione in $p(y_t \mid \alpha_t)$ o in $p(\alpha_{t+1} \mid \alpha_t)$ è non lineare, il

modello sarà definito come *state space* non lineare Gaussiano. Infine se le le relazioni descritte in $p(y_t \mid \alpha_t)$ e in $p(\alpha_{t+1} \mid \alpha_t)$ sono lineari e le densità $p(y_t \mid \alpha_t)$ e $p(\alpha_{t+1} \mid \alpha_t)$ sono Gaussiane, si ritorna al modello (3.1).

3.4.1 Modelli con segnale lineare Gaussiano

S'introduce il modello multivariato con segnale lineare Gaussiano. Quest'ultimo ha una struttura *state space* simile a quella del modello (3.1). Infatti il vettore delle osservazioni è determinato dalla relazione seguente

$$p(y_t \mid \alpha_1, \ldots, \alpha_t, y_1, \ldots, y_{t-1}) = p(y_t \mid Z_t\alpha_t)$$

con il vettore degli stati α_t che è regolato dall'espressione

$$\alpha_{t+1} = T_t\alpha_t + R_t\eta_t, \quad \eta_t \sim N(0, Q_t), \tag{3.22}$$

con i disturbi η_t serialmente indipendenti per $t = 1, \ldots, n$. Si definisce poi

$$\theta_t = Z_t\alpha_t, \tag{3.23}$$

e ci si riferisce a θ_t come segnale. Quest'ultimo è il predittore lineare connesso con il valore atteso $\mathbb{E}[y_t] = \mu_t$ attraverso la *link function* $g(\mu_t) = \theta_t$. Ad esempio nel caso in cui $p(y_t \mid \theta_t)$ sia Binomiale con probabilità di successo π_t si usa il *link* logit, tale per cui $\theta_t = \text{logit}(\pi_t)$.

La densità $p(y_t \mid \theta_t)$ può essere non Gaussiana. Nel caso in cui $p(y_t \mid \theta_t)$ sia Normale e θ_t lineare in y_t, il modello si riduce a quello lineare Gaussiano descritto in (3.1). Se le osservazioni sono generate da una distribuzione che appartiene alla famiglia delle distribuzioni esponenziali, allora avranno una densità che si può scrivere nella seguente maniera

$$p(y_t \mid \theta_t) = \exp\{y_t^\top\theta_t - b_t(\theta_t) + c_t(y_t)\}, \quad -\infty < \theta_t < \infty, \tag{3.24}$$

con $b_t(\theta_t)$ che è differenziabile due volte e $c_t(y_t)$ che è una funzione solo di y_t. Il modello (3.24) insieme alle relazioni descritte in (3.22) e (3.23), con η_t assunto Gaussiano, è stato introdotto con il nome di *dynamic generalized linear model*.

3.4.2 Modelli state space per la famiglia esponenziale

Per il modello (3.24), si definiscono le seguenti quantità

$$\dot{b}_t(\theta_t) = \frac{\partial b_t(\theta_t)}{\partial \theta_t} \qquad \text{and} \qquad \ddot{b}_t(\theta_t) = \frac{\partial^2 b_t(\theta_t)}{\partial \theta_t \partial \theta_t^\top}.$$

Assumendo che le principali di condizioni di regolarità siano soddisfatte, si dimostra che vale

$$\mathbb{E}\big[y_t\big] = \dot{b}_t(\theta_t) \quad \text{e} \quad \text{Var}(y_t) = \ddot{b}_t(\theta_t).$$

Come esempio di distribuzione appartenete alla famiglia esponenziale si fa uso della Binomiale, in quanto in seguito sarà utile nell'analisi. Per altri esempi si legga Durbin & Koopman (2012). L'osservazione y_t ha una distribuzione Binomiale se è uguale al numero di successi in k_t prove indipendenti con una probabilità di successo pari a π_t. La densità di y_t è

$$p(y_t \mid \pi_t) = \binom{k_t}{y_t} \pi_t^{y_t} (1 - \pi_t)^{k_t - y_t},$$

con $y_t = 0, \dots, k_t$. Dunque si calcola la log-verosimiglianza

$$\log p(y_t \mid \pi_t) = y_t \big[\log \pi_t - \log(1 - \pi_t)\big] + k_t \log(1 - \pi_t) + \log \binom{k_t}{y_t}.$$

Riscriviamo questa espressione nella forma mostrata in equazione (3.24) ponendo $\theta_t = \log[\pi_t/(1 - \pi_t)]$ e $b_t(\theta_t) = k_t \log(1 + \exp \theta_t)$

$$p(y_t \mid \theta_t) = \exp\left[y_t \theta_t - k_t \log(1 + \exp \theta_t) + \log \binom{k_t}{y_t}\right]. \tag{3.25}$$

Da quest'ultima espressione (3.25) è facile ricavare media e varianza,

$$\begin{aligned} \dot{b}_t &= k_t \frac{\exp \theta_t}{1 + \exp \theta_t} = k_t \pi_t \\ \ddot{b}_t &= k_t \frac{\exp \theta_t}{(1 + \exp \theta_t)^2} = k_t \pi_t (1 - \pi_t). \end{aligned} \tag{3.26}$$

3.5 Filtro approssimato

Nel momento in cui si ha a che fare con modelli non Gaussiani o non lineari le equazioni del filtro di Kalman non esistono più in forma chiusa. Dunque è necessario fare affidamento a soluzioni alternative. Un possibile approccio al problema è quello

di fare uso di un'approssimazione. Un esempio è il Filtro di Kalman Esteso che si basa sulla linearizzazione delle equazioni di stato e di osservazione attraverso un'espansione di Taylor per poi applicare in modo diretto il filtro al modello linearizzato ottenuto. Un'altra idea, più semplice, è quella di applicare una trasformazione alle osservazioni tale per cui il modello lineare gaussiano possa essere applicato come approssimazione. Tuttavia, il metodo approfondito, utilizzato poi anche nell'applicazione in quanto alla base degli algoritmi usati nel pacchetto R **KFAS** (Helske, 2017), è quello dell'approssimazione attraverso la stima della moda. Con questo metodo, per fare inferenza sui modelli non gaussiani, si trova un modello gaussiano che abbia la stessa moda a posteriori di $p(\theta \mid y)$. Per ottenere questo risultato si utilizza un processo iterativo basato sull'approssimazione di Laplace di $p(\theta \mid y)$, con le stime aggiornate di θ calcolate attraverso l'applicazione del filtro e dello smoother del modello gaussiano approssimante. Nel modello gaussiano approssimante l'equazione delle osservazioni è rimpiazzata da

$$x_t = Z_t\alpha_t + \epsilon_t \qquad \epsilon_t \sim N(0, A_t)$$

con le pseudo-osservazioni x_t e varianze A_t basate sulla prima e seconda derivata di $\log p(y_t \mid \theta_t)$ rispetto a θ_t. Le stime finali $\hat{\theta}_t$ corrispondono alla moda di $p(\theta \mid y)$. Nel caso gaussiano la moda è anche la media mentre nel caso binomiale la differenza è spesso contenuta, ma può diventare non trascurabile per probabilità vicine a $0/1$ o campioni piccoli. Spesso si è interessati a α_t piuttosto che al predittore lineare θ_t. In quanto la funzione legame $g(\cdot)$ non è lineare la trasformazione $\hat{\mu}_t = g(\hat{\theta}_t)$ comporta della distorsione, ma **KFAS** sviluppa metodi che la mitigano, per i dettagli si legga Helske (2017).

3.5.1 Approssimazione attraverso stima della moda

Si consideri il modello *state space* introdotto in sotto-sezione 3.4.1 con struttura

$$\begin{aligned} y_t \mid \theta_t &\sim p(y_t \mid \theta_t), \\ \theta_t &= Z_t\alpha_t, \\ \alpha_{t+1} &= T_t\alpha_t + R_t\eta_t, \qquad \eta_t \sim N(0, Q_t), \end{aligned}$$

per $t = 1, \ldots, n$. La dinamica di stato è lineare e gaussiana, mentre $p(y_t \mid \theta_t)$ è una distribuzione non gaussiana. Nel caso di particolare interesse essa appartiene alla famiglia delle distribuzioni esponenziali, pertanto la sua verosimiglianza può essere espressa come in equazione (3.24) e valgono le proprietà illustrate in sotto-sezione

3.4.2. Il calcolo della moda dei θ_t condizionatamente alle osservazioni $y_1, \dots, y_n$ porta alla costruzione di un modello *state space* lineare e gaussiano approssimante.

Gli input dell'algoritmo di approssimazione via stima della moda sono, oltre al modello *state space* definito precedentemente, i dati osservati $Y_n = (y_1, \dots, y_n)$ e un vettore di valori iniziali per il segnale $\tilde{\theta}^{(0)} = (\tilde{\theta}_1^{(0)}, \dots, \tilde{\theta}_n^{(0)})^\top$. L'obiettivo è stimare la moda della densità $p(\theta \mid Y_n)$, che in generale non possiede un'espressione chiusa da cui derivare la moda analiticamente.

La distribuzione a posteriori congiunta del segnale è

$$p(\theta \mid Y_n) \propto p(Y_n \mid \theta)\, p(\theta),$$

con $p(Y_n \mid \theta) = \prod_{t=1}^{n} p(y_t \mid \theta_t)$ e $p(\theta) = N(\mu, \Omega)$. La log-posterior è, a meno di una costante additiva,

$$\ell(\theta) = \log p(\theta \mid Y_n) = \log p(Y_n \mid \theta) - \frac{1}{2}(\theta - \mu)^\top \Omega^{-1} (\theta - \mu).$$

Per massimizzare questa espressione si utilizza l'algoritmo di Newton–Raphson. Indicando con

$$\dot{\ell}(\theta) = \frac{\partial \ell(\theta)}{\partial \theta}, \qquad \ddot{\ell}(\theta) = \frac{\partial^2 \ell(\theta)}{\partial \theta\, \partial \theta^\top},$$

un passo di Newton a partire dal valore corrente $\tilde{\theta}$ è dato da

$$\tilde{\theta}^{+} = \tilde{\theta} - \big[\ddot{\ell}(\theta)\big|_{\theta=\tilde{\theta}}\big]^{-1} \dot{\ell}(\theta)\big|_{\theta=\tilde{\theta}}.$$

In Durbin & Koopman (2012) si dimostra che questo passo di aggiornamento può essere riscritto nella forma

$$\tilde{\theta}^{+} = (\Omega^{-1} + A^{-1})^{-1} \big(A^{-1} x + \Omega^{-1} \mu\big),$$

dove

$$A = -\big[\ddot{p}(Y_n \mid \theta)\big|_{\theta=\tilde{\theta}}\big]^{-1}, \qquad x = \tilde{\theta} + A\, \dot{p}(Y_n \mid \theta)\big|_{\theta=\tilde{\theta}},$$

con

$$\dot{p}(Y_n \mid \theta) = \frac{\partial \log p(Y_n \mid \theta)}{\partial \theta} \qquad \ddot{p}(Y_n \mid \theta) = \frac{\partial^2 \log p(Y_n \mid \theta)}{\partial \theta \partial \theta^T}.$$

La chiave del metodo è reinterpretare il passo di Newton–Raphson come la moda di un modello lineare e gaussiano. Infatti, in un modello puramente lineare e gaussiano

che si può scrivere come

$$x = \theta + \varepsilon, \qquad \varepsilon \sim N(0, A), \qquad \theta \sim N(\mu, \Omega),$$

la moda (che coincide con la media) della posterior $p(\theta \mid x)$ è

$$\hat{\theta} = (\Omega^{-1} + A^{-1})^{-1}\big(A^{-1}x + \Omega^{-1}\mu\big). \tag{3.27}$$

In altre parole, un passo di Newton–Raphson per $\ell(\theta)$ può essere visto come il calcolo della media (moda) di un opportuno modello lineare gaussiano fittizio con osservazioni x e matrice di varianza A.

Nel caso in cui $p(y_t \mid \theta_t)$ appartenga alla famiglia esponenziale, si ha

$$\log p(y_t \mid \theta_t) = y_t^\top \theta_t - b_t(\theta_t) + c_t(y_t),$$

da cui discendono, per ogni t,

$$\frac{\partial \log p(y_t \mid \theta_t)}{\partial \theta_t} = y_t - \dot{b}_t(\theta_t), \qquad \frac{\partial^2 \log p(y_t \mid \theta_t)}{\partial \theta_t^2} = -\ddot{b}_t(\theta_t),$$

dove $\dot{b}_t$ e $\ddot{b}_t$ indicano rispettivamente prima e seconda derivata di b_t. Alla j-esima iterazione dell'algoritmo, valutando queste derivate in $\tilde{\theta}_t^{(j)}$, si definiscono

$$A_t^{(j)} = \big[\ddot{b}_t(\tilde{\theta}_t^{(j)})\big]^{-1}, \qquad x_t^{(j)} = \tilde{\theta}_t^{(j)} + A_t^{(j)}\big[y_t - \dot{b}_t(\tilde{\theta}_t^{(j)})\big].$$

Nel caso binomiale con $y_t \mid \pi_t \sim \mathrm{Bin}(k_t, \pi_t)$ e

$$\theta_t = \log \frac{\pi_t}{1 - \pi_t},$$

si ha $b_t(\theta_t) = k_t \log(1 + e^{\theta_t})$, da cui le derivate espresse in (3.26). Si ottiene quindi

$$A_t^{(j)} = \frac{\big(1 + e^{\tilde{\theta}_t^{(j)}}\big)^2}{k_t e^{\tilde{\theta}_t^{(j)}}}, \qquad x_t^{(j)} = \tilde{\theta}_t^{(j)} + \frac{\big(1 + e^{\tilde{\theta}_t^{(j)}}\big)^2}{e^{\tilde{\theta}_t^{(j)}}} \frac{y_t}{k_t} - \big(1 + e^{\tilde{\theta}_t^{(j)}}\big).$$

Il modello gaussiano approssimante alla iterazione j è quindi

$$x_t^{(j)} = \theta_t + \varepsilon_t, \qquad \varepsilon_t \sim N\big(0, A_t^{(j)}\big),$$

con la stessa equazione di stato lineare e gaussiana per α_t e la definizione $\theta_t = Z_t\alpha_t$. Su questo modello si calcola l'aggiornamento del passo di Newton–Raphson, cioè la nuova traiettoria $\tilde{\theta}^{(j+1)}$, ottenuta come in equazione (3.27).

In un modello *state space* lineare e gaussiano, la media condizionata e la matrice di varianza degli stati e del segnale si ottengono attraverso le ben note ricorsioni del filtro di Kalman e dello smoother di Kalman. Nel contesto dell'approssimazione via moda, il ruolo di questi algoritmi è duplice. In primo luogo, all'interno dell'algoritmo iterativo, dato il modello gaussiano approssimante definito da $(x_t^{(j)}, A_t^{(j)})$, si applicano filtro e smoother per ottenere la nuova traiettoria $\tilde{\theta}^{(j+1)}$, che rappresenta un passo di Newton–Raphson sulla log-posterior del segnale. Si itera fino alla convergenza, ottenendo la moda congiunta $\hat{\theta}$. In secondo luogo, una volta fissato il modello approssimante finale (cioè il modello lineare gaussiano costruito attorno a $\hat{\theta}$), le ricorsioni di Kalman forniscono le stime filtrate degli stati

$$\hat{\alpha}_{t|t} = E(\alpha_t \mid X_t), \qquad \hat{\theta}_{t|t} = Z_t\hat{\alpha}_{t|t},$$

Nel caso binomiale, applicando il logit inverso si ottengono infine le stime filtrate nella scala delle probabilità π_t.

3.6 Applicazione

Una variabile casuale discreta X si definisce Binomiale se essa rappresenta il numero dei successi che si verificano in una sequenza di n sottoprove indipendenti nelle quali è costante la probabilità π di un successo; tale famiglia di variabili casuali discrete dipende dai due parametri (n, π) e sarà indicata con $X \sim \text{Bin}(n, \pi)$. Quando $n = 1$ evidentemente la v.c. Binomiale coincide con la v.c. Bernoulli. In generale la distribuzione di probabilità delle v.c. $X \sim \text{Bin}(n, \pi)$ è:

$$P(X = x) = \binom{n}{x}\pi^x(1-\pi)^{n-x}, \tag{3.28}$$

per ulteriori dettagli si consulti ad esempio Piccolo (2010).

La distribuzione Binomiale è appropriata nel contesto di studio in quanto le risposte di ogni intervistato possono essere considerate prove di Bernoulli indipendenti: una prova della quale interessa esclusivamente constatare se un evento E (l'intervistato afferma che voterà il partito in questione nelle prossime elezioni) si è verificato oppure no (in tal caso si è verificata la sua negazione $\bar{E}$, cioè l'intervistato afferma che non

voterà il partito in questione alle prossime elezioni). Aggregando queste risposte, lo schema di un sondaggio sulle intenzioni di voto per un singolo partito alle prossime elezioni è quello associato ad una variabile casuale Binomiale.

L'obiettivo è quello di, attraverso il filtro di Kalman, interpolare la serie storica delle intenzioni di voto di un determinato partito per ogni agenzia in modo da ottenere una stima nella finestra di tempo prima di un'elezione in cui non è possibile pubblicare sondaggi. Infine si adotterà la stima ottenuta corrispondente al giorno delle elezioni come valore da confrontare con il risultato effettivo di quest'ultime. L'indice scelto per il confronto è l'errore relativo definito in sezione 2.5. Una scelta rilevante è stata quella di suddividere ogni serie storica considerata in tre sezioni. La prima inizia dalla prima elezione disponibile, le Politiche del 04/03/2018, e arriva fino alla seconda, le Europee del 26/05/2019. Applicando il filtro a questa parte di sondaggi si ottiene la stima delle intenzioni di voto per ogni partito per ogni agenzia per l'elezioni Europee del 2019. La seconda sezione inizia con la seconda elezione e arriva alla terza disponibile, le Politiche del 25/09/2022. In questo caso attraverso il filtro di Kalman, applicato a questa parte di sondaggi, si ottiene la stima delle intenzioni di voto per ogni partito e per ogni agenzia per l'elezioni Politiche del 2022. Infine la terza sezione spazia dalla terza all'ultima elezione disponibile, le Europee del 09/06/2024. Attraverso il filtro di Kalman, applicato a questa parte di sondaggi, si ottiene la stima delle intenzioni di voto per ogni partito e per ogni agenzia proprio per quest'ultime elezioni. Le ragioni di questa scelta sono due. In primo luogo, come anticipato, l'obiettivo ultimo è valutare l'abilità di ogni istituto demoscopico di misurare con efficacia le preferenze di voto della popolazione attraverso i sondaggi pre-elettorali. Se si volesse, ad esempio, stimare questo errore per l'elezioni Europee del 2019, non servirebbero a nulla i sondaggi pubblicati posteriormente a quest'ultima. Pertanto risulta, oltre che corretto, schematicamente più pulito suddividere la serie storica di conseguenza. Secondariamente, impostando il problema in questa maniera, si ha la possibilità di sfruttare il valore della percentuale di voto di un determinato partito per una determinata agenzia ad una determinata elezione come valore iniziale per lo sviluppo del filtro di Kalman.

3.6.1 Il modello utilizzato

Dato che non sono disponibili sondaggi quotidianamente, per dare una continuità temporale alla stima delle intenzioni di voto per ciascun partito assumiamo l'esistenza,

per ogni giorno t, di un vero valore di intenzione di voto $\pi_t \in (0,1)$ che evolve nel tempo in modo regolare.

Per il k-esimo partito, indichiamo con y_{ijkt} il numero di intervistati che dichiarano di votarlo nelle prossime elezioni j, nel sondaggio condotto dall'i-esimo istituto nel giorno t, su un campione di ampiezza u_{ijkt}. Per non appesantire la notazione più del dovuto d'ora in avanti si farà riferimento solamente all'indice temporale, quindi ad esempio al posto di y_{ijkt} si usa y_t. Se nel giorno t non è disponibile alcun sondaggio dell'istituto i, il corrispondente valore è posto come mancante, pertanto le serie storiche presentano in generale molti più dati mancanti che osservati.

Per descrivere l'evoluzione temporale delle intenzioni di voto si lavora sulla scala logit:

$$\theta_t = \text{logit}(\pi_t),$$

e si assume per θ_t un modello basato sul LLT - si veda sezione 3.2.1. Dunque il modello utilizzato è:

$$\begin{aligned} p(y_t \mid \theta_t) &= \text{Bin}\left(u_t, \frac{\exp\{\theta_t\}}{\exp\{1+\theta_t\}}\right), \quad u_t = \text{campione}_t \\ \theta_{t+1} &= \theta_t + \nu_t + \eta_t, & \eta_t &\sim N(0, \sigma_\eta^2) \\ \nu_{t+1} &= \nu_t + \xi_t & \xi_t &\sim N(0, \sigma_\xi^2) \end{aligned}$$

con ν_t che rappresenta la velocità di variazione dell'intenzioni di voto. In altre parole, il logit delle intenzioni di domani è uguale a quello di oggi più il termine ν_t e uno shock casuale, mentre il ν_t stesso si evolve nel tempo come un *random walk*. Questo garantisce una dinamica sufficientemente flessibile ma al tempo stesso regolare, che permette di colmare razionalmente i giorni senza sondaggi, sfruttando l'informazione dei giorni precedenti.

Nel complesso otteniamo quindi un modello *state-space* binomiale con *local linear trend* per le intenzioni di voto. I parametri σ_η^2 e σ_ξ^2 vengono stimati attraverso la massima verosimiglianza e l'inferenza sulla traiettoria latente di π_t viene effettuata tramite il filtro di Kalman approssimato attraverso il metodo descritto in sotto-sezione 3.5.1.

Per quanto riguarda l'inizializzazione del filtro di Kalman, per ogni sezione si specifica uno stato iniziale $(\theta_1, \nu_1)^\top$ coerente con quanto specificato precedentemente. In particolare, per il livello θ_1 si assume una distribuzione iniziale gaussiana centrata sul risultato dell'elezione all'inizio dell'intervallo: se indichiamo con π_{elez} il risultato

in quell'elezione, allora la media iniziale del livello è fissata pari al suo logit,

$$\theta_1 \sim N\left(\log\frac{\pi_{\text{elez}}}{1-\pi_{\text{elez}}}, \sigma_\theta^2\right),$$

con la varianza σ_θ^2 che è posta uguale alla varianza empirica delle intenzioni osservate Y_t/u_t nell'intervallo considerato. Per la *slope* iniziale ν_1 si adotta invece un'inizializzazione diffusa: la sua media è posta pari a zero, mentre la varianza viene considerata infinita, come spiegato in 3.3.4.

In conclusione, si calcolano anche intervalli di confidenza per le stime filtrate degli stati. Nel caso di modelli non gaussiani, gli intervalli vengono costruiti innanzitutto sulla scala del predittore lineare e successivamente trasformati sulla scala delle probabilità. Nel caso di studio, se $\hat{\theta}_t$ indica la stima filtrata del logit delle intenzioni di voto e P_t la corrispondente varianza, l'intervallo di confidenza al 95% per $\hat{\pi}_t = \text{logit}^{-1}(\hat{\theta}_t)$ si ottiene applicando la trasformazione logistica (logit inverso) agli estremi:

$$\pi_t^{\text{L}} = \frac{\exp\left\{\hat{\theta}_t - z_{0.975}\sqrt{P_t}\right\}}{1+\exp\left\{\hat{\theta}_t - z_{0.975}\sqrt{P_t}\right\}},$$

$$\pi_t^{\text{U}} = \frac{\exp\left\{\hat{\theta}_t + z_{0.975}\sqrt{P_t}\right\}}{1+\exp\left\{\hat{\theta}_t + z_{0.975}\sqrt{P_t}\right\}},$$

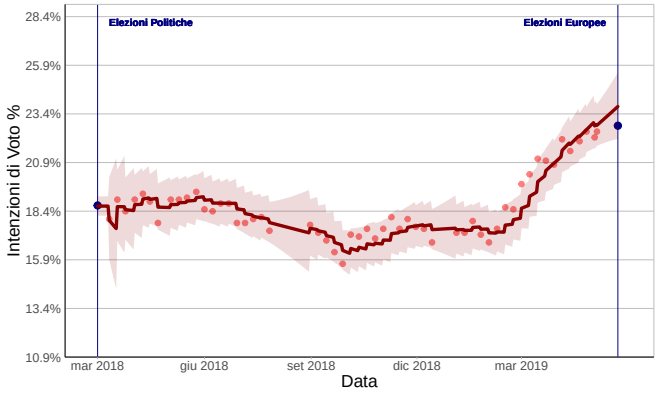
con $z_{0.975}$ che è il quantile al 97.5% della distribuzione Normale standard e π_t^{L}, π_t^{U} indicano rispettivamente il limite inferiore e superiore dell'intervallo di confidenza per π_t.

3.7 Risultati filtro e costruzione dataset per l'analisi

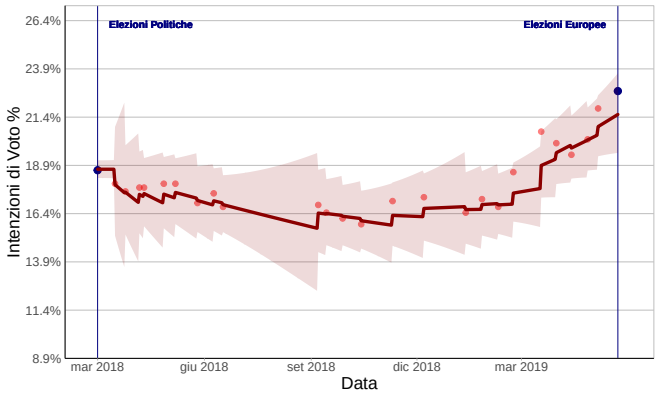
In ultimo si discute di alcuni risultati legati all'applicazione del filtro di Kalman al problema di studio descritto in sezione precedente 3.6. Infatti per come è stato schematizzato il problema si ottiene per ogni istituto demoscopico ed ogni partito, le stime filtrate del livello delle intenzioni di voto negli intervalli di tempo d'interesse definiti precedentemente.

In figura 3.1a si riporta per l'intervallo che spazia dalle elezioni Politiche del 2018 alle elezioni Europee del 2019, le stime filtrate del livello delle intenzioni di voto per il Partito Democratico associate all'agenzia Swg. Non solo, vengono mostrati anche i sondaggi pubblicati da quest'ultima e gli intervalli di confidenza. In figura 3.1b, si riporta, invece, per lo stesso intervallo di tempo di riferimento e sempre

per il Partito Democratico, le stime filtrate del livello delle intenzioni di voto, ma calcolate interpolando i sondaggi pubblicati da Euromedia. É immediato notare come le rilevazioni demoscopiche diffuse da Euromedia in quest'intervallo siano meno numerose rispetto a quelle dell'agenzia Swg. Questo fatto implica che gli intervalli di confidenza siano più ampi, in quanto meno osservazioni si ha a disposizione maggiore incertezza sulle stime filtrate si ottiene.



(a) Stima filtrata del livello delle intenzioni di voto per il partito PD nel primo intervallo d'interesse per l'agenzia Swg.



(b) Stima filtrata del livello delle intenzioni di voto per il partito PD nel primo intervallo d'interesse per l'agenzia Euromedia.

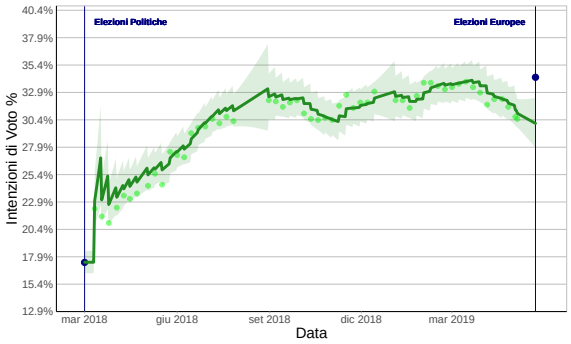
Figura 3.1: Stima filtrata del livello delle intenzioni di voto per il partito PD nel primo intervallo d'interesse (Politiche 2018 - Europee 2019) rispettivamente per **(a)** l'agenzia Swg e **(b)** l'agenzia Euromedia. I punti rossi sono i sondaggi pre-elettorali pubblicati rispettivamente dai due istituti nel periodo considerato.

Anche in figura 3.2, per un altro partito, si nota che quando si ha un periodo all'interno dell'intervallo senza rilevazioni, gli intervalli diventano più ampi. Nei lunghi periodi senza osservazioni, il filtro si basa esclusivamente sulle dinamiche predittive del modello, il che porta ad un ampliamento degli intervalli di confidenza. Una

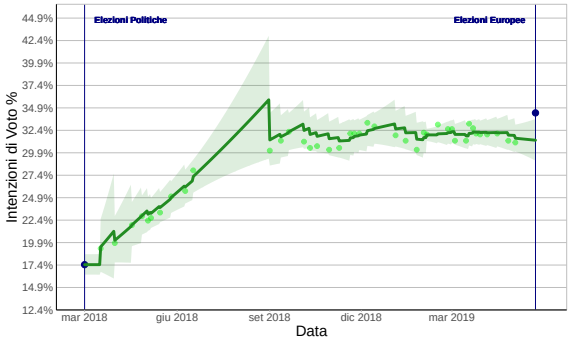
volta che una nuova osservazione diventa disponibile, il filtro si ricalibra, adeguando nettamente la stima del livello e correggendo eventuali deviazioni accumulate. Si consulti l'appendice C per una panoramica completa su tutti i grafici prodotti attraverso l'applicazione di questo processo.

Si procede con la valutazione dell'affidabilità degli istituti demoscopici selezionati, confrontando per ciascuna elezione, la stima filtrata del livello delle intenzioni di voto ottenuta per ogni agenzia con il risultato effettivo ottenuto dal partito in questione.

Pertanto il dataset ottenuto è una copia di quello definito in sezione 2.5. É composto dalle tre variabili categoriche indipendenti `Istituto`, `Partito` ed `Elezione` e dalla variabile risposta definita come in equazione (2.1), con la modifica che ora $\hat{V}_{ikj}$ è la stima filtrata del livello delle intenzioni di voto per l'istituto i, per il partito k all'elezione j, ottenuta attraverso l'applicazione del filtro di Kalman.



(a) Stima filtrata del livello delle intenzioni di voto per il partito Lega nel primo intervallo d'interesse per l'agenzia Swg.



(b) Stima filtrata del livello delle intenzioni di voto per il partito Lega nel primo intervallo d'interesse per l'agenzia Tecnè.

Figura 3.2: Stima filtrata del livello delle intenzioni di voto per il partito Lega nel primo intervallo d'interesse (Politiche 2018 - Europee 2019) rispettivamente per **(a)** l'agenzia Swg e **(b)** l'agenzia Tecnè. I punti verdi sono i sondaggi pre-elettorali pubblicati rispettivamente dai due istituti nel periodo considerato.

Capitolo 4

Analisi dell'errore

In questo capitolo, l'ultimo della tesi, si descrive l'analisi portata avanti sui due dataset ottenuti rispettivamente alla fine del capitolo 2 e 3. Precisamente i due dataset sono speculari, in quanto condividono le stesse variabili indipendenti `Istituto`, `Partito` e `Elezione`. Ciò che li distingue è la formulazione della variabile risposta; in particolare, la scelta della stima delle intenzioni di voto fornita da un'agenzia, per un partito ad una determinata elezione da confrontare con gli effettivi risultati dell'elezione stessa. Infatti, il dataset ottenuto con il capitolo 2 prende a riferimento, come stima, l'ultimo sondaggio pubblicato da ogni agenzia, mentre quello ricavato alla fine del capitolo 3 prende a riferimento l'interpolazione dei sondaggi pubblicati da una determinata agenzia, per un determinato partito prima di una certa elezione, calcolata il giorno dell'elezioni stesso. Quest'ultima viene ottenuta attraverso il filtro di Kalman. La variabile risposta è dunque definita come in (2.1).

In sezione 4.1 si studia l'eventuale presenza di distorsioni sistematiche nelle previsioni delle intenzioni di voto fornite dalle agenzie demoscopiche per ogni partito e turno elettorale. Mentre in sezione 4.2 ci si concentra sulla precisione della previsione di ogni istituto demoscopico, in particolare quello che si vuole capire è se esistono agenzie più affidabili di altre.

4.1 Analisi sulla presenza di distorsioni sistematiche

La prima analisi riguarda la verifica della presenza di eventuali distorsioni sistematiche nelle previsioni delle intenzioni di voto fornite dalle agenzie demoscopiche. In figura 4.1 si mostrano i boxplot della variabile risposta, definita in 4.1a come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo e in 4.1b come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni, condizionata alla variabile `Istituto`. In entrambi i casi non emergono sistematicità evidenti,

suggerendo che le agenzie demoscopiche non differiscano metodicamente in termini di errore relativo.

Diversa è la situazione illustrata nella figura 4.2, in cui sono riportati i boxplot della variabile risposta condizionata congiuntamente alle variabili `Partito` ed `Elezione`. In questo caso compaiono chiari segnali di sistematicità, indicando che tali fattori contribuiscono a spiegare la variabilità dell'errore. Inoltre, i due grafici mostrano come la risposta cambi in modo diverso a seconda della combinazione tra `Partito` ed `Elezione`. Ciò suggerisce che l'effetto dei due fattori insieme non coincida con la somma dei loro effetti separati: in altre parole, le due variabili fattoriali interagiscono. Un modello adeguato deve quindi includere anche il relativo termine d'interazione, poiché l'effetto di un fattore risulta dipendere dal livello assunto dall'altro.

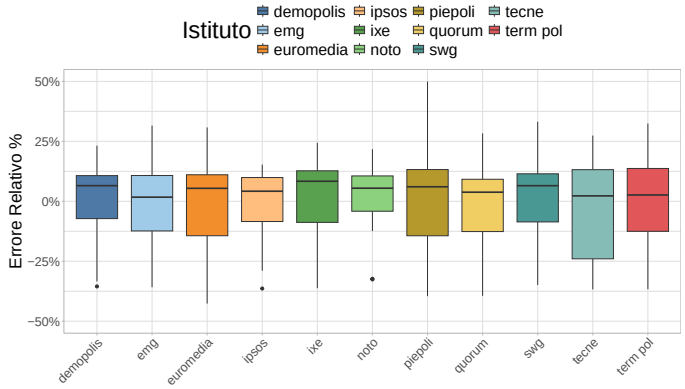
Il modello adottato è un modello lineare semplice. Inizialmente si considerano tutte le variabili indipendenti a disposizione: `Istituto`, `Partito` ed `Elezione`, includendo anche il loro termine d'interazione. Pertanto si può scrivere il modello come:

$$\begin{aligned}\texttt{rel.err}_i = \beta_0 &+ \beta_1\,\texttt{Istituto}_i + \beta_2\,\texttt{Partito}_i \\ &+ \beta_3\,\texttt{Elezione}_i + \beta_4\,(\texttt{Partito}_i \times \texttt{Elezione}_i).\end{aligned} \tag{4.1}$$

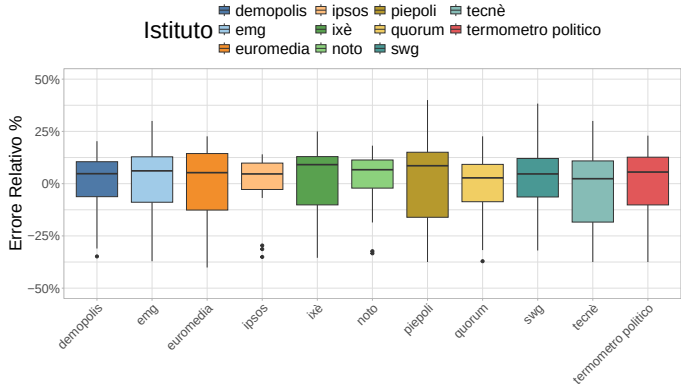
con $i = 1, \ldots, 165$. Per verificare quanto suggerito dal grafico 4.1 — ovvero che le diverse agenzie demoscopiche non contribuiscono a spiegare la variabilità dell'errore — si ricorre all'analisi della varianza (ANOVA). Come discusso in Gelman & Hill (2006), l'ANOVA consente di valutare l'importanza relativa delle diverse fonti di variazione presenti nel dataset. In questo contesto, essa permette di quantificare quanta parte della variabilità nei dati sia attribuibile rispettivamente ai fattori `Istituto`, `Partito` ed `Elezione`, e di verificare quanta variabilità residua rimanga una volta che tutti questi effetti sono stati inclusi nel modello.

Nelle tabelle 4.1 sono riportati i risultati dell'analisi della varianza per il modello lineare specificato in equazione (4.1) considerando entrambe le definizioni di variabile risposta. Specificatamente, nella tabella 4.1a l'errore relativo è calcolato rispetto all'interpolazione nel giorno delle elezioni, mentre nella tabella 4.1b rispetto all'ultimo sondaggio disponibile prima del voto.

Ogni riga delle tabelle si riferisce a un insieme di coefficienti associati alle componenti additive del modello, che in questo caso corrispondono alle tre covariate e al termine di interazione tra `Partito` ed `Elezione`. La colonna Df indica i gradi di libertà associati a ciascun insieme di coefficienti, determinati dal numero di categorie meno il vincolo di identificabilità imposto in un modello lineare classico. Ad esempio,



(a) Boxplot della variabile risposta, definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo, per `Istituto`.



(b) Boxplot della variabile risposta, definita come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni, per `Istituto`.

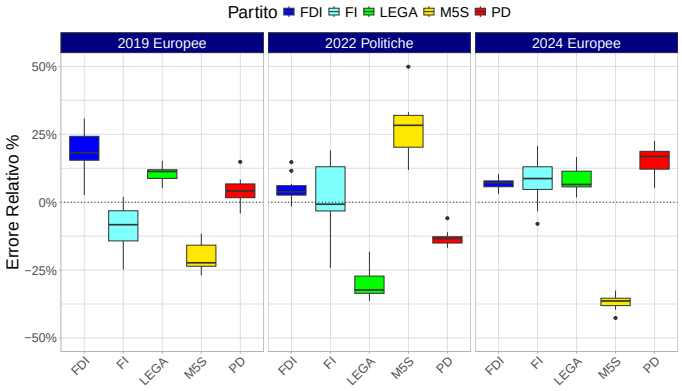
Figura 4.1: Boxplot della variabile risposta per la variabile indipendente `Istituto`. In **(a)** viene definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo, mentre in **(b)** come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

gli 11 istituti presenti nel dataset generano 10 gradi di libertà, poiché uno di essi deve fungere da categoria di riferimento.

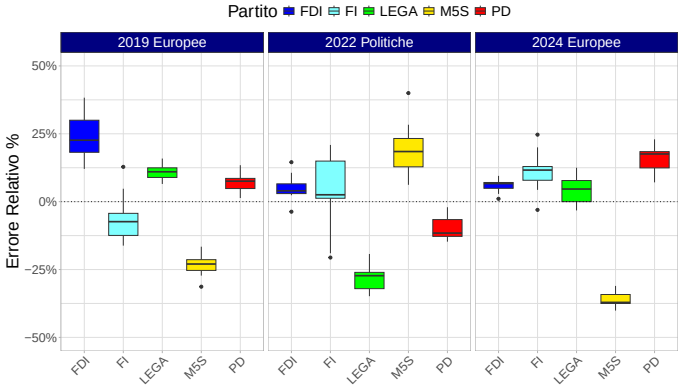
I valori riportati nella colonna Sum Sq rappresentano la quantità di variabilità nella risposta attribuibile a ciascuna componente del modello. In un dataset bilanciato, la somma di tali quantità coincide con la *total sum of squares*, ossia

$$\sum_{i=1}^{165} (\text{rel.err}_i - \overline{\text{rel.err}_i})^2.$$

La colonna Mean Sq è ottenuta dividendo la somma dei quadrati per i relativi gradi di libertà, e ricopre un ruolo centrale nell'ANOVA poiché è utilizzata per



(a) Boxplot della variabile risposta, definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo, per ogni combinazione di `Partito` - `Elezione`.



(b) Boxplot della variabile risposta, definita come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni, per ogni combinazione di `Partito` - `Elezione`.

Figura 4.2: Boxplot della variabile risposta per ogni combinazione di `Partito` - `Elezione`. In **(a)** viene definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo, mentre in **(b)** come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

costruire le statistiche F. Per ogni riga della tabella, infatti, la statistica F è il rapporto tra la *mean square* dell'effetto e quella dei residui: se tale rapporto non risulta significativamente maggiore di 1, l'effetto considerato non fornisce un contributo apprezzabile alla spiegazione della variabilità dei dati.

Nella tabella 4.1a, ad esempio, il valore della *mean square* associata alla variabile `Istituto` non differisce in modo statisticamente significativo da quella dei residui. Ciò indica che la variabilità della risposta tra gli istituti non supera quella che ci si aspetterebbe, dati i livelli di variabilità presenti nel dataset. Dunque la variabile

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Istituto	10	0.0418	0.00418	0.9234	0.5137
Partito	4	0.7314	0.18286	40.3507	$< 2 \cdot 10^{-16}$
Elezione	2	0.0211	0.01056	2.3306	0.1010
Partito:Elezione	8	4.2531	0.53163	117.3138	$< 2 \cdot 10^{-16}$
Residui	140	0.6344	0.00453		

(a) ANOVA per il modello (4.1), con la variabile risposta definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo.

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Istituto	10	0.0246	0.00246	0.5552	0.8478
Partito	4	1.1866	0.29665	67.0736	$< 2.2 \cdot 10^{-16}$
Elezione	2	0.0491	0.02457	5.5558	0.0048
Partito:Elezione	8	3.5106	0.43882	99.2202	$< 2.2 \cdot 10^{-16}$
Residui	140	0.6192	0.00442		

(b) ANOVA per il modello (4.1), con la variabile risposta definita come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

Tabella 4.1: Risultati dell'ANOVA per il modello lineare definito come in (4.1). In **(a)** la variabile risposta è definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo. In **(b)** la variabile risposta è definita come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

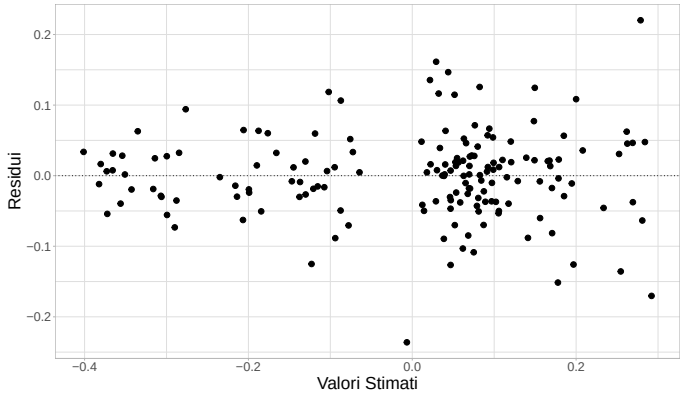
`Istituto` non fornisce un contributo informativo rilevante e la sua inclusione non migliora in modo significativo la capacità del modello di spiegare la variabile risposta. Al contrario, per la variabile `Partito` si osserva un p-value estremamente piccolo: questo porta a rifiutare l'ipotesi nulla che sostiene che la variabile in questione non abbia effetto sulla risposta e implica che la variabilità tra partiti è ben superiore a quella attesa.

Considerazioni analoghe valgono anche per la tabella 4.1b, ottenuta utilizzando la seconda definizione di variabile risposta.

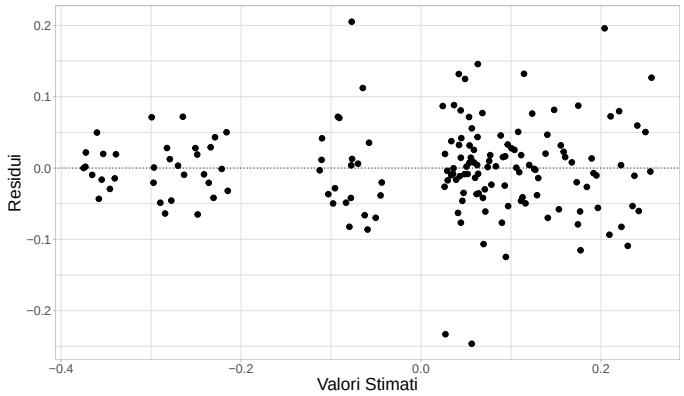
Alla luce dei risultati ottenuti dall'analisi della varianza, appare ragionevole stimare un nuovo modello lineare che escluda la variabile indipendente `Istituto`, mantenendo unicamente `Partito`, `Elezione` e il loro termine di interazione. Dunque si definisce il seguente modello:

$$\begin{aligned} \texttt{rel.err}_i = \beta_0 &+ \beta_1 \, \texttt{Partito}_i \\ &+ \beta_2 \, \texttt{Elezione}_i + \beta_3 \, (\texttt{Partito}_i \times \texttt{Elezione}_i). \end{aligned} \tag{4.2}$$

con $i = 1, \ldots, 165$.



(a) Residui vs Valori stimati per il modello (4.1), con la variabile risposta definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo.



(b) Residui vs Valori stimati per il modello (4.1), con la variabile risposta definita come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

Figura 4.3: Grafico dei Residui vs Valori Stimati per il modello (4.1). In **(a)** la variabile risposta viene definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo, mentre in **(b)** come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

Tuttavia, dall'analisi grafica dei residui riportata in figura 4.3, con i due grafici 4.3a e 4.3b, emerge la possibile presenza di eteroschedasticità, ovverosia la varianza degli errori non risulta costante su tutte le osservazioni. In presenza di eteroschedasticità, le stime OLS continuano ad essere non distorte, ma gli usuali test possono risultare inappropriati. Pertanto è necessario tenere conto di questa possibile violazione delle assunzioni, altrimenti si potrebbe arrivare a conclusioni inferenziali errate.

Dunque si confrontano i modelli (4.1) e (4.2) utilizzando una procedura robusta all'eteroschedasticità. In particolare, il confronto è stato formulato come un test di

Wald sui vincoli che impongono a zero i coefficienti associati alla variabile `Istituto`, utilizzando una matrice di varianza–covarianza robusta all'eteroschedasticità (si legga Long & Ervin (2000)). In questo contesto, il test verifica l'ipotesi nulla secondo cui l'inclusione della variabile `Istituto` non apporta un contributo informativo significativo e che quindi risulta preferibile un modello più parsimonioso che comprenda esclusivamente `Partito`, `Elezione` e la loro interazione. La statistica test si può scrivere come:

$$F = \frac{1}{k_2}(R\hat{\beta} - r)^\top \left[R\widehat{Var}(\hat{\beta})R^\top\right]^{-1}(R\hat{\beta} - r) \sim F_{k_2, n-k_1-k_2} \tag{4.3}$$

con R matrice $k_2 \times (k_1 + k_2)$ che seleziona i coefficienti associati alla variabile esplicativa `Istituto`, r che è un vettore $k_2 \times 1$ di zeri e k_2 che rappresenta il numero di vincoli imposti Inoltre la matrice di covarianza è stimata nel seguente modo,

$$\widehat{Var}(\hat{\beta}) = (X^\top X)^{-1} X^\top \text{diag}\left[\frac{\hat{\epsilon}_i^2}{(1-h_{ii})^2}\right] X(X^\top X)^{-1}, \tag{4.4}$$

con $\hat{\epsilon}_i$ residuo i-esimo e $h_{ii} = x_i(X^\top X)x_i^\top$ elemento i-esimo sulla diagonale della matrice di proiezione H.

Il test di Wald robusto restituisce una statistica F rispettivamente pari a circa 0.8707 e 0.8483 per le due definizioni di variabile risposta, con associati i p-value pari a 0.5622 e 0.5832. Pertanto, anche adottando una procedura inferenziale robusta rispetto all'eteroschedasticità, l'ipotesi nulla non viene rifiutata. Questo risultato conferma quanto già emerso dall'analisi ANOVA classica: l'inclusione della variabile `Istituto` non migliora in modo statisticamente significativo la capacità esplicativa del modello.

È importante sottolineare che, sebbene le conclusioni sostanziali rimangano invariate, l'utilizzo di standard error robusti rafforza l'affidabilità dei risultati, rendendo il confronto tra modelli meno sensibile a eventuali violazioni delle assunzioni del modello lineare. Alla luce di queste evidenze, appare dunque giustificato adottare il modello più parsimonioso che esclude la variabile `Istituto` per la continuazione dell'analisi.

Una volta identificato il modello di interesse, definito in (4.2), è necessario analizzare la possibile presenza di distorsioni sistematiche nelle stime. A tal fine, l'attenzione si concentra sugli intervalli di confidenza delle medie previste della variabile risposta per ciascuna combinazione di `Partito` ed `Elezione`. L'obiettivo è verificare se tali medie risultino statisticamente diverse da zero.

Nel caso in cui l'intervallo di confidenza associato a una data combinazione includa lo zero, vorrebbe dire che non vi sono evidenze di *bias* sistematici, in altre parole indicherebbe che le agenzie demoscopiche tendono a sovrastimare o sottostimare il risultato elettorale in modo puramente casuale. Al contrario, medie significativamente diverse da zero fornirebbero indicazioni della presenza di distorsioni sistematiche, suggerendo un errore medio non nullo ed in una certa direzione per quella specifica combinazione di fattori. Tutte le combinazioni di `Partito` ed `Elezione` considerate nell'analisi sono riportate nella tabella 4.2.

Tabella 4.2: Combinazioni di `Partito` ed `Elezione`.

Partito	**Elezione**
FDI	2019 Europee
FDI	2022 Politiche
FDI	2024 Europee
FI	2019 Europee
FI	2022 Politiche
FI	2024 Europee
LEGA	2019 Europee
LEGA	2022 Politiche
LEGA	2024 Europee
M5S	2019 Europee
M5S	2022 Politiche
M5S	2024 Europee
PD	2019 Europee
PD	2022 Politiche
PD	2024 Europee

Tuttavia, come discusso in precedenza, nei dati è presente eteroschedasticità. Sebbene in tale contesto le stime OLS dei coefficienti rimangano non distorte, consistenti e asintoticamente normali, lo stimatore usuale della matrice di covarianza non è più consistente e risulta distorto. Di conseguenza, per svolgere correttamente inferenza sui parametri della regressione — sia sui singoli coefficienti, come nel caso precedente, sia su combinazioni lineari di essi, come si intende fare ora — è necessario introdurre opportuni aggiustamenti e ricorrere a uno stimatore consistente della matrice di covarianza.

Esistono diversi stimatori *heteroskedasticity-consistent* della matrice di covarianza degli stimatori OLS; tra questi rientra quello definito in (4.4). Tali stimatori sono consistenti sia in presenza di omoschedasticità sia in presenza di eteroschedasticità di forma ignota.

Il modello d'interesse è quello lineare:

$$y = X\beta + \epsilon,$$

con y e ϵ vettori $n \times 1$ della risposta e degli errori rispettivamente, X matrice $n \times k$ dei regressori e $\beta = (\beta_0, ..., \beta_{k-1})^\top$ vettore $k \times 1$ dei coefficienti della regressione. L'i-esimo errore ha media zero e varianza σ_i^2; inoltre gli errori sono incorrelati, i.e. $\mathbb{E}[\epsilon_i \epsilon_j] = 0$, $\forall i \neq j$. Dunque la matrice di covarianza di ϵ è $\Omega = \text{diag}(\sigma_1^2, \ldots, \sigma_n^2)$. Lo stimatore OLS dei parametri è $\hat{\beta} = (X^\top X)^{-1} X^\top y$ e la sua matrice di covarianza è $\Psi = (X^\top X)^{-1} X^\top \sigma^2 X (X^\top X)^{-1}$. Se l'assunzione di omoschedasticità non è violata, $\Psi = \sigma^2 (X^\top X)^{-1}$ e $\hat{\Psi}$ si stima sotituendo a σ^2, che è comune a tutti gli errori, l'espressione $\hat{\sigma}^2 = \hat{\epsilon}^\top \hat{\epsilon}/(n-k)$. Nel caso, invece, dovesse essere presente eteroschedasticità il vettore β continua ad essere stimato con OLS, ma è accoppiato con uno stimatore della struttura di covarianza consistente che permetta di performare test d'ipotesi e costruire intervalli di confidenza coerenti. Poiché la forma di eteroschedasticità non è nota è necessario utilizzare una matrice HCCM[1]. In White (1980) si è ottenuto un primo stimatore di questo genere

$$\text{HC0} = \hat{\Psi} = (X^\top X)^{-1} X^\top \hat{\Omega} X (X^\top X)^{-1}$$

con $\hat{\Omega} = \text{diag}\{\hat{\epsilon}_1^2, \ldots \hat{\epsilon}_n^2\}$. Lo stimatore HC0 tende ad essere piuttosto distorto per campioni di dimensioni piccola o moderata. In particolare propende a sottostimare le vere varianze. In questo senso sono state proposte delle varianti di questo stimatore. Quest'ultime sono allo stesso modo consistenti in presenza o assenza di eteroschedasticità. Lo stimatore HC1 è espresso come:

$$\text{HC1} = \hat{\Psi}_1 = (X^\top X)^{-1} X^\top E_1 \hat{\Omega} X (X^\top X)^{-1},$$

con $E_1 = n/(n-k)I$. Lo stimatore HC2 è espresso come

$$\text{HC2} = \hat{\Psi}_2 = (X^\top X)^{-1} X^\top E_2 \hat{\Omega} X (X^\top X)^{-1},$$

con $E_2 = \text{diag}\{1/(1-h_{ii})\}$. Infine il terzo stimatore, quello utilizzato in (4.4) - e che verrà utilizzato anche successivamente -, si può scrivere come

$$\text{HC3} = \hat{\Psi}_3 = (X^\top X)^{-1} X^\top E_3 \hat{\Omega} X (X^\top X)^{-1},$$

[1] *Heteroscedasticity consistent covariance matrix.*

con $E_3 = \text{diag}\{1/(1-h_{ii})^2\}$. Si noti che sia lo stimatore HC2 che HC3 correggono per il termine h_{ii}, che rappresenta l'influenza dell'osservazione i-esima.

L'impiego di queste matrici di covarianza corrisponde, nella pratica, a quella che viene comunemente definita correzione per standard error robusti. Infatti, la radice quadrata degli elementi sulla diagonale di tali stimatori fornisce stime consistenti degli errori standard dei coefficienti anche in presenza di eteroschedasticità. È per questo motivo che tali errori standard vengono detti robusti: in campioni sufficientemente grandi, essi permettono di costruire test d'ipotesi e intervalli di confidenza affidabili facendo assunzioni minime sulla struttura dei dati e del modello.

Nel contesto applicativo considerato, per ciascuna combinazione delle variabili esplicative `Partito` ed `Elezione`, identificata dal vettore c, si è interessati a testare l'ipotesi nulla

$$H_0 : c^\top \beta = 0.$$

Poiché l'inferenza viene corretta per la presenza di eteroschedasticità, la statistica test associata non segue più esattamente una distribuzione t di Student. Il test risultante non è quindi un t-test in senso stretto, ma viene indicato come quasi-t test (si veda Cribari-Neto & Silva (2011)). Tuttavia, nella pratica statistica si continua a utilizzare il quantile della distribuzione t di Student per la costruzione degli intervalli di confidenza[2], ottenendo quindi l'intervallo riportato di seguito.

$$\text{IC}_{1-\alpha}(c^\top \hat{\beta}) = c^\top \hat{\beta} \ \pm \ t_{1-\alpha/2,\, n-k} \sqrt{c^\top \widehat{\text{Var}}(\hat{\beta}) c} \tag{4.5}$$

con k numero di coefficienti della regressione ed n numero di osservazioni.

Nelle tabelle 4.3a e 4.3b vengono mostrati gli intervalli di confidenza e la media prevista per ogni combinazione tra le modalità delle variabili indipendenti `Partito` ed `Elezione`. Tuttavia, l'esecuzione di una sequenza di test d'ipotesi comporta un aumento della probabilità di incorrere in risultati apparentemente significativi ma in realtà spurii, ossia in errori di Tipo I. Questo fenomeno, noto come problema delle comparazioni multiple, implica che la probabilità di osservare almeno un rifiuto erroneo dell'ipotesi nulla cresce all'aumentare del numero di ipotesi testate. Di conseguenza, considerare i risultati riportati nelle tabelle in 4.3 senza tenere conto di tale aspetto può condurre a interpretazioni fuorvianti e a conclusioni non giustificate dai dati.

[2] Il pacchetto di R Fox & Weisberg (2019) li costruisce in questa maniera.

Tabella 4.3: Intervalli di confidenza per la media prevista associata ad ogni combinazione `Partito - Elezione`. In **(a)** l'errore relativo è definito come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo. In **(b)** la variabile risposta viene definita come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

(a) Intervalli di confidenza per la media prevista associata ad ogni combinazione `Partito - Elezione`. L'errore relativo è definito tra interpolazione nel giorno delle elezioni e risultato effettivo.

Partito	Elezione	fit	lwr	upr
FDI	2019 Europee	0.183	0.114	0.252
FI	2019 Europee	-0.089	-0.159	-0.020
LEGA	2019 Europee	0.104	0.078	0.130
M5S	2019 Europee	-0.201	-0.245	-0.158
PD	2019 Europee	0.045	0.004	0.086
FDI	2022 Politiche	0.051	0.013	0.089
FI	2022 Politiche	0.027	-0.079	0.133
LEGA	2022 Politiche	-0.301	-0.346	-0.257
M5S	2022 Politiche	0.267	0.177	0.357
PD	2022 Politiche	-0.132	-0.156	-0.108
FDI	2024 Europee	0.067	0.050	0.085
FI	2024 Europee	0.080	0.011	0.149
LEGA	2024 Europee	0.085	0.049	0.121
M5S	2024 Europee	-0.368	-0.391	-0.344
PD	2024 Europee	0.154	0.110	0.198

(b) Intervalli di confidenza per la media prevista associata ad ogni combinazione tra `Partito-Elezione`. L'errore relativo è definito tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

Partito	Elezione	fit	lwr	upr
FDI	2019 Europee	0.237	0.172	0.303
FI	2019 Europee	-0.062	-0.134	0.009
LEGA	2019 Europee	0.108	0.084	0.132
M5S	2019 Europee	-0.234	-0.267	-0.200
PD	2019 Europee	0.071	0.042	0.101
FDI	2022 Politiche	0.051	0.012	0.090
FI	2022 Politiche	0.044	-0.071	0.159
LEGA	2022 Politiche	-0.282	-0.322	-0.243
M5S	2022 Politiche	0.192	0.114	0.270
PD	2022 Politiche	-0.095	-0.133	-0.058
FDI	2024 Europee	0.059	0.039	0.080
FI	2024 Europee	0.111	0.051	0.172
LEGA	2024 Europee	0.041	-0.002	0.085
M5S	2024 Europee	-0.358	-0.380	-0.335
PD	2024 Europee	0.155	0.116	0.195

Per questo motivo è cruciale controllare il *family-wise error rate* (FWER), definito come la probabilità di commettere almeno un errore di Tipo I nell'insieme delle comparazioni effettuate. Esistono diversi metodi per adattare il livello di significatività al fine di mantenere il FWER al livello desiderato. Nel presente studio si è adottato l'approccio più semplice e diretto, ossia la correzione di Bonferroni. L'idea alla base è immediata: il livello di significatività prefissato α viene diviso per il numero di test condotti m, ottenendo una soglia di significatività corretta pari a α/m, che garantisce il controllo del FWER. La correzione di Bonferroni, però, può risultare piuttosto conservativa. Infatti, la riduzione della probabilità di incorrere in errori di Tipo I non è priva di costi, in quanto comporta un aumento della probabilità di commettere errori di Tipo II, ossia di non rifiutare l'ipotesi nulla quando essa è in realtà falsa.

Dunque, l'applicazione della correzione di Bonferroni comporta che gli intervalli di confidenza riportati nelle tabelle in 4.4 risultino, come atteso, leggermente più ampi rispetto a quelli presentati in 4.3. Tale correzione riflette la natura conservativa della procedura. In ogni modo, in questo caso, non altera in maniera sostanziale i risultati ottenuti, ma rende l'inferenza più prudente.

In figura 4.4 sono rappresentati graficamente gli intervalli di confidenza ottenuti per entrambe le analisi considerate: quella basata sull'interpolazione dei sondaggi fino al giorno delle elezioni in 4.4a e quella che utilizza l'ultimo sondaggio rilevato per ciascuna agenzia in 4.4b. La rappresentazione grafica consente di apprezzare con maggiore immediatezza le differenze tra partiti ed elezioni in termine di distorsioni.

Per quanto riguarda l'analisi basata sull'interpolazione dei sondaggi, il Movimento 5 Stelle risulta sistematicamente stimato in modo errato: sovrastimato nelle elezioni Europee del 2019, sottostimato nelle elezioni Politiche del 2022 e nuovamente sovrastimato nelle elezioni Europee del 2024. La Lega mostra un andamento speculare e contrario, risultando sottostimata nel 2019, fortemente sovrastimata nel 2022 e leggermente sottostimata nel 2024. Al contrario, Forza Italia non sembra risentire di distorsioni sistematiche: in tutte e tre le tornate elettorali considerate lo zero è incluso negli intervalli di confidenza, per cui non vi è evidenza sufficiente per rifiutare, al livello di confidenza α, l'ipotesi di assenza di una deviazione sistematica nella stima di questo partito da parte delle agenzie demoscopiche.

Per Fratelli d'Italia emerge una tendenza a una stima schematicamente distorta in tutte e tre le elezioni, sebbene nel caso delle Politiche del 2022 l'intervallo di confidenza risulti prossimo allo zero. Per il Partito Democratico, invece, nell'elezione europea del 2019 l'intervallo di confidenza comprende lo zero, mentre nelle successive

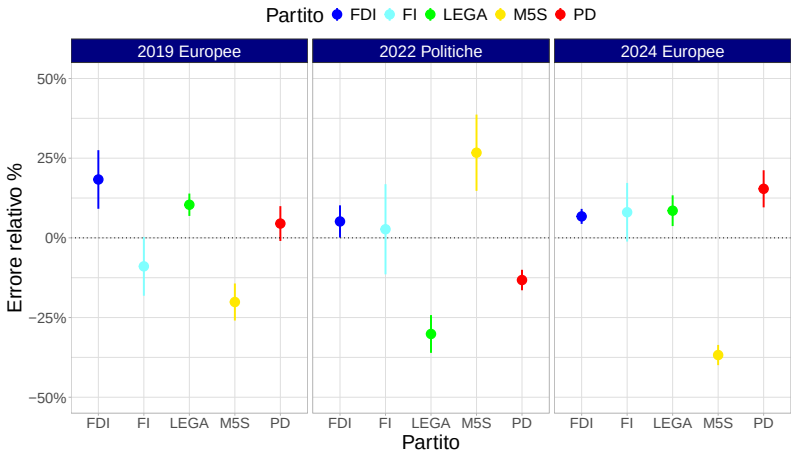
Tabella 4.4: Intervalli di confidenza per la media prevista con correzione di Bonferroni. In **(a)** la variabile risposta è definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo. In **(b)** la variabile risposta viene definita come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

(a) Intervalli di confidenza per la media prevista, con l'errore relativo definito tra interpolazione nel giorno delle elezioni e risultato effettivo con correzione di Bonferroni.

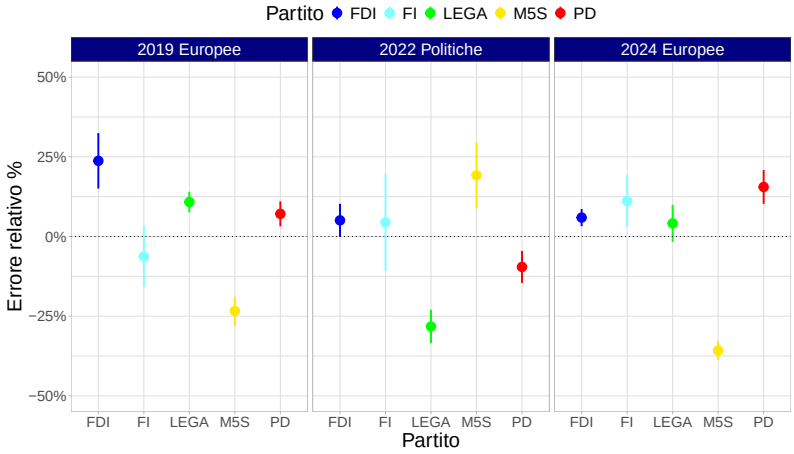
Partito	Elezione	fit	lwr	upr
FDI	2019 Europee	0.183	0.091	0.275
FI	2019 Europee	-0.089	-0.182	0.003
LEGA	2019 Europee	0.104	0.069	0.139
M5S	2019 Europee	-0.201	-0.259	-0.143
PD	2019 Europee	0.045	-0.010	0.100
FDI	2022 Politiche	0.051	0.001	0.102
FI	2022 Politiche	0.027	-0.114	0.168
LEGA	2022 Politiche	-0.301	-0.361	-0.242
M5S	2022 Politiche	0.267	0.147	0.387
PD	2022 Politiche	-0.132	-0.164	-0.100
FDI	2024 Europee	0.067	0.044	0.091
FI	2024 Europee	0.080	-0.011	0.172
LEGA	2024 Europee	0.085	0.037	0.134
M5S	2024 Europee	-0.368	-0.399	-0.336
PD	2024 Europee	0.154	0.096	0.212

(b) Intervalli di confidenza per la media prevista, con l'errore relativo definito tra ultimo sondaggio rilevato e risultato effettivo alle elezioni con correzione di Bonferroni.

Partito	Elezione	fit	lwr	upr
FDI	2019 Europee	0.237	0.150	0.324
FI	2019 Europee	-0.062	-0.158	0.033
LEGA	2019 Europee	0.108	0.076	0.140
M5S	2019 Europee	-0.234	-0.279	-0.189
PD	2019 Europee	0.071	0.032	0.110
FDI	2022 Politiche	0.051	-0.001	0.102
FI	2022 Politiche	0.044	-0.109	0.197
LEGA	2022 Politiche	-0.282	-0.335	-0.230
M5S	2022 Politiche	0.192	0.088	0.296
PD	2022 Politiche	-0.095	-0.146	-0.045
FDI	2024 Europee	0.059	0.032	0.086
FI	2024 Europee	0.111	0.030	0.192
LEGA	2024 Europee	0.041	-0.017	0.099
M5S	2024 Europee	-0.358	-0.388	-0.328
PD	2024 Europee	0.155	0.102	0.209



(a) Intervalli di confidenza corretti per Bonferroni per la media prevista per ogni combinazione di `Partito-Elezione`, con l'errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo.



(b) Intervalli di confidenza corretti per Bonferroni per la media prevista per ogni combinazione di `Partito-Elezione`, con l'errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

Figura 4.4: Intervalli di confidenza corretti per Bonferroni per la media prevista per ogni combinazione di `Partito-Elezione`. In **(a)** la variabile risposta viene definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo, mentre in **(b)** come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

due tornate elettorali si osserva rispettivamente una sovrastima e una sottostima sistematiche.

Passando all'analisi condotta sull'errore relativo basato sull'ultimo sondaggio rilevato, i risultati appaiono nel complesso coerenti con quelli precedenti. Il Movimento 5 Stelle risulta sovrastimato nel 2019, sottostimato nel 2022 e nuovamente sovrastimato nel 2024, mentre la Lega segue un andamento opposto, con una sottostima nel 2019

e una sovrastima nel 2022. Nel 2024, diversamente da quanto osservato nell'analisi basata sull'interpolazione, la Lega non mostra evidenza significativa di distorsione sistematica.

Anche per Forza Italia i risultati sono sostanzialmente analoghi: nelle elezioni del 2019 e del 2024 lo zero è incluso negli intervalli di confidenza, mentre per le politiche del 2022 emerge una lieve sottostima. Fratelli d'Italia risulta invece stimata in modo distorto nella prima e nella terza elezione, ma non nella seconda. Infine, per il Partito Democratico si osserva in sequenza una sottostima, sovrastima e nuovamente sottostima, quest'ultima di entità più contenuta, sistematica.

Nel complesso, la coerenza dei risultati tra le due diverse definizioni di errore relativo rafforza l'evidenza della presenza di distorsioni sistematiche specifiche per partito ed elezione, suggerendo che tali schemi non siano imputabili a scelte arbitrarie nella costruzione della misura di errore, ma riflettano caratteristiche strutturali della stima delle intenzioni di voto provenienti da sondaggi pre-elettorali.

4.2 Analisi affidabilità degli istituti demoscopici

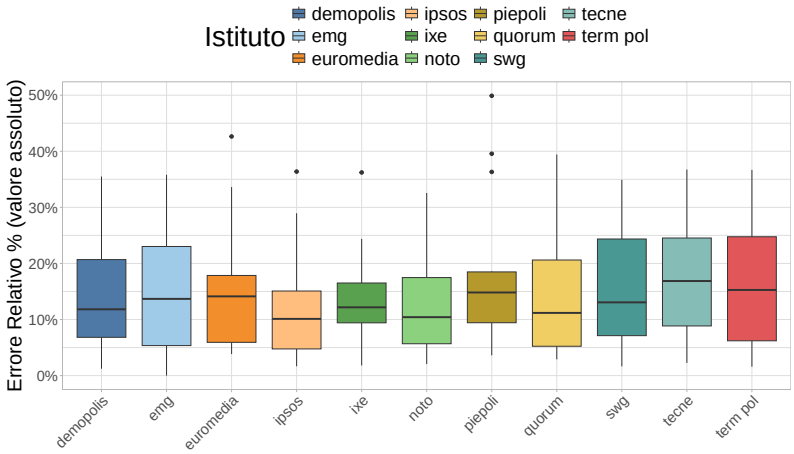
La seconda analisi si concentra sulla differenza di affidabilità tra le agenzie demoscopiche, ossia sulla possibilità che alcune di esse siano più precise di altre nel stimare le intenzioni di voto per una determinata elezione. Nella sezione precedente, come mostrato in tabella 4.1b, è emerso che la variabile `Istituto` non ha un effetto significativo sulla variabile risposta definita in (2.1), indipendentemente dalla modalità con cui viene calcolata la previsione.

Tuttavia, definire la variabile risposta come in (2.1) implica distinguere tra errori di sovrastima ed errori di sottostima. In questo contesto, però, tale distinzione risulta ridondante e secondaria: l'obiettivo non è valutare la direzione dell'errore, bensì la precisione complessiva delle stime fornite dalle diverse agenzie. Per questo motivo l'analisi viene riformulata considerando il valore assoluto dell'errore relativo, così che sovrastime e sottostime della stessa entità vengano trattate in modo equivalente.

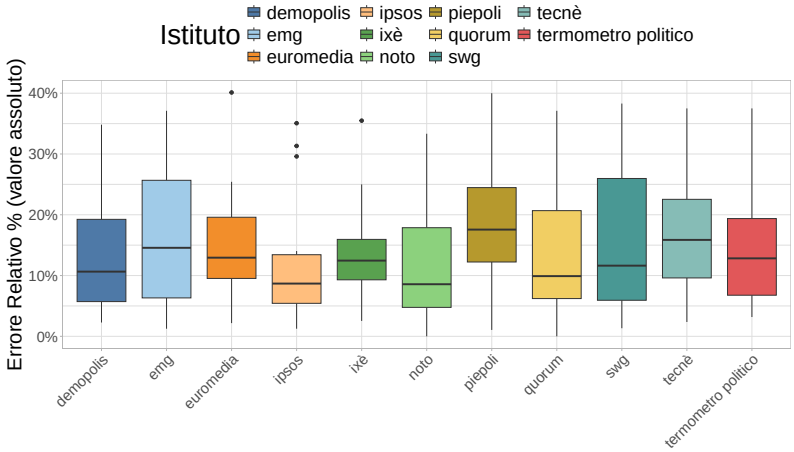
Il passaggio al valore assoluto dell'errore relativo compromette le informazioni dei dati sulla direzione del bias, che vengono totalmente perdute, a favore però di concentrare l'attenzione esclusivamente sulla variabilità e sull'intensità dell'errore.

L'analisi procede quindi in modo analogo a quanto fatto in precedenza in 4.1.

I grafici in figura 4.5 mostrano in modo chiaro come, per entrambe le definizioni della variabile risposta presa in valore assoluto, non sembra vi siano eterogeneità nell'errore, e quindi nell'affidabilità, commesso da parte delle agenzie demoscopiche



(a) Boxplot del valore assoluto della variabile risposta, definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo, per `Istituto`.

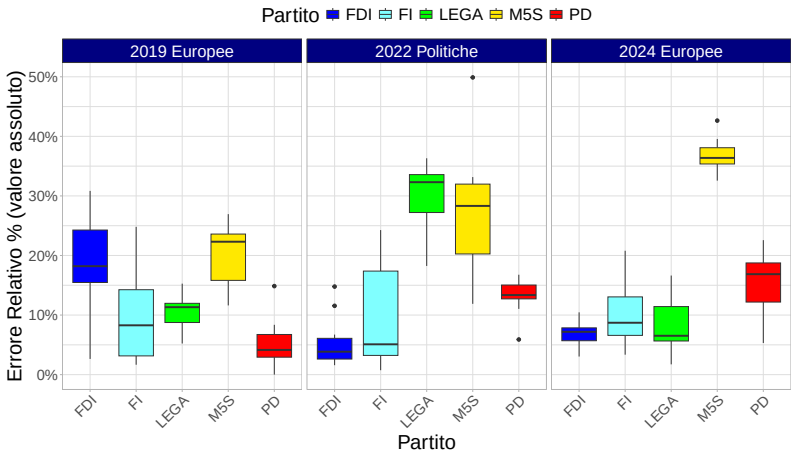


(b) Boxplot del valore assoluto della variabile risposta, definita come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni, per `Istituto`.

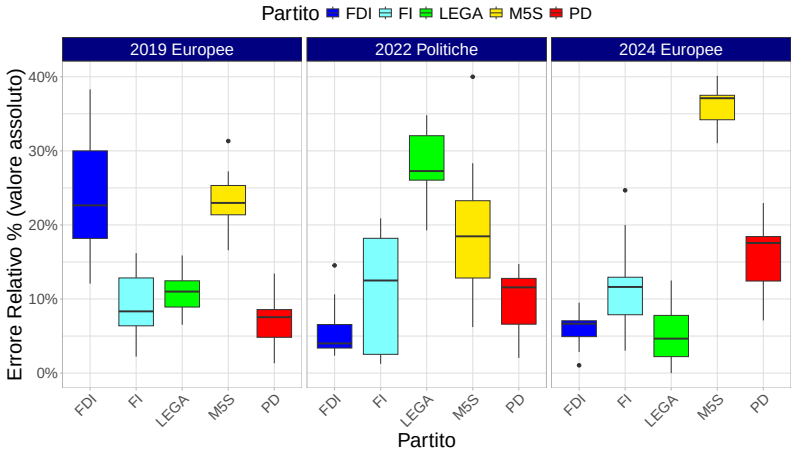
Figura 4.5: Boxplot del valore assoluto della variabile risposta condizionato alla variabile indipendente `Istituto`. In **(a)** viene definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo, mentre in **(b)** come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

facenti parte della variabile `Istituto`. Anche in queste circostanze, quindi, un'analisi grafica preliminare evidenzia segnali e un comportamento sistematico già discussi precedentemente in sezione 4.1. Sembrerebbe quindi che le agenzie demoscopiche non differiscano in modo sistematico in termini di affidabilità.

Nella figura 4.6 sono riportati i boxplot del valore assoluto della variabile risposta condizionata congiuntamente alle variabili `Partito` ed `Elezione`. In questo caso compaiono chiari segnali di sistematicità, indicando che tali fattori contribuiscono a



(a) Boxplot del valore assoluto della variabile risposta, definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo, per ogni combinazione di `Partito` - `Elezione`.



(b) Boxplot del valore assoluto della variabile risposta, definita come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni, per ogni combinazione di `Partito` - `Elezione`.

Figura 4.6: Boxplot del valore assoluto della variabile risposta per ogni combinazione di `Partito` - `Elezione`. In **(a)** viene definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo, mentre in **(b)** come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

spiegare la variabilità dell'errore. Non solo, i due grafici 4.6a e 4.6b, come nell'analisi della distorsione condotta nella sezione precedente, mostrano come la risposta cambi in modo diverso a seconda della combinazione tra `Partito` ed `Elezione`. Questo aspetto suggerisce l'implementazione di un modello che, per essere adeguato, deve includere anche il relativo termine d'interazione tra i due fattori. Dunque vengono stimati due modelli lineare classici; il primo modello contiene tutte le variabili indipendenti disponibili: `Istituto`, `Partito`, `Elezione` e l'interazione di queste

ultime due; mentre il secondo esclude `Istituto`. Quindi possono essere scritti come:

$$\begin{aligned}\text{abs}(\texttt{rel.err}_i) = \beta_0 + \beta_1\, \texttt{Istituto}_i + \beta_2\, \texttt{Partito}_i \\ + \beta_3\, \texttt{Elezione}_i + \beta_4\, (\texttt{Partito}_i \times \texttt{Elezione}_i)\end{aligned} \tag{4.6}$$

$$\begin{aligned}\text{abs}(\texttt{rel.err}_i) = \beta_0 + \beta_1\, \texttt{Partito}_i + \beta_2\, \texttt{Elezione}_i \\ + \beta_3\, (\texttt{Partito}_i \times \texttt{Elezione}_i)\end{aligned} \tag{4.7}$$

con $i = 1, \ldots, 165$. Si vuole testare l'ipotesi che la variabile `Istituto` non fornisca un contributo informativo rilevante capace di spiegare la variabile risposta, ovverosia che non vi siano differenze in termini di affidabilità da parte delle diverse agenzie demoscopiche selezionate nel capitolo 2. É possibile condurre questo tipo di test confrontando i due modelli definiti in (4.6) ed (4.7) attraverso l'ANOVA come fatto nella sezione precedente. Infatti l'ipotesi nulla del test costruito in tal maniera afferma che il modello più semplice, in questo caso (4.7), sia quello corretto, ossia che i coefficienti associati alle covariate aggiuntive, vale a dire `Istutito`, nel modello più grande (4.6) siano tutti nulli.

I risultati del confronto tra i due modelli considerati nel presente studio sono riportati nelle tabelle in 4.5.

	Res.Df	RSS	Df	Sum Sq	F	Pr(>F)
Modello in (4.7)	150	0.51542				
Modello in (4.6)	140	0.47052	10	0.044897	1.3359	0.2171

(a) Confronto ANOVA tra modelli annidati per la variabile risposta definita come valore assoluto dell'errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo.

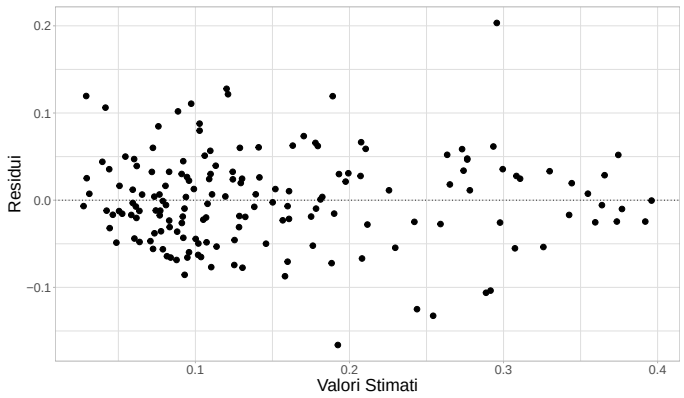
	Res.Df	RSS	Df	Sum Sq	F	Pr(>F)
Modello in (4.7)	150	0.42971				
Modello in (4.6)	140	0.37896	10	0.050752	1.8750	0.05346

(b) Confronto ANOVA tra modelli annidati per la variabile risposta definita come valore assoluto dell'errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

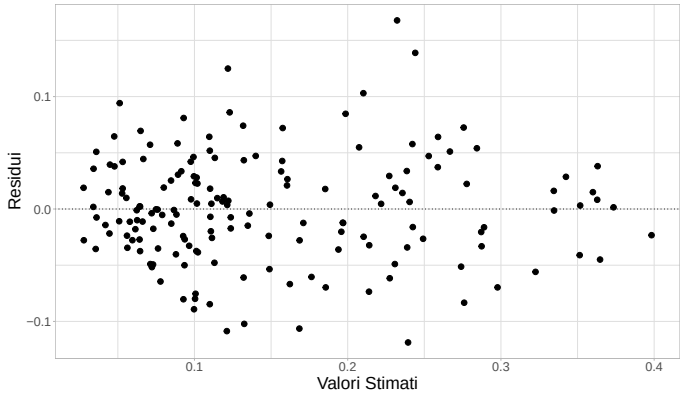
Tabella 4.5: Risultati del confronto tramite ANOVA per i due modelli lineari definiti rispettivamente come in (4.7) e in (4.6). In **(a)** la variabile risposta è definita come valore assoluto dell'errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo. In **(b)** la variabile risposta è definita come valore assoluto dell'errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

Dall'analisi del confronto emerge che l'ipotesi nulla non viene rifiutata in entrambi i casi essendo i p-value sopra la soglia critica. Questo risultato suggerisce che

l'inclusione della variabile `Istituto` non migliora in modo significativo la capacità del modello di spiegare la variabile risposta. Non solo, infatti verifica ciò che si era anticipato e cioè che non vi è eterogeneità in termini di affidabilità, e quindi di precisione, da parte delle varie agenzie demoscopiche considerate.



(a) Residui vs Valori stimati per il modello (4.6), con la variabile risposta definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo.



(b) Residui vs Valori stimati per il modello (4.6), con la variabile risposta definita come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

Figura 4.7: Grafico dei Residui vs Valori Stimati per il modello (4.6). In **(a)** la variabile risposta viene definita come errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo, mentre in **(b)** come errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

Tuttavia, anche in questo caso, dall'analisi grafica dei residui riportata in figura 4.7, con i due grafici 4.7a e 4.7b, emerge la possibile presenza di eteroschedasticità. Dunque, anche se le stime OLS continuano ad essere non distorte, gli usuali test

d'ipotesi possono risultare inappropriati. Per garantire robustezza all'analisi è necessario tenere conto di questa possibile violazione delle assunzioni.

Per queste ragioni il confronto tra i due modelli è stato ripetuto utilizzando una procedura robusta all'eteroschedasticità, specularmente a quanto fatto nella sezione precedente. Il test realizzato verifica l'ipotesi secondo cui l'inclusione della variabile `Istituto` non apporta un contributo informativo significativo rispetto al modello più parsimonioso che include esclusivamente `Partito`, `Elezione` e la loro interazione. La statistica test è quella descritta in (4.3) con la matrice di covarianza è stimata come in (4.4).

	Res.Df	Df	F	Pr(>F)
Modello in (4.7)	150			
Modello in (4.6)	140	10	1.0102	0.438

(a) Confronto tra modelli annidati tramite test di Wald con standard error robusti per la variabile risposta definita come valore assoluto dell'errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo.

	Res.Df	Df	F	Pr(>F)
Modello in (4.7)	150			
Modello in (4.6)	140	10	1.4847	0.151

(b) Confronto tra modelli annidati tramite test di Wald con standard error robusti per la variabile risposta definita come valore assoluto dell'errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

Tabella 4.6: Risultati del confronto tra modelli annidati mediante test di Wald con matrice di covarianza robusta. In **(a)** la variabile risposta è il valore assoluto dell'errore relativo tra interpolazione nel giorno delle elezioni e risultato effettivo. In **(b)** la variabile risposta è il valore assoluto dell'errore relativo tra ultimo sondaggio rilevato e risultato effettivo alle elezioni.

Il test di Wald robusto restituisce risultati, mostrati nelle tabelle 4.6a e 4.6b, analoghi a quelli precedentemente ottenuti con il confronto ANOVA. Si può concludere che l'ipotesi nulla non viene rifiutata: la covariata `Istituto` non è statisticamente significativa e non migliora in alcun modo la capacità esplicativa del modello.

Alla luce di queste evidenze, si può dunque argomentare che, una volta controllato per il contesto politico ed elettorale attraverso le variabili `Partito`, `Elezione` e la loro interazione, le diverse agenzie demoscopiche non mostrano differenze sistematiche in termini di affidabilità. In altre parole, l'errore commesso nelle stime delle intenzioni di voto non dipende dall'istituto che realizza il sondaggio, ma dal contesto specifico in cui la previsione viene formulata.

Dunque eventuali differenze osservate nella precisione delle previsioni elettorali non sono dovute ad una maggiore o minore competenza delle singole agenzie, ma piuttosto a fattori strutturali legati al partito e al turno elettorale considerato o pura casualità. Di conseguenza appare che le agenzie demoscopiche selezionate per quest'analisi risultano sostanzialmente intercambiabili in termini di affidabilità: a parità di informazione e di contesto, un'agenzia vale l'altra.

4.3 Discussione dei risultati e futuri sviluppi

In questa tesi, lo studio è stato condotto su binari paralleli e alla fine le analisi principali sono state due: la prima con l'obiettivo di comprendere se esistano distorsioni sistematiche nelle stime delle intenzioni di voto, per un determinato partito in una specifica elezione, basate su sondaggi pre-elettorali; la seconda con l'obiettivo di verificare la presenza di differenze strutturali nell'affidabilità delle principali agenzie demoscopiche in Italia.

Dal punto di vista metodologico, la variabile risposta è stata definita come errore relativo tra la stima delle intenzioni di voto e il risultato elettorale effettivo, la sua espressione è mostrata nell'equazione (2.1). Non solo, sono state considerate due diverse modalità di costruzione della previsione: nel primo caso è stato utilizzato l'ultimo sondaggio disponibile prima delle elezioni, mentre nel secondo si è fatto ricorso a una stima delle intenzioni di voto ottenuta tramite l'applicazione del filtro di Kalman, valutata nel giorno effettivo del turno elettorale. Per la seconda analisi, in cui l'interesse era focalizzato sulla precisione delle stime piuttosto che sulla direzione dell'errore, la variabile risposta è stata considerata in valore assoluto.

I risultati della prima analisi mostrano che la variabile Istituto non risulta statisticamente significativa nello spiegare l'errore relativo. Questo implica che l'errore non varia in modo sostanziale tra le diverse agenzie demoscopiche oltre quanto ci si aspetterebbe per semplice variabilità casuale. In altri termini, gli errori commessi dai vari istituti tendono ad assomigliarsi. Tuttavia, estendendo l'analisi alle combinazioni di `Partito` ed `Elezione`, emergono evidenze chiare di distorsioni sistematiche. Ciò suggerisce che i sondaggi pre-elettorali presentino problemi strutturali nella stima delle intenzioni di voto, con errori che dipendono dal contesto politico-elettorale e che risultano comuni a tutte le agenzie. In sostanza, i partiti vengono stimati con un certo errore che non è specifico del singolo istituto, ma condiviso.

La seconda analisi, in cui non si distingue più tra sovrastima e sottostima, conferma ulteriormente questo risultato. Una volta controllato per il contesto politico ed elettorale attraverso le variabili `Partito`, `Elezione` e la loro interazione, le diverse agenzie demoscopiche non mostrano differenze sistematiche in termini di affidabilità. L'errore commesso nelle stime delle intenzioni di voto non dipende quindi dall'istituto che realizza il sondaggio, ma piuttosto dal contesto specifico in cui la previsione viene formulata. Eventuali differenze osservate nella precisione delle stime non sono riconducibili a una maggiore o minore competenza delle singole agenzie, bensì a fattori strutturali legati al partito, al turno elettorale considerato o alla componente casuale. Di conseguenza, gli istituti demoscopici analizzati appaiono sostanzialmente intercambiabili in termini di affidabilità: a parità di informazione e di contesto, un'agenzia vale l'altra.

Questi risultati consentono di formulare una critica all'approccio comunemente noto come *Poll of Polls*. Un sondaggio elettorale rappresenta un'osservazione singola, ma si inserisce in un contesto più ampio caratterizzato dalla presenza di numerose agenzie demoscopiche che pubblicano indagini con frequenze eterogenee. In prossimità di un'elezione, è quindi possibile disporre di più sondaggi riferiti allo stesso periodo temporale. Diverse metodologie sfruttano questa abbondanza informativa, tra cui il *Poll of Polls*, che per la sua semplicità ed immediatezza viene spesso adottata dai media per comunicare quello che è l'approccio statistico al contesto politico. Alla base del metodo vi è l'ipotesi che la combinazione di risultati provenienti da più istituti consenta di ottenere una stima più precisa delle intenzioni di voto, diluendo eventuali distorsioni metodologiche specifiche di un singolo sondaggio o di una singola agenzia.

Tuttavia, come evidenziato da Shirani-Mehr et al. (2018), questo approccio richiede cautela. Hanno infatti dimostrato che i sondaggi relativi a una stessa elezione tendono a condividere una distorsione sistematica comune, probabilmente dovuta al fatto che, pur provenendo da istituti diversi, quest'ultimi affrontano problemi simili e si basano su procedure metodologiche affini. L'aggregazione dei sondaggi non elimina questa componente di distorsione condivisa, che rimane inalterata anche quando si media su un gran numero di osservazioni. Ne consegue che un modello aggregativo è, in ultima analisi, della stessa qualità dei sondaggi di cui è composto: se tutte le agenzie sono sistematicamente in errore nella stessa direzione, il processo di aggregazione rischia di produrre risultati fuorvianti.

I risultati ottenuti in questa tesi, sebbene riferiti al contesto italiano, sono coerenti con le evidenze riportate da Shirani-Mehr et al. (2018). Le distorsioni sistematiche

osservate dipendono dal contesto politico-elettorale e sono comuni a tutte le agenzie, suggerendo che il *Poll of Polls* non rappresenti una soluzione ai limiti strutturali dei sondaggi pre-elettorali. In questo senso, approcci alternativi che integrino non solo informazioni provenienti da più sondaggi, ma anche variabili di contesto (come indicatori economici o risultati di elezioni precedenti), appaiono più promettenti.

Per quanto riguarda i futuri sviluppi, un aspetto che meriterebbe un'analisi più approfondita è il fenomeno dell'*herding* tra istituti demoscopici nel contesto italiano. È noto in letteratura che, specialmente nelle fasi finali di una campagna elettorale, le agenzie tendano a pubblicare risultati sempre più simili tra loro. Questo comportamento riduce i benefici associati alla cosiddetta "saggezza della folla", che presuppone l'indipendenza delle fonti informative. Un'agenzia che si limita a replicare o mediare risultati precedenti non fornisce informazione indipendente, aggravando ulteriormente i limiti del *Poll of Polls* tra le altre cose. Come suggerito da Whiteley (2016) e da Silver (2014), un modo per individuare questo fenomeno potrebbe essere l'analisi della variabilità dei sondaggi all'inizio e alla fine della campagna elettorale: una riduzione significativa dell'eterogeneità tra le stime man mano che le elezioni si avvicinano potrebbe costituire un indizio di *herding*.

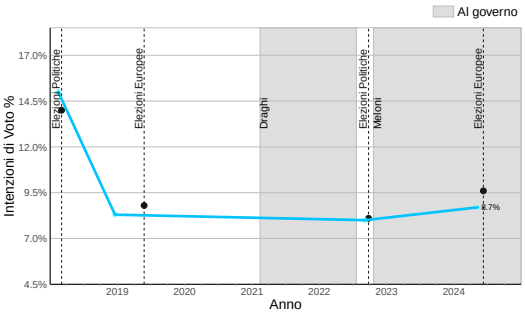
Un ulteriore ambito di sviluppo riguarda il modello *state space* adottato in questa tesi. L'analisi è stata condotta utilizzando un approccio univariato, trattando ciascun partito separatamente. Questa scelta rappresenta un'approssimazione, in quanto i partiti costituiscono parti di un totale e le loro dinamiche sono intrinsecamente collegate. Un'estensione naturale sarebbe quindi l'adozione di modelli *state space* multivariati, in grado di catturare le interdipendenze tra le serie. Inoltre, potrebbero essere esplorate metodologie alternative, come il *particle filtering*, che consentono una maggiore flessibilità nella modellazione di sistemi complessi e non lineari.

Appendice A

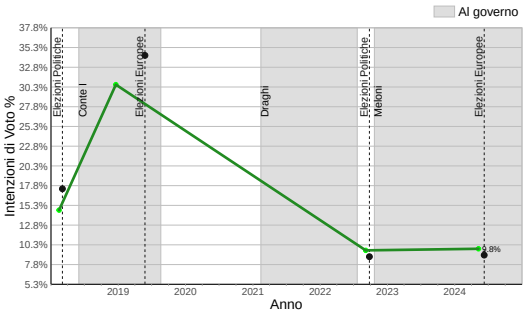
Integrazione grafici capitolo 2

A.1 Serie storica intenzioni di voto stimate

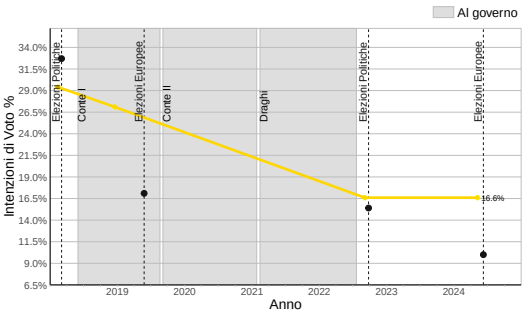
Viene riportato il grafico della serie storica del consenso sull'intero arco temporale per ogni combinazione di agenzia-partito. La parte di sfondo colorato in grigio indica quando il partito è stato al governo.



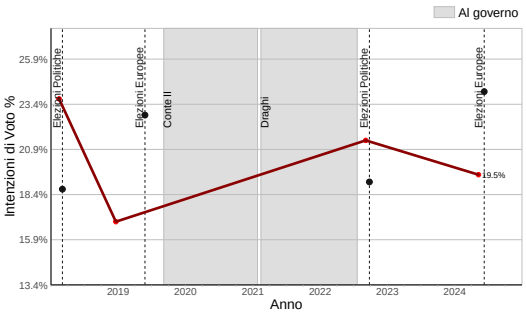
Cise - FI



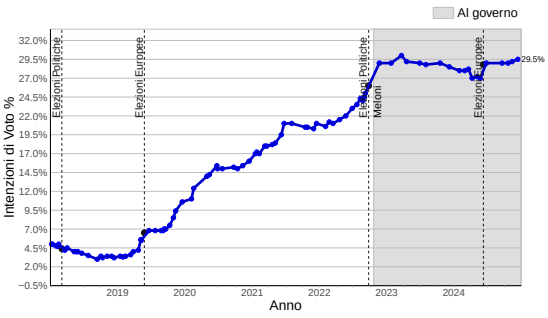
Cise - Lega



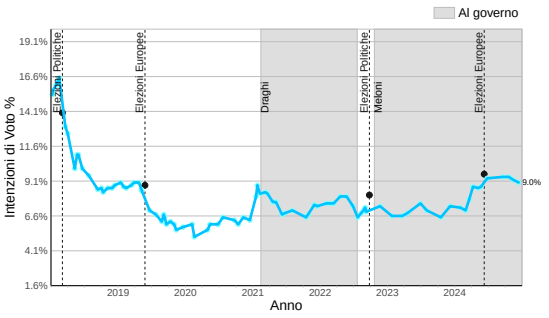
Cise - M5S



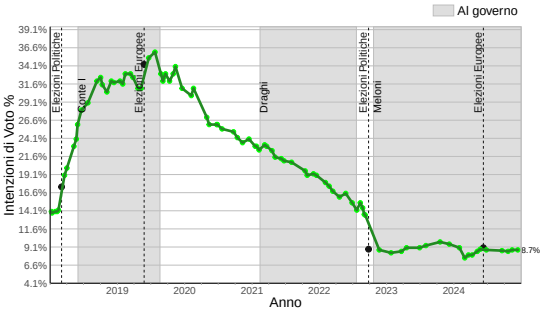
Cise - PD



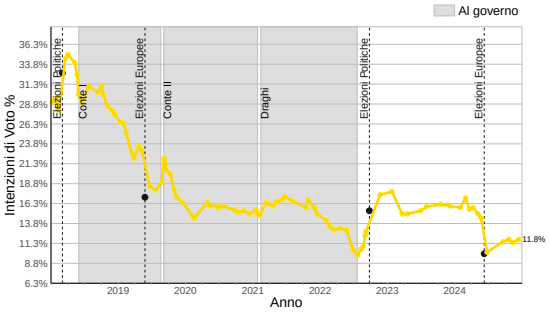
Demopolis - FDI



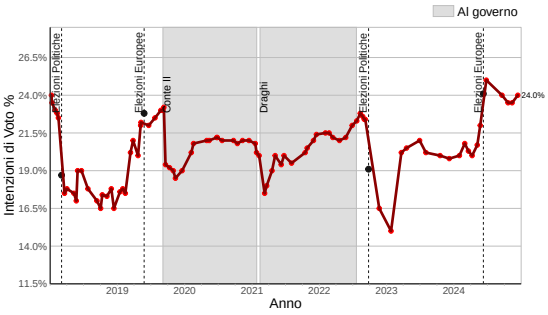
Demopolis - FI



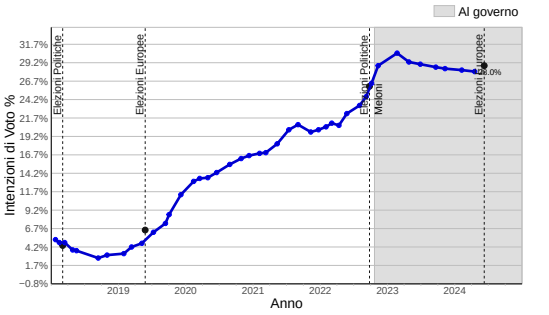
Demopolis - Lega



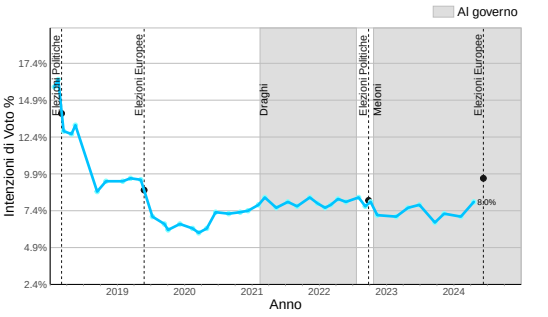
Demopolis - M5S



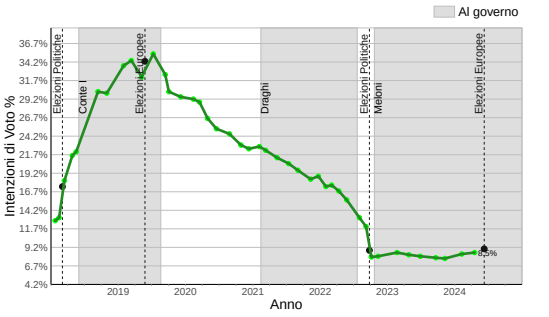
Demopolis - PD



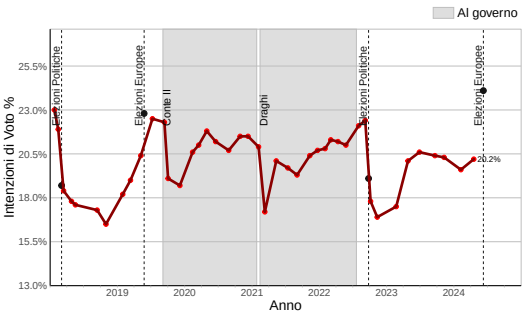
Demos - FDI



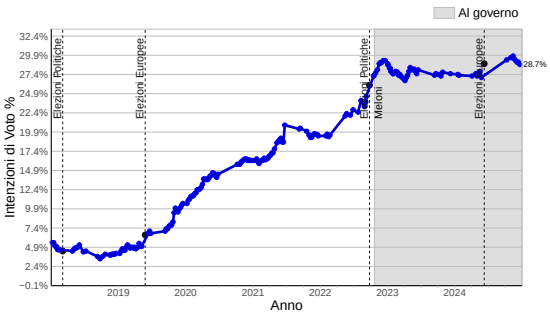
Demos - FI



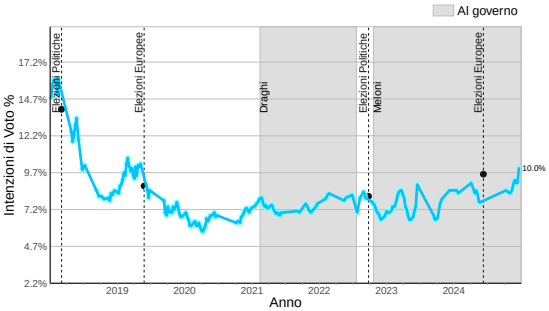
Demos - Lega



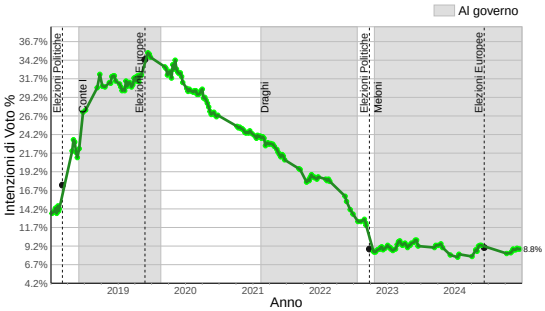
Demos - PD



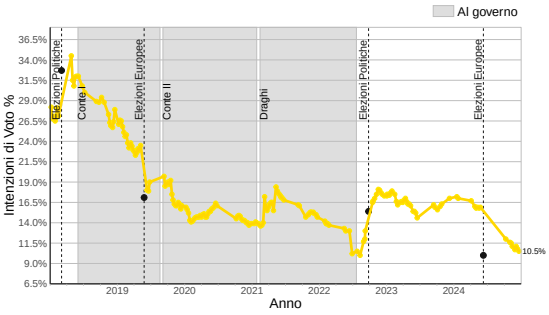
Emg - FDI



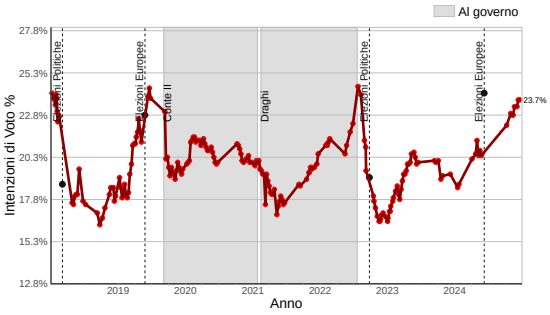
Emg - FI



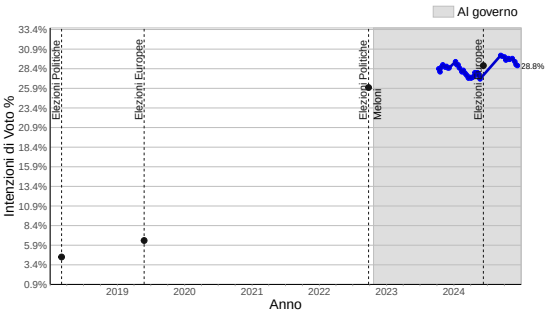
Emg - Lega



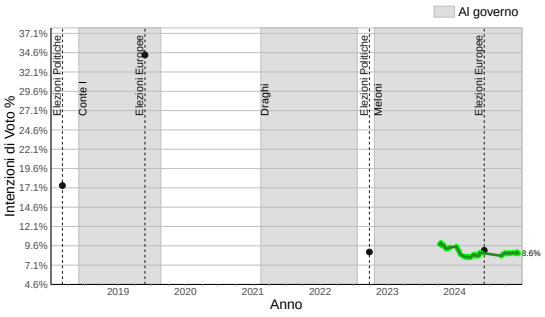
Emg - M5S



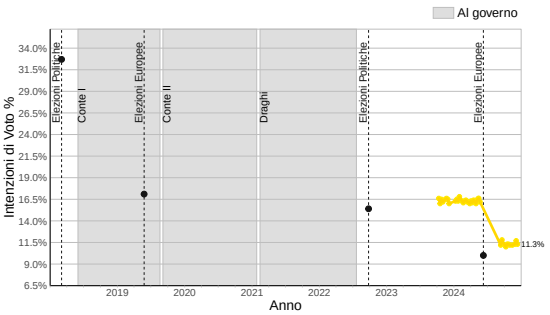
Emg - PD



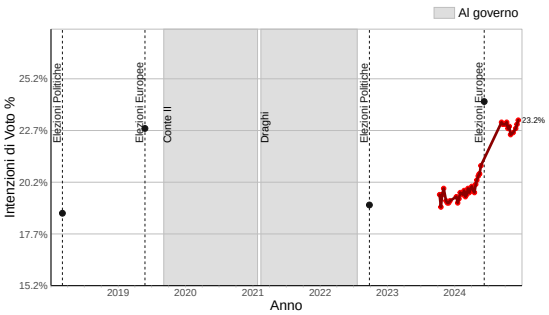
Eumetra - FDI



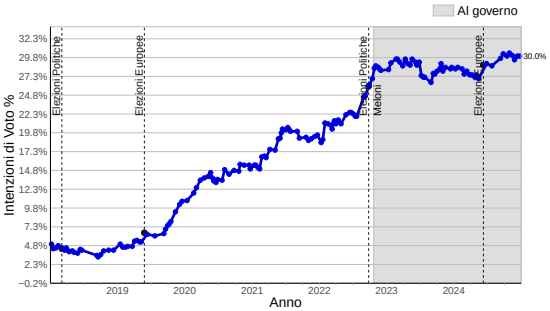
Eumetra - Lega



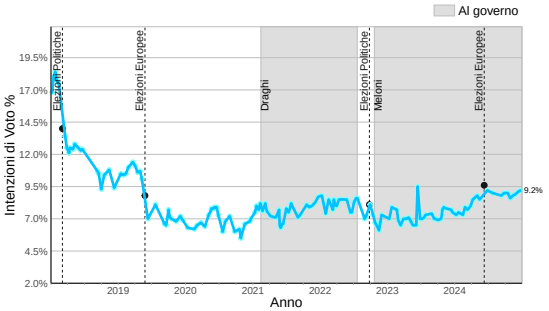
Eumetra - M5S



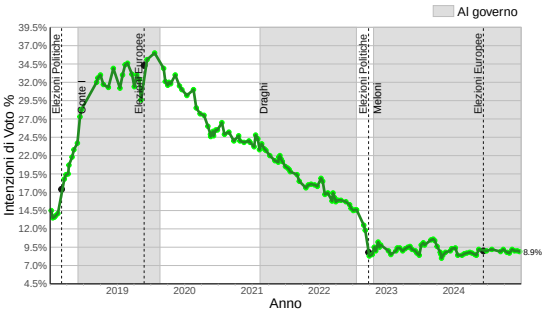
Eumetra - PD



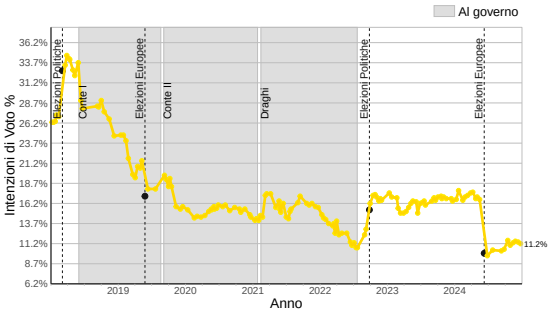
Euromedia - FDI



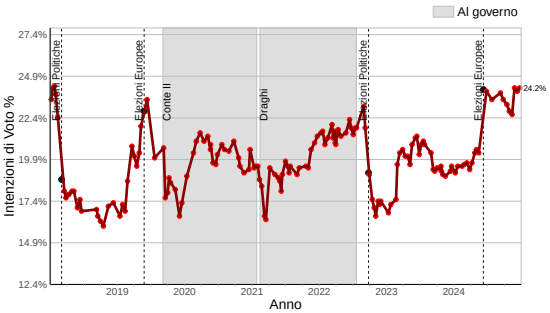
Euromedia - FI



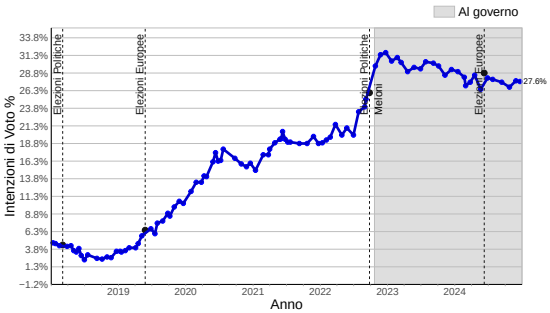
Euromedia - Lega



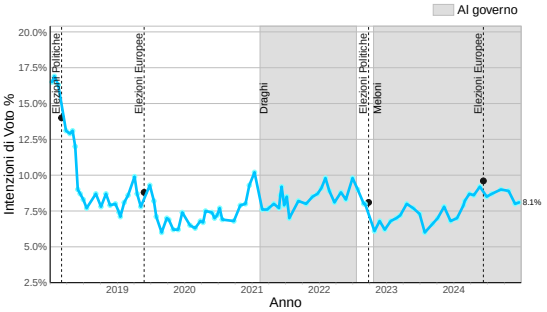
Euromedia M5S



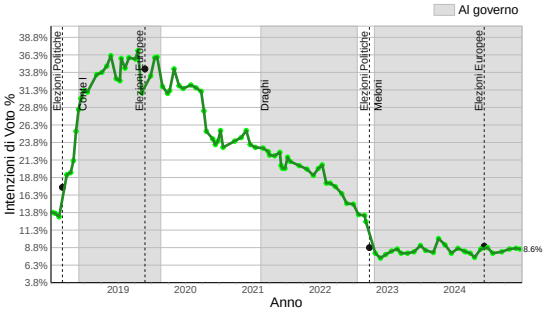
Euromedia - PD



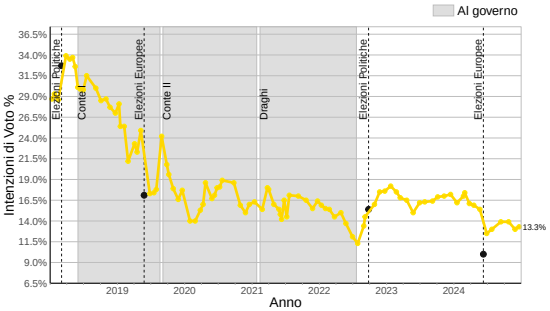
Ipsos - FDI



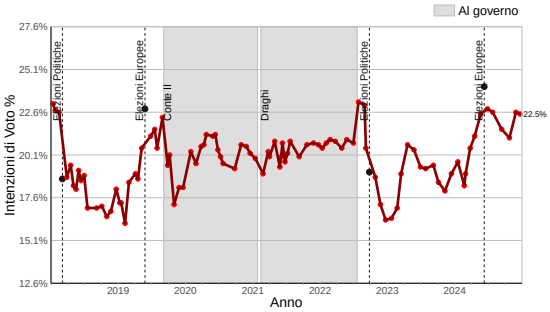
Ipsos - FI



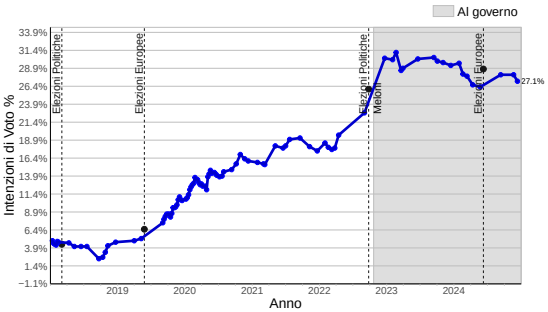
Ipsos - Lega



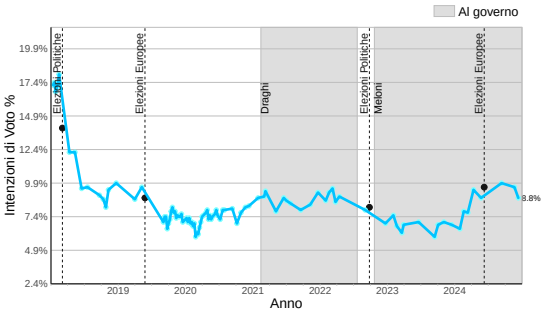
Ipsos - M5S



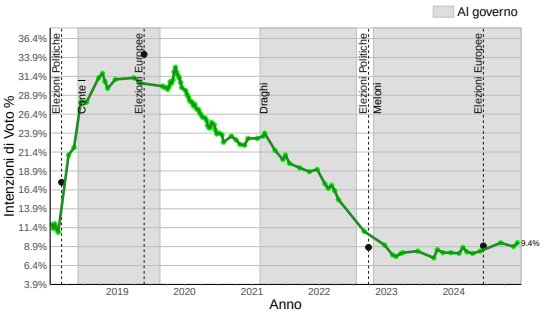
Ipsos - PD



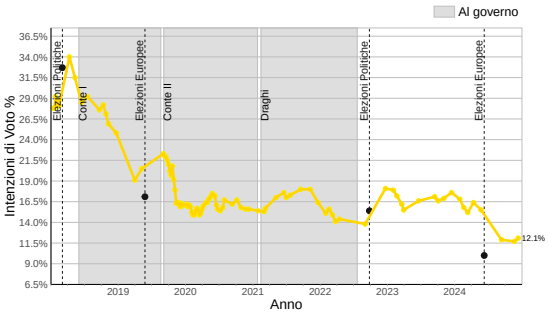
Ixè - FDI



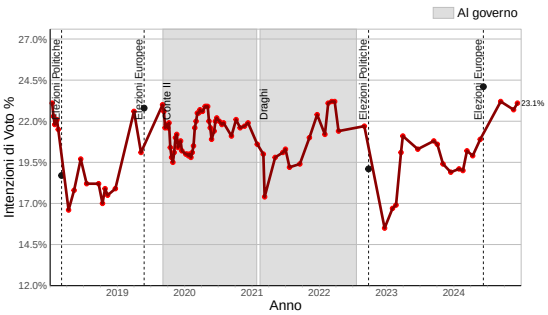
Ixe - FI



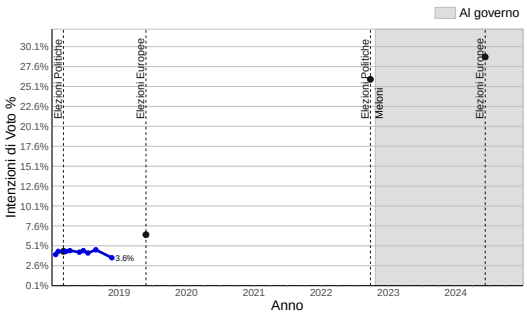
Ixè - Lega



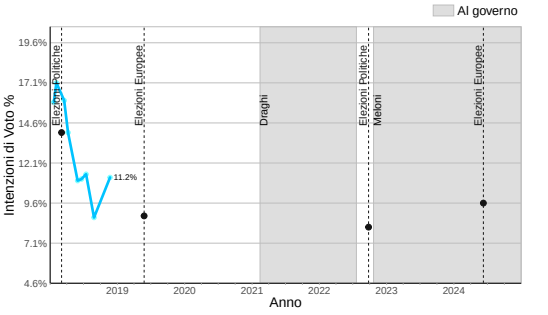
Ixè - M5S



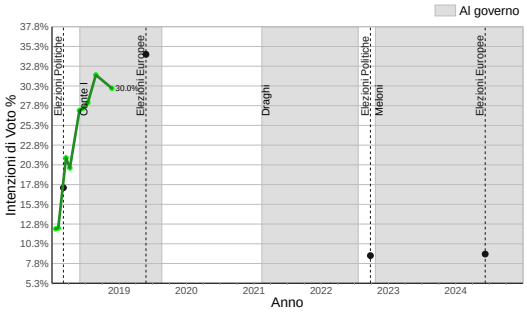
Ixè - PD



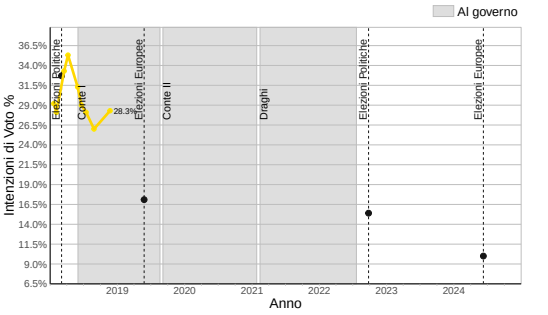
Lorien - FDI



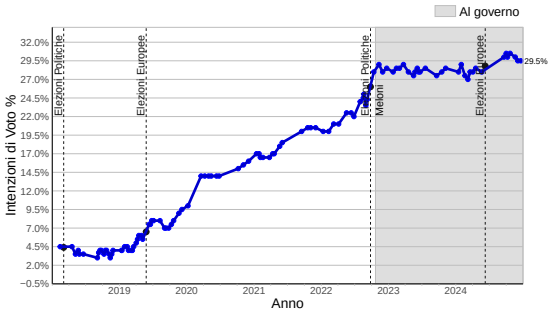
Lorien - FI



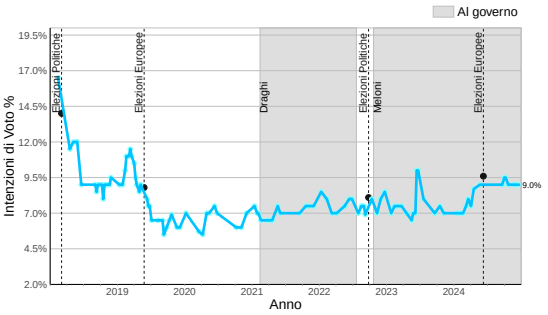
Lorien - Lega



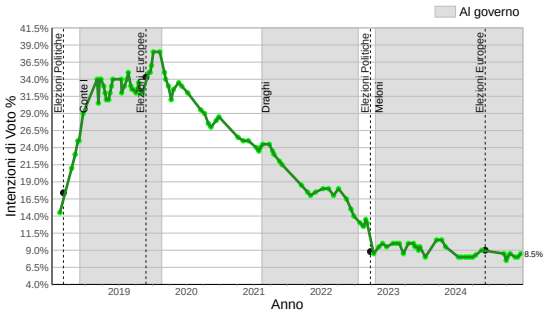
Lorien - M5S



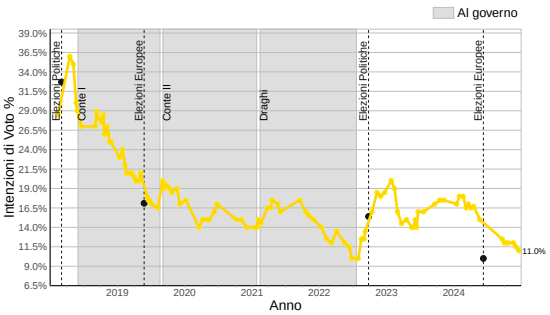
Noto - FDI



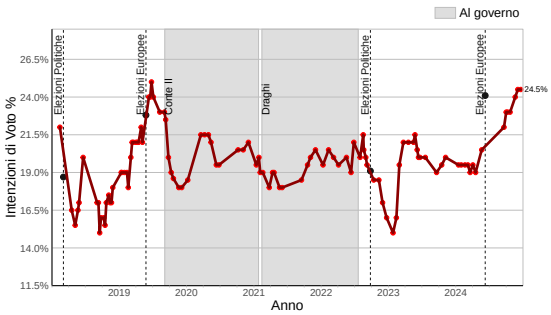
Noto - FI



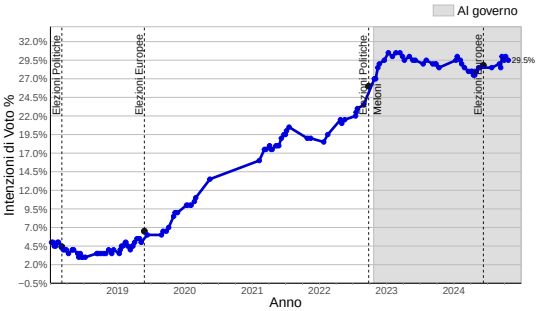
Noto - Lega



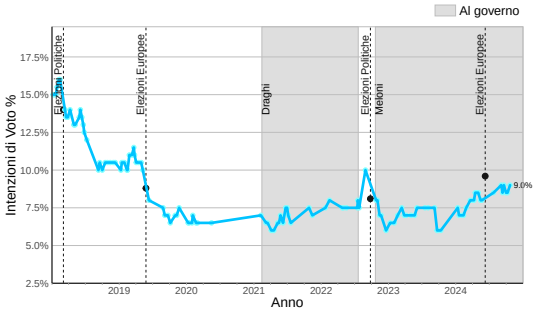
Noto - M5S



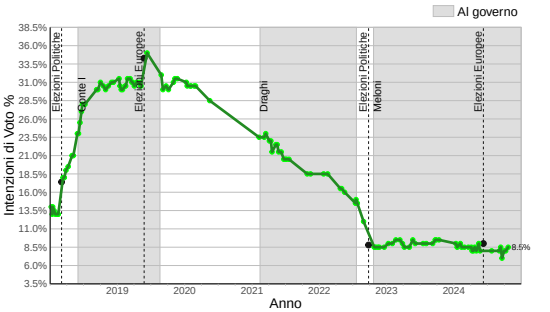
Noto - PD



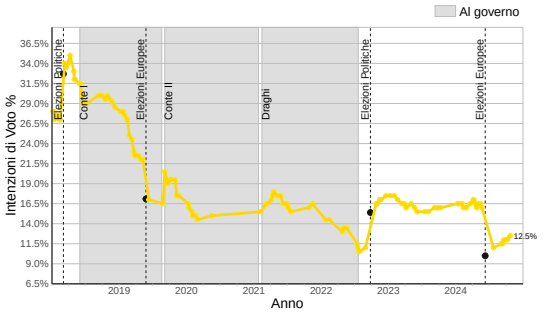
Piepoli - FDI



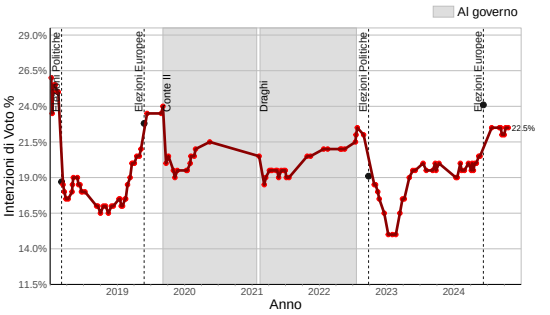
Piepoli - FI



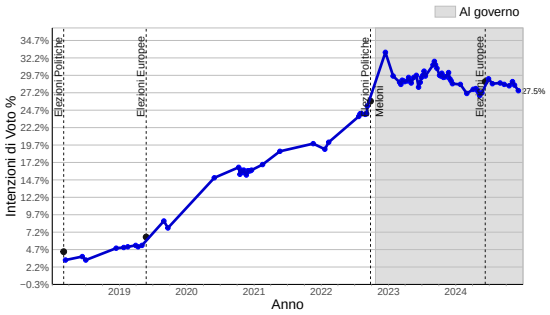
Piepoli - Lega



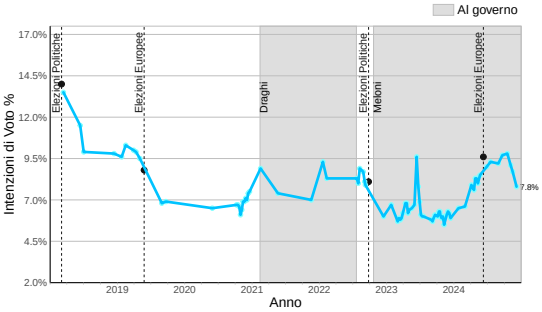
Piepoli - M5S



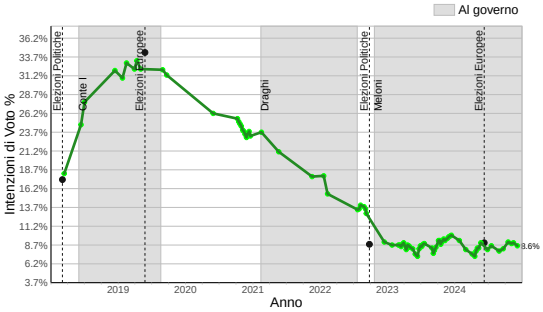
Piepoli - PD



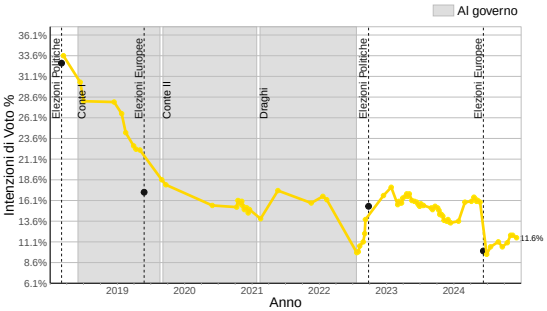
Quorum - FDI



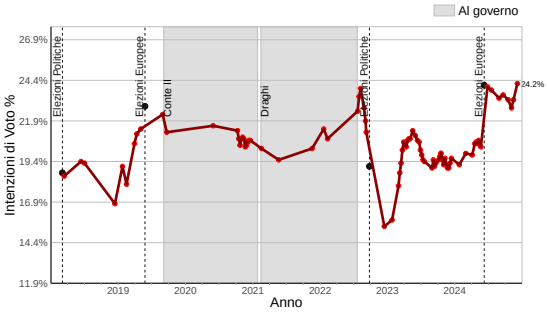
Quorum - FI



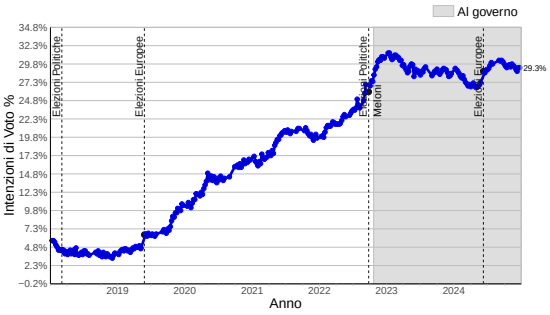
Quorum - Lega



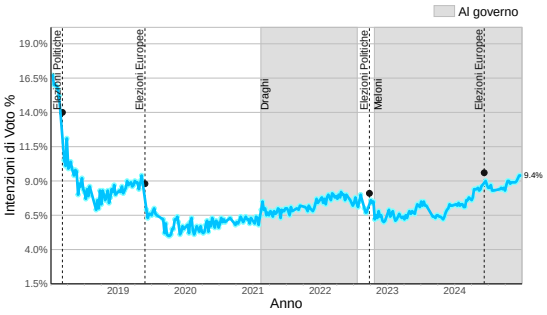
Quorum - M5S



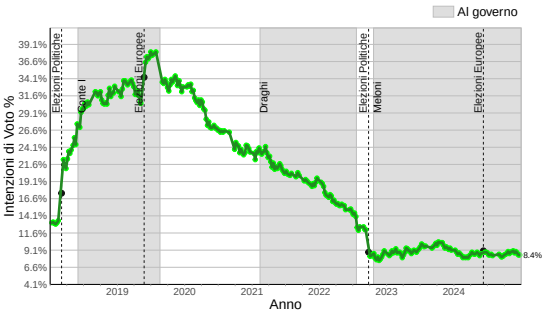
Quorum - PD



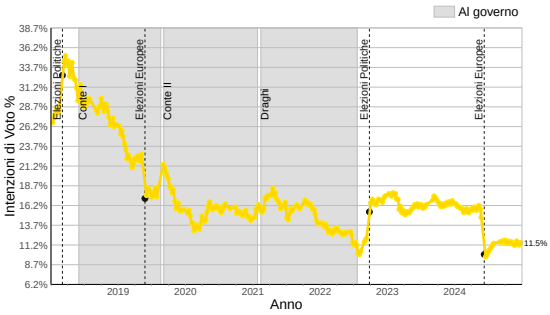
Swg - FDI



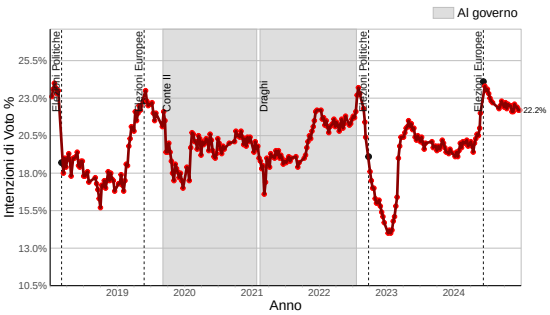
Swg - FI



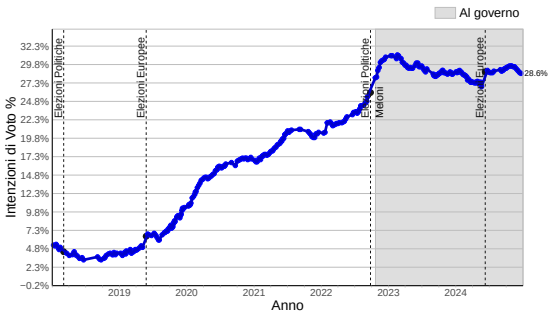
Swg - Lega



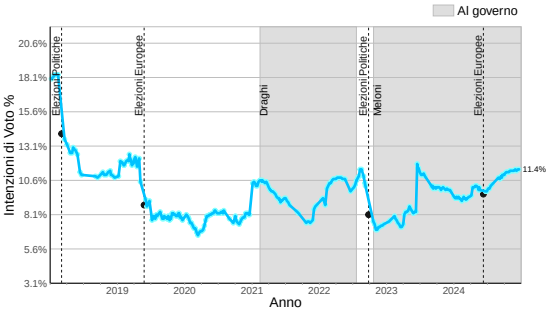
Swg - M5S



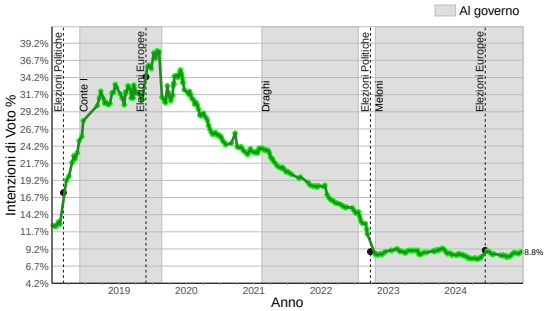
Swg - PD



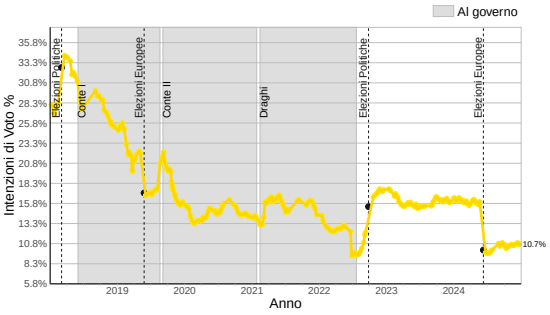
Tecnè - FDI



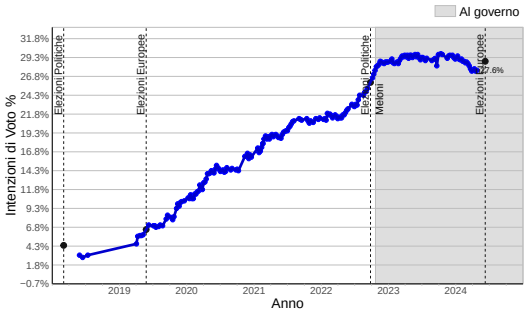
Tecnè - FI



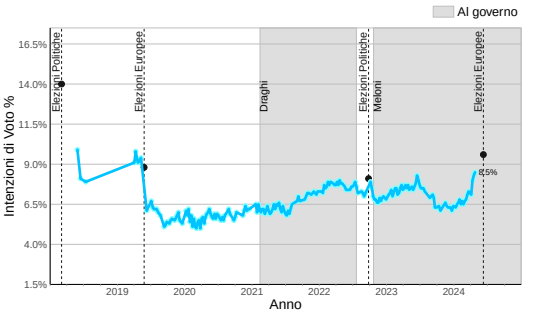
Tecnè - Lega



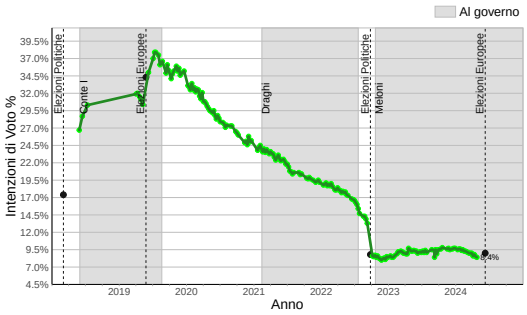
Tecnè - M5S



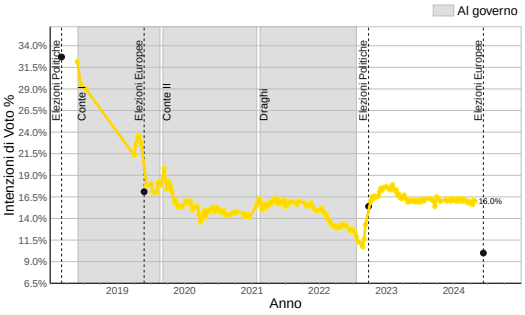
TP - FDI



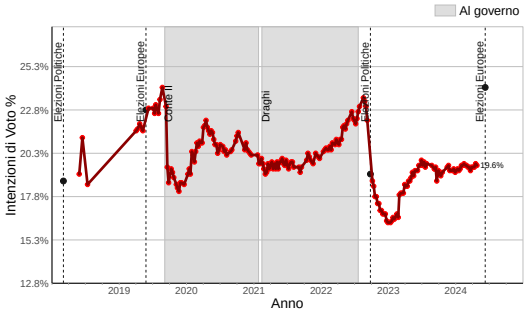
TP - FI



TP - Lega



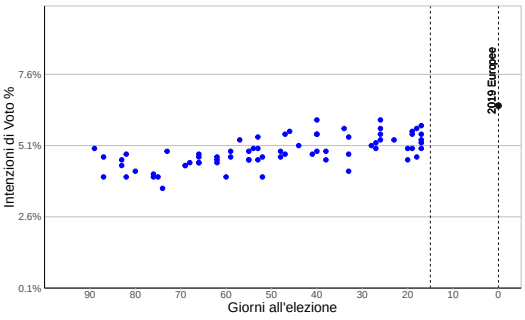
TP - M5S



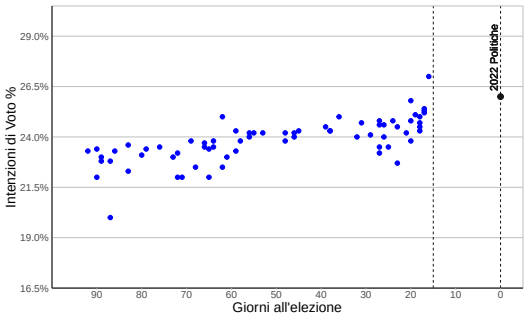
TP - PD

A.2 Intenzioni di voto stimate nei tre mesi precedenti all'elezione

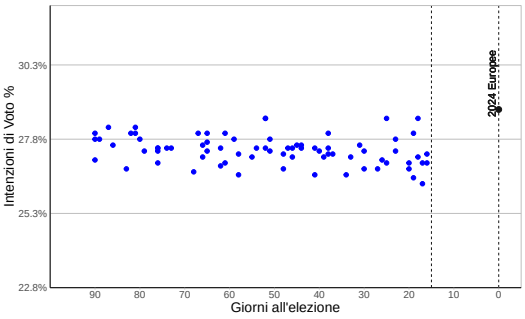
Viene riportato il grafico delle intenzioni di voto stimate da tutte le agenzie selezionate nei tre mesi precedenti ad un elezione per ogni combinazione di partito-elezione. La prima linea verticale tratteggiata indica la soglia entro cui poi non è più possibile pubblicare indagini politiche (15 giorni prima del voto), mentre la seconda indica il giorno dell'elezione.



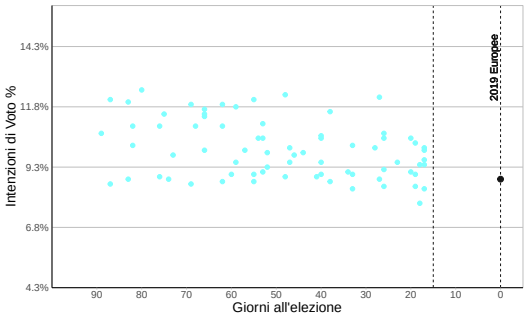
FDI - Europee 2019



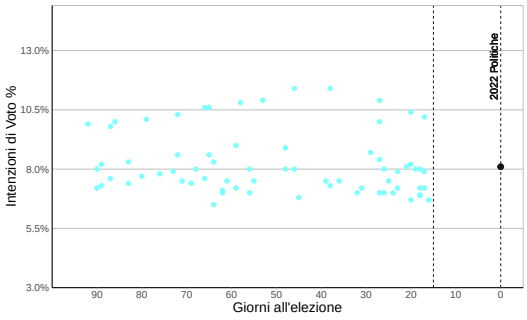
FDI - Politiche 2022



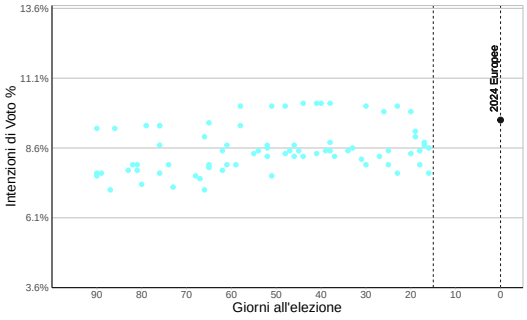
FDI - Europee 2024



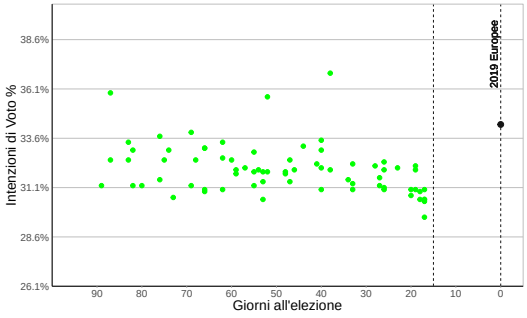
FI - Europee 2019



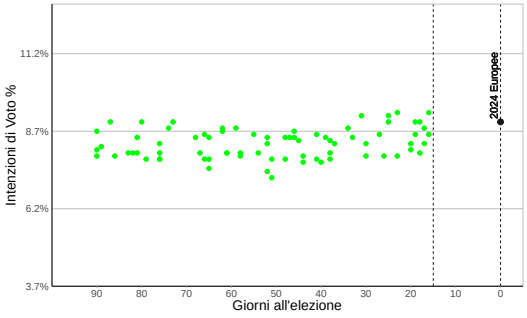
FI - Politiche 2022



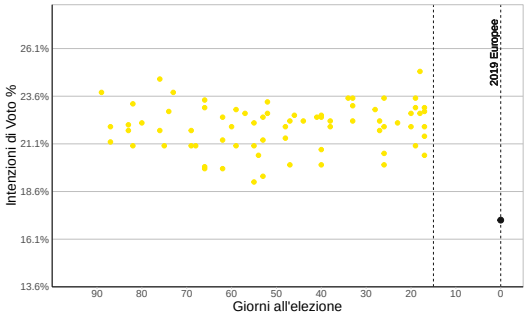
FI - Europee 2024



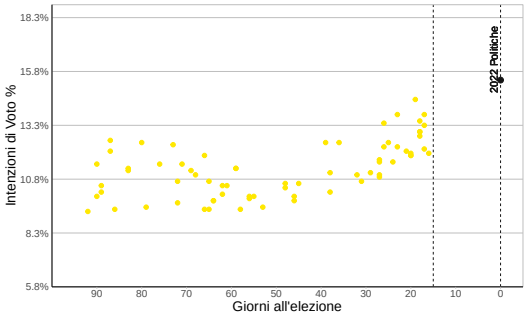
Lega - Europee 2019



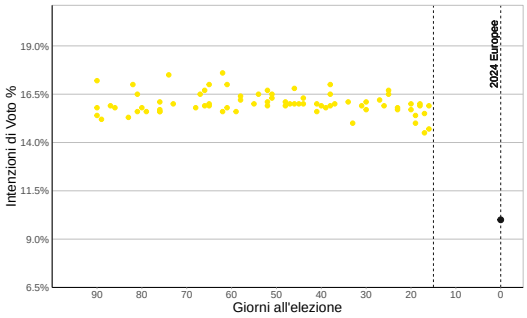
Lega - Europee 2024



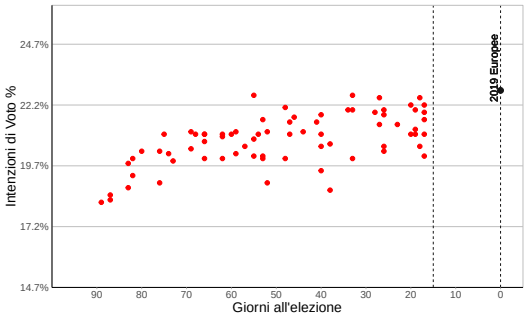
M5S - Europee 2019



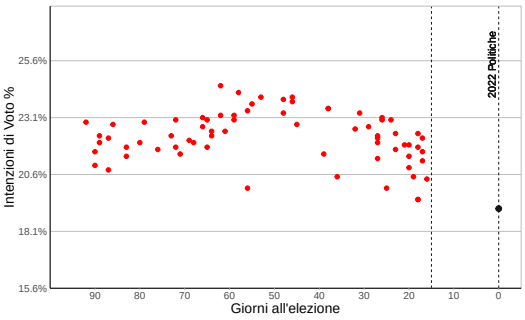
M5S - Politiche 2022



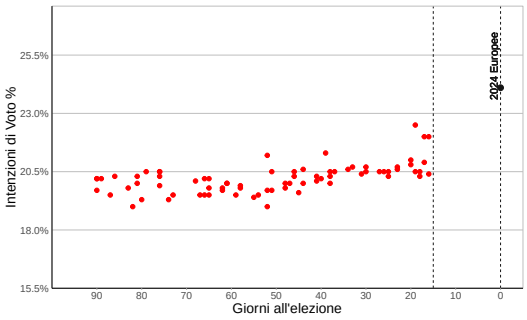
M5S - Europee 2024



PD - Europee 2019



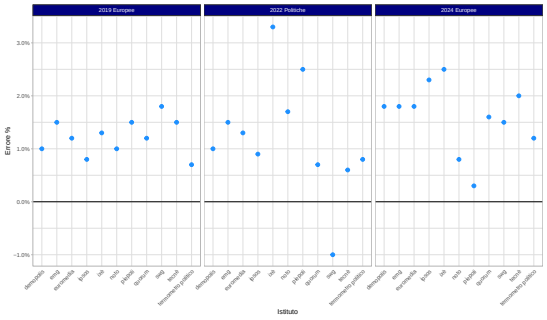
PD - Politiche 2022



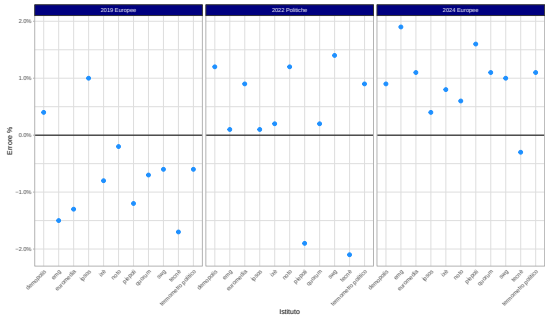
PD - Europee 2024

A.3 Differenza per intenzioni di voto

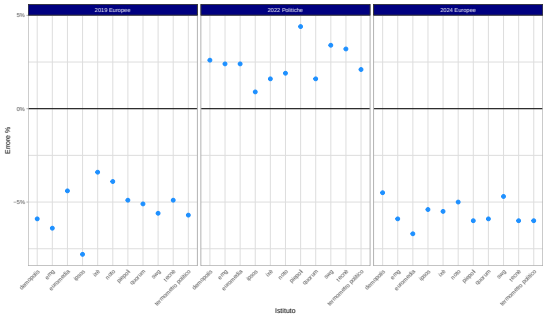
Viene riportato il grafico della differenza tra le intenzioni di voto stimate dall'ultimo sondaggio disponibile per agenzia rispetto al risultato delle elezioni corrispondenti per ogni partito.



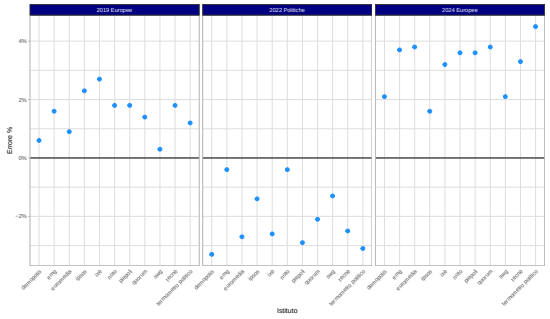
FDI



FI



M5S



PD

Appendice B

Dimostrazioni risultati principali capitolo 3

B.1 Dimostrazione lemma 1

Dimostrazione. Sia

$$z = x - \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y), \qquad A = \Sigma_{xy}\Sigma_{yy}^{-1}.$$

Poiché la trasformazione da (x, y) a (y, z) è lineare e (x, y) è distribuita Normalmente, allora la distribuzione congiunta di y e z è Normale. Si ha che

$$\begin{aligned}\mathbb{E}\big[z\big] &= \mathbb{E}\big[x - A(y - \mu_y)\big] \\ &= \mathbb{E}\big[x\big] - A\,\mathbb{E}\big[y - \mu_y\big] \\ &= \mu_x.\end{aligned}$$

Per la varianza,

$$\begin{aligned}\mathrm{Var}(z) &= \mathrm{Var}\big(x - A(y - \mu_y)\big) \\ &= \mathrm{Var}(x) + \mathrm{Var}\big(A(y - \mu_y)\big) - 2\,\mathrm{Cov}\big(x, A(y - \mu_y)\big) \\ &= \Sigma_{xx} + A\Sigma_{yy}A^\top - 2\Sigma_{xy}A^\top \\ &= \Sigma_{xx} + \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yy}\Sigma_{yy}^{-1}\Sigma_{yx} - 2\Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx} \\ &= \Sigma_{xx} - \Sigma_{xy}\Sigma_{yy}^{-1}\Sigma_{yx}. \qquad \text{(B.1)}\end{aligned}$$

Infine, per la covarianza tra y e z,

$$\begin{aligned}\mathrm{Cov}(y, z) &= \mathrm{Cov}\big(y, x - A(y - \mu_y)\big) \\ &= \mathrm{Cov}(y, x) - \mathrm{Cov}\big(y, A(y - \mu_y)\big)\end{aligned}$$

$$
\begin{aligned}
&= \Sigma_{yx} - \mathrm{Cov}(y, Ay) \\
&= \Sigma_{yx} - \Sigma_{yy} A^\top \\
&= \Sigma_{yx} - \Sigma_{yy} (\Sigma_{xy} \Sigma_{yy}^{-1})^\top \\
&= \Sigma_{yx} - \Sigma_{yy} \Sigma_{yy}^{-1} \Sigma_{yx} \\
&= \Sigma_{yx} - \Sigma_{yx} = 0.
\end{aligned} \tag{B.2}
$$

Due vettori casuali distribuiti Normalmente incorrelati sono anche indipendenti. Pertanto dall'equazione (B.2) si conclude che z sia indipendente da y. Dal momento che la distribuzione di z non dipende da y, la sua distribuzione condizionata ad y coincide con quella non condizionata, che è una Normale con vettore delle medie μ_x e matrice di varianza covarianza pari a quella in equazione (B.1) che è identica a quella mostrata in (3.5). In conclusione poiché $z = x - \Sigma_{xy}\Sigma_{yy}^{-1}(y - \mu_y)$ segue che la distribuzione condizionata di x dato y è una Normale con vettore delle medie (3.4) e matrice di varianza covarianza (3.5).

□

B.2 Calcoli estesi derivazione filtro

In questa sezione si riportano i calcoli estesi per alcuni passaggi coinvolti nella derivazione del filtro di Kalman e nella gestione della previsione con il filtro.

$$
\begin{aligned}
\mathrm{Cov}(\alpha_t, v_t) &= \mathbb{E}(\alpha_t v_t^\top \mid Y_{t-1}) - \mathbb{E}(\alpha_t \mid Y_{t-1}) E(v_t \mid Y_{t-1}) \\
&= \mathbb{E}(\alpha_t v_t^\top \mid Y_{t-1}) \\
&= \mathbb{E}(\alpha_t (Z_t \alpha_t + \epsilon_t - Z_t a_t)^\top \mid Y_{t-1}) \\
&= \mathbb{E}(\alpha_t (\alpha_t - a_t)^\top Z_t^\top \mid Y_{t-1}) \\
&= \left[\mathbb{E}(\alpha_t \alpha_t^\top \mid Y_{t-1}) - \mathbb{E}(\alpha_t \mid Y_{t-1}) a_t^\top\right] Z_t^\top \\
&= \left[\mathbb{E}(\alpha_t \alpha_t^\top \mid Y_{t-1}) - a_t a_t^\top\right] Z_t^\top \\
&= P_t Z_t^\top
\end{aligned}
$$

$$
\begin{aligned}
P_{t+1} &= T_t P_{t|t} T_t^\top + R_t Q_t R_t^\top \\
&= T_t \left[P_t - P_t Z_t^\top F_t^{-1} Z_t P_t\right] T_t^\top + R_t Q_t R_t^\top \\
&= T_t P_t T_t^\top - T_t P_t Z_t^\top F_t^{-1} Z_t P_t T_t^\top + R_t Q_t R_t^\top
\end{aligned}
$$

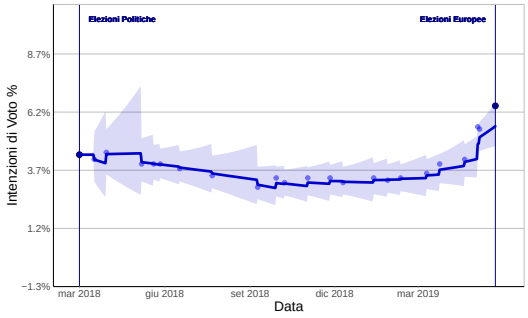
$$
\begin{aligned}
&= T_t P_t \left(T_t - T_t P_t Z_t^\top F_t^{-1} Z_t\right)^\top + R_t Q_t R_t^\top \\
&= T_t P_t \left(T_t - K_t Z_t\right)^\top + R_t Q_t R_t^\top
\end{aligned}
$$

$$
\begin{aligned}
\bar{P}_{n+j+1} &= \mathbb{E}\big[(\bar{a}_{n+j+1} - \alpha_{n+j+1})(\bar{a}_{n+j+1} - \alpha_{n+j+1})^\top \mid Y_n\big] \\
&= \mathbb{E}\Big[(T_{n+j}\bar{a}_{n+j} - T_{n+j}\alpha_{n+j} - R_{n+j}\eta_{n+j}) \\
&\quad (T_{n+j}\bar{a}_{n+j} - T_{n+j}\alpha_{n+j} - R_{n+j}\eta_{n+j})^\top \;\Big|\; Y_n\Big] \\
&= T_{n+j}\mathbb{E}\big[(\bar{a}_{n+j} - \alpha_{n+j})(\bar{a}_{n+j} - \alpha_{n+j})^\top \mid Y_n\big] T_{n+j}^\top + \\
&\quad R_{n+j}\mathbb{E}\big[\eta_{n+j}\eta_{n+j}^\top\big] R_{n+j}^\top \\
&\quad - 2\mathrm{Cov}\big(T_{n+j}(\bar{a}_{n+j} - \alpha_{n+j}), R_{n+j}\eta_{n+j}\big) \\
&= T_{n+j}\mathbb{E}\big[(\bar{a}_{n+j} - \alpha_{n+j})(\bar{a}_{n+j} - \alpha_{n+j})^\top \mid Y_n\big] T_{n+j}^\top + \\
&\quad R_{n+j}\mathbb{E}\big[\eta_{n+j}\eta_{n+j}^\top\big] R_{n+j}^\top \\
&= T_{n+j}\bar{P}_{n+j} T_{n+j}^\top + R_{n+j} Q_{n+j} R_{n+j}^\top
\end{aligned}
$$

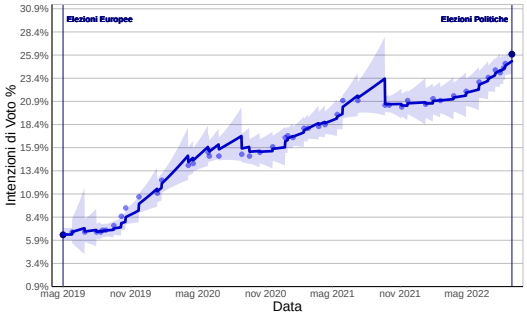
Appendice C

Integrazione grafici capitolo 3

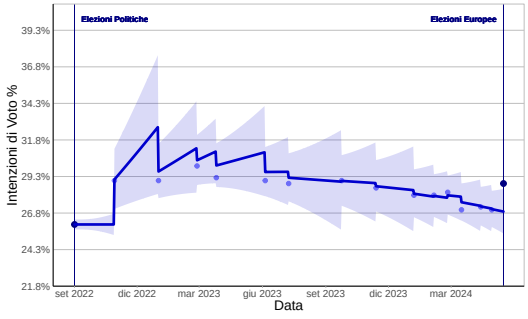
Vengono riportati tutti i grafici ottenuti attraverso il processo descritto nel capitolo 3. In particolare in ogni grafico la linea corrisponde alla stima filtrata giornaliera del livello delle intenzioni di voto per il partito ottenuta tramite l'applicazione del filtro di Kalman ai sondaggi pubblicati dall'agenzia nell'intervallo d'interesse. I punti, in questo senso, corrispondono proprio alle rilevazioni dei sondaggi pre-elettorali. Vengono riportati anche gli intervalli di confidenza associati ad ogni stima.



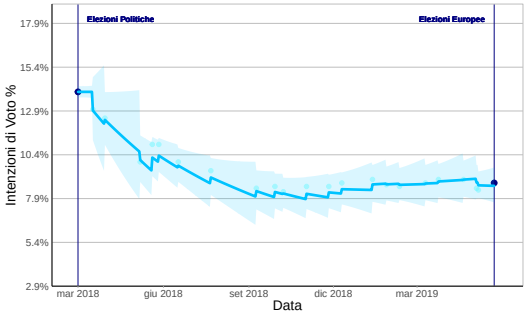
Demopolis - FDI - Intervallo 1



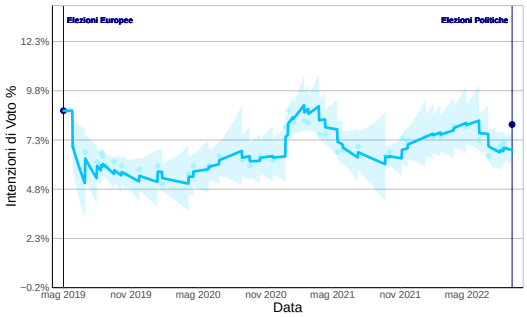
Demopolis - FDI - Intervallo 2



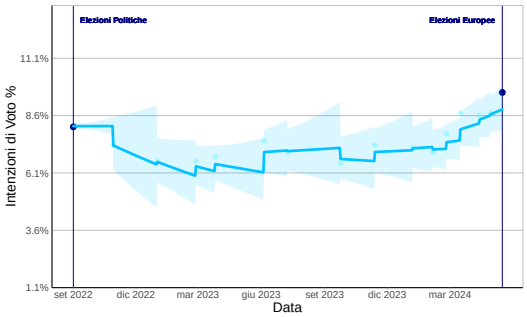
Demopolis - FDI - Intervallo 3



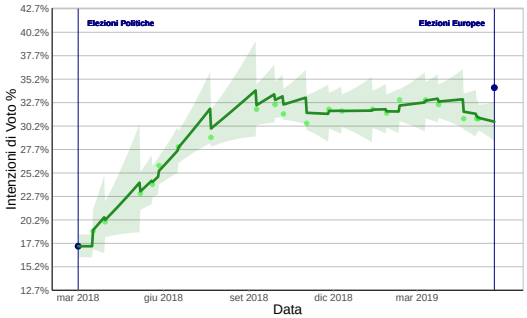
Demopolis - FI - Intervallo 1



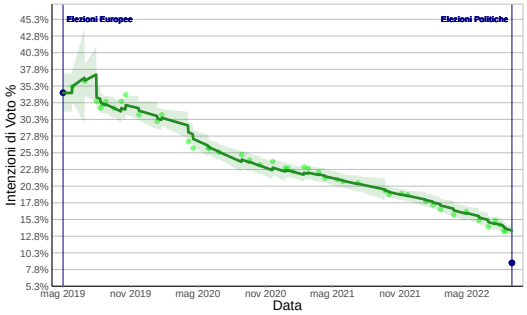
Demopolis - FI - Intervallo 2



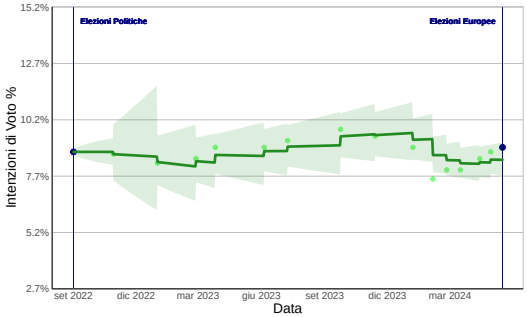
Demopolis - FI - Intervallo 3



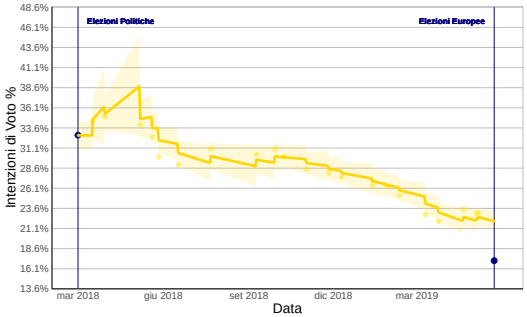
Demopolis - Lega - Intervallo 1



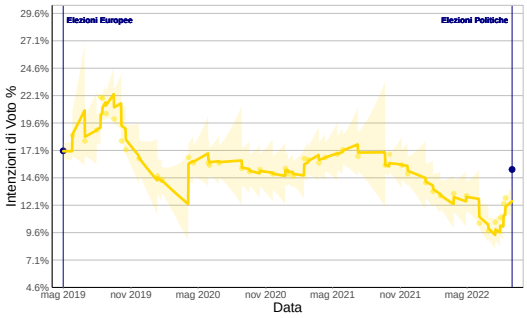
Demopolis - Lega - Intervallo 2



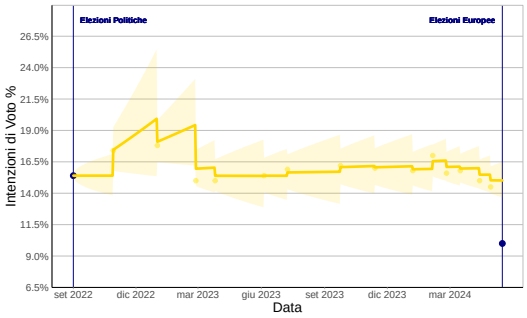
Demopolis - Lega - Intervallo 3



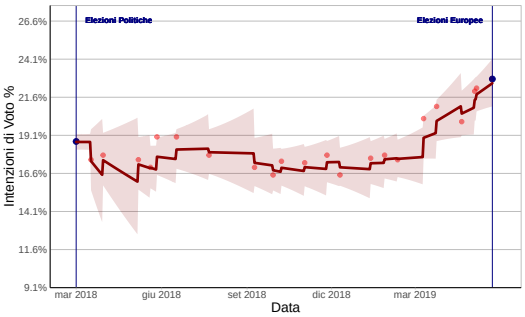
Demopolis - M5S - Intervallo 1



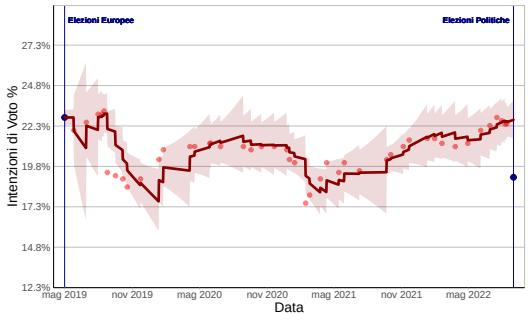
Demopolis - M5S - Intervallo 2



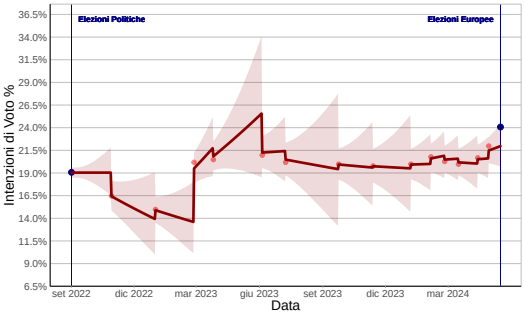
Demopolis - M5S - Intervallo 3



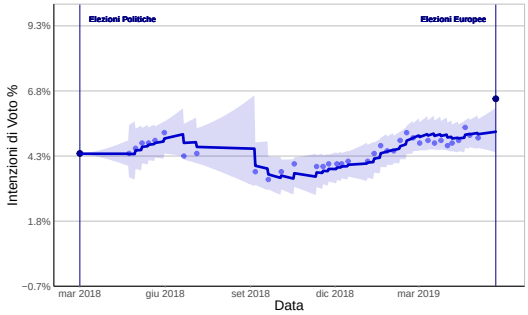
Demopolis - PD - Intervallo 1



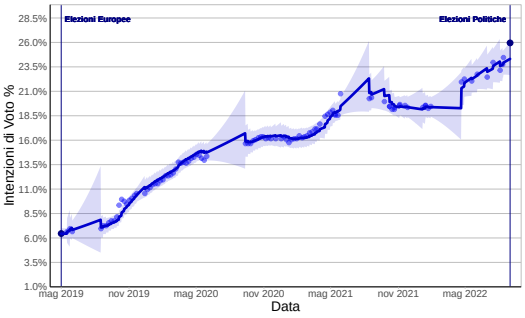
Demopolis - PD - Intervallo 2



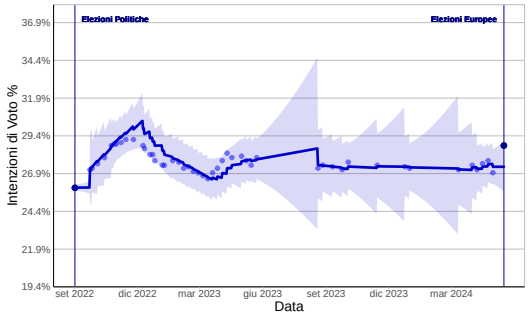
Demopolis - PD - Intervallo 3



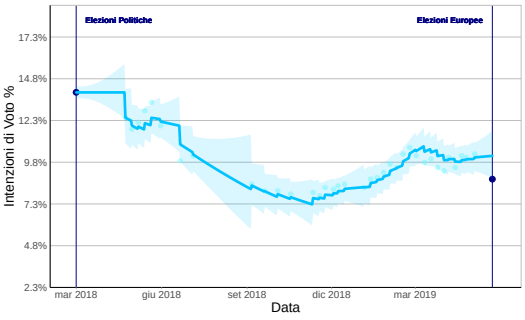
EMG - FDI - Intervallo 1



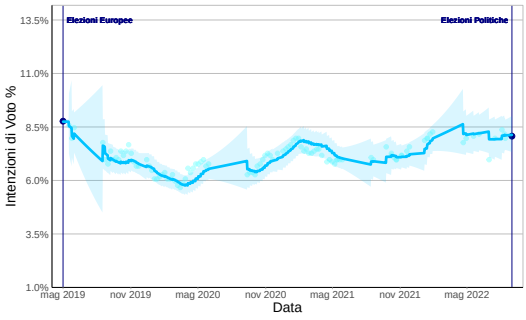
EMG - FDI - Intervallo 2



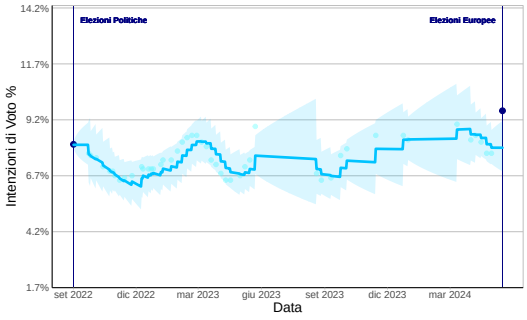
EMG - FDI - Intervallo 3



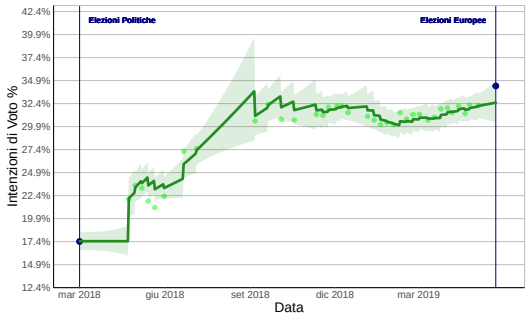
EMG - FI - Intervallo 1



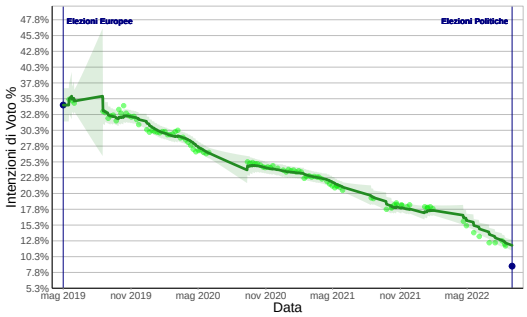
EMG - FI - Intervallo 2



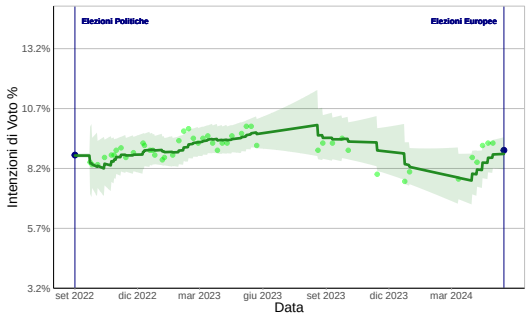
EMG - FI - Intervallo 3



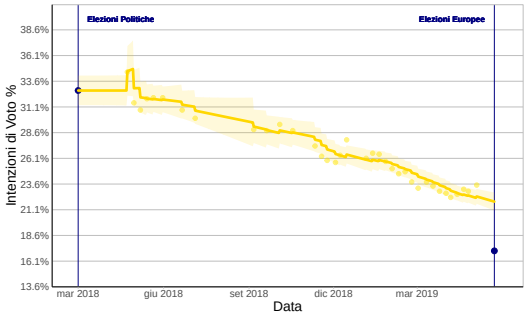
EMG - Lega - Intervallo 1



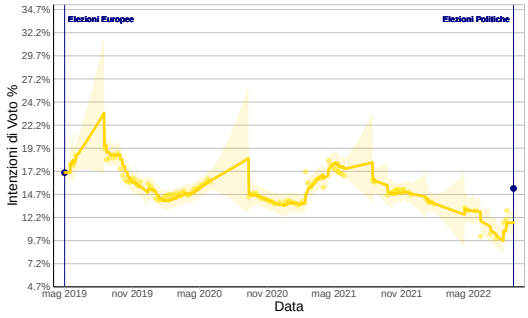
EMG - Lega - Intervallo 2



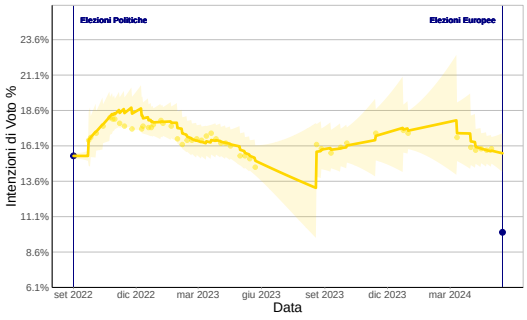
EMG - Lega - Intervallo 3



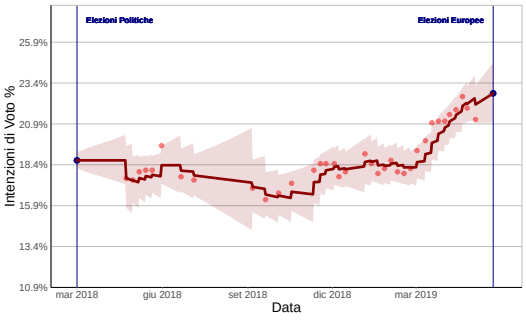
EMG - M5S - Intervallo 1



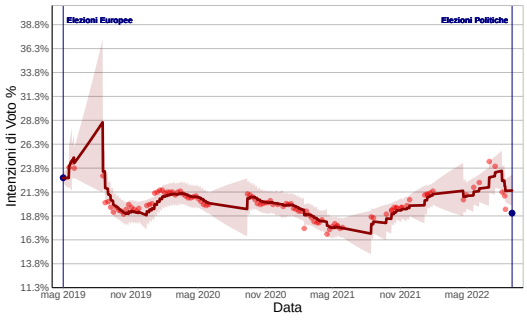
EMG - M5S - Intervallo 2



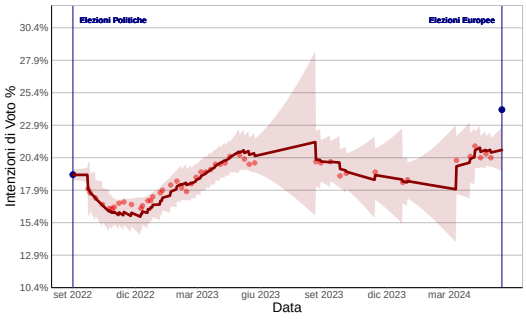
EMG - M5S - Intervallo 3



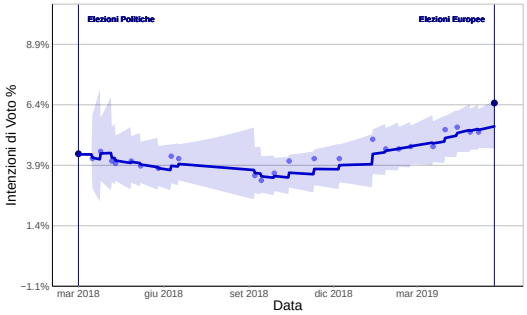
EMG - PD - Intervallo 1



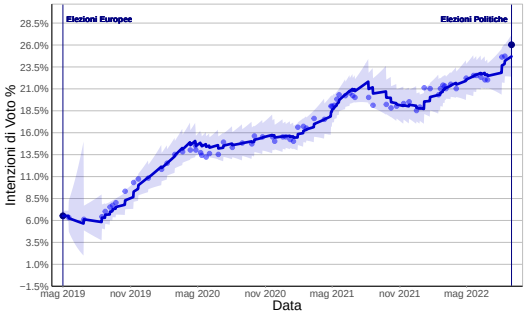
EMG - PD - Intervallo 2



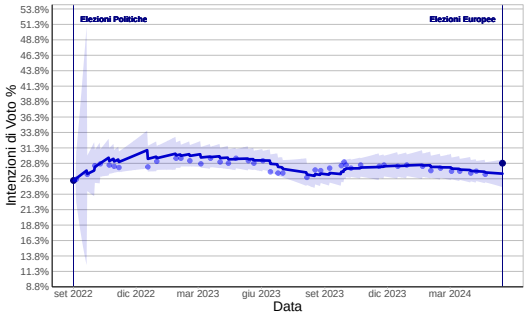
EMG - PD - Intervallo 3



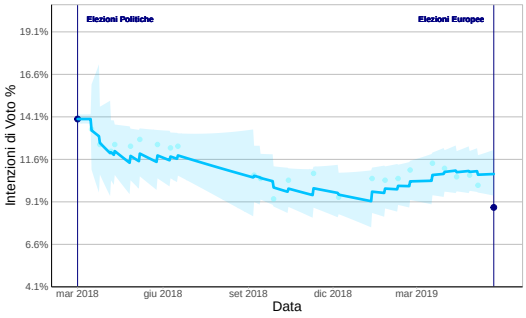
Euromedia - FDI - Intervallo 1



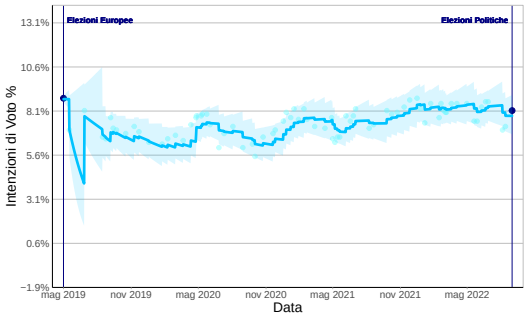
Euromedia - FDI - Intervallo 2



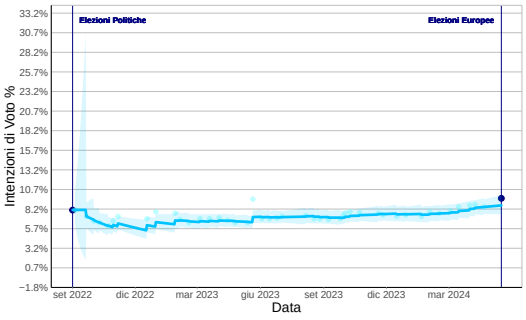
Euromedia - FDI - Intervallo 3



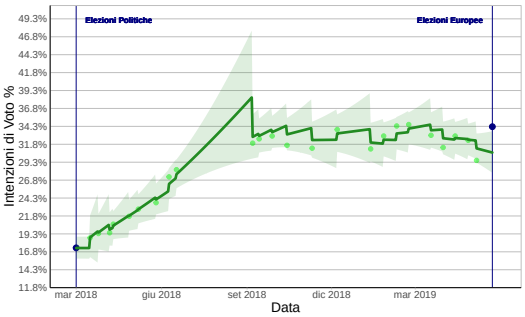
Euromedia - FI - Intervallo 1



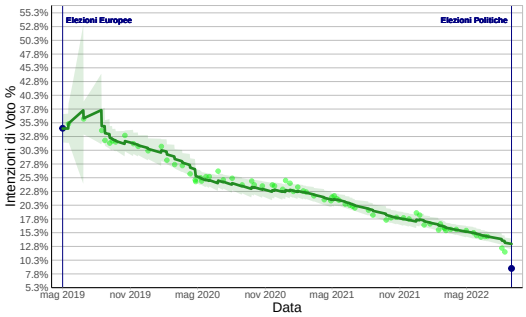
Euromedia - FI - Intervallo 2



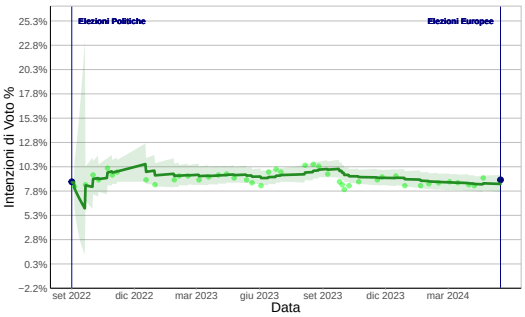
Euromedia - FI - Intervallo 3



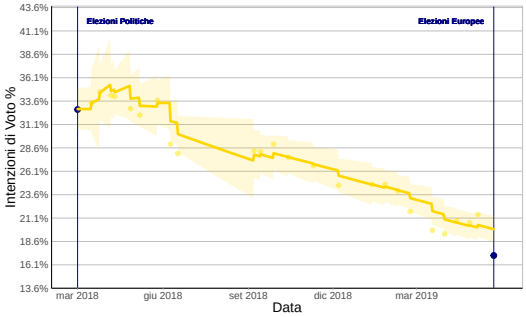
Euromedia - Lega - Intervallo 1



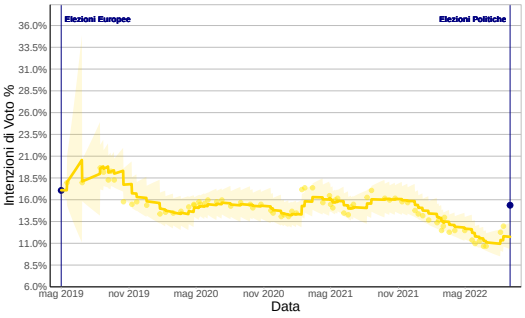
Euromedia - Lega - Intervallo 2



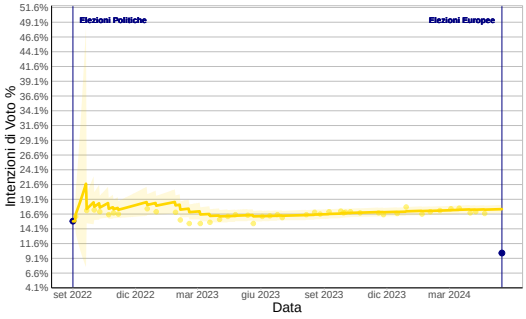
Euromedia - Lega - Intervallo 3



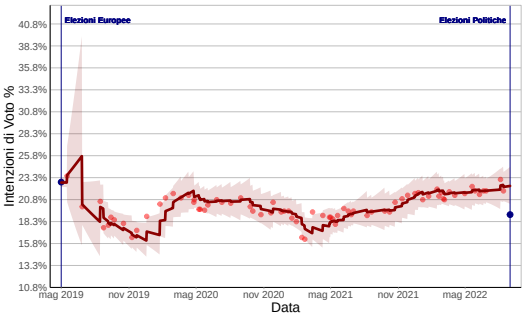
Euromedia - M5S - Intervallo 1



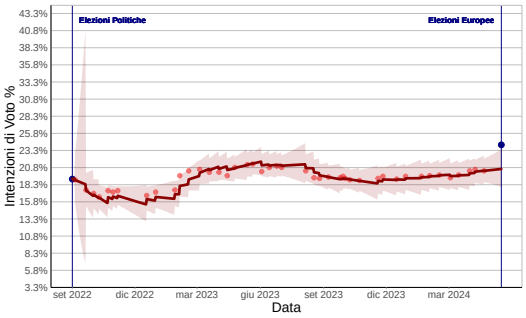
Euromedia - M5S - Intervallo 2



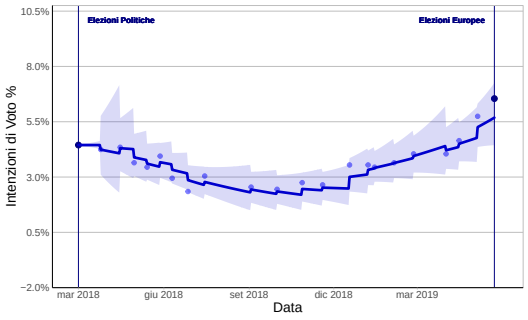
Euromedia - M5S - Intervallo 3



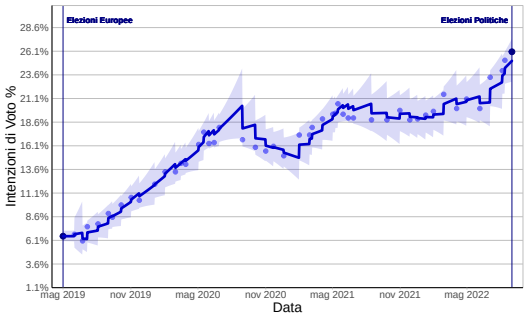
Euromedia - PD - Intervallo 2



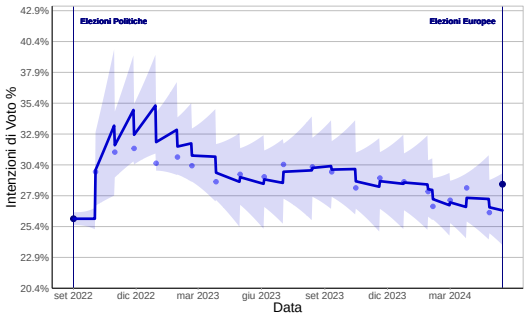
Euromedia - PD - Intervallo 3



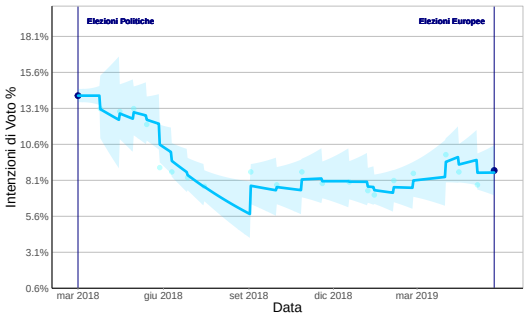
Ipsos - FDI - Intervallo 1



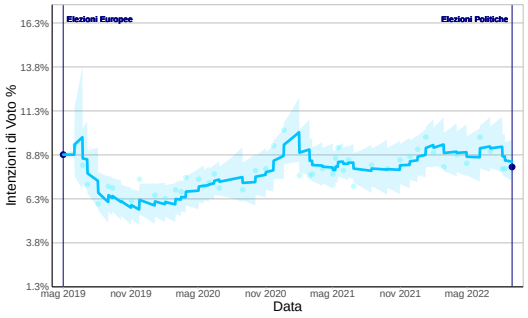
Ipsos - FDI - Intervallo 2



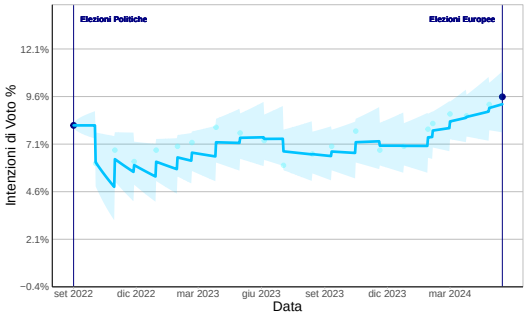
Ipsos - FDI - Intervallo 3



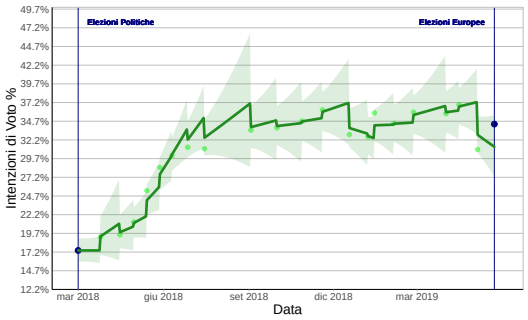
Ipsos - FI - Intervallo 1



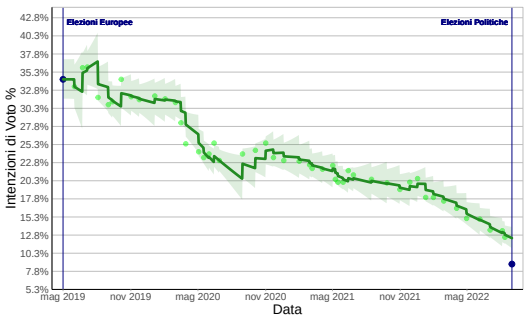
Ipsos - FI - Intervallo 2



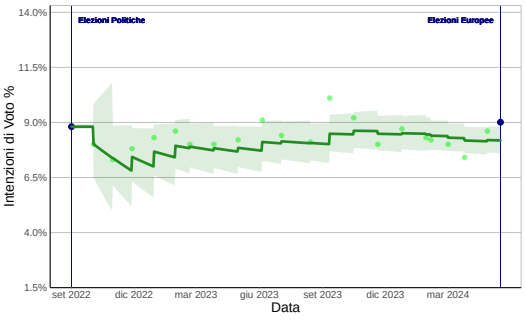
Ipsos - FI - Intervallo 3



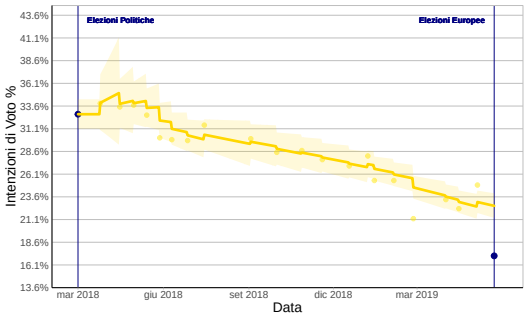
Ipsos - Lega - Intervallo 1



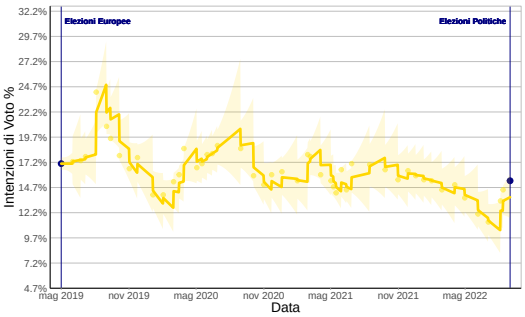
Ipsos - Lega - Intervallo 2



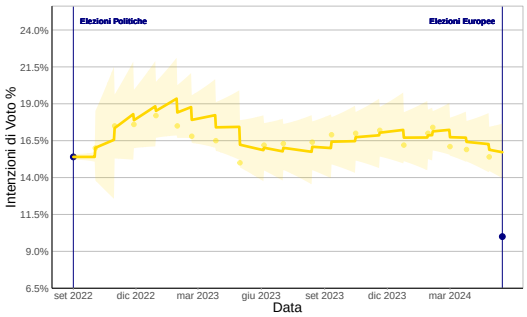
Ipsos - Lega - Intervallo 3



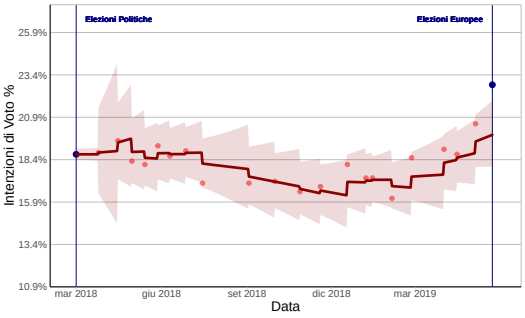
Ipsos - M5S - Intervallo 1



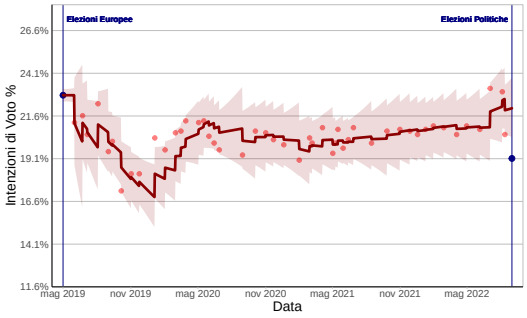
Ipsos - M5S - Intervallo 2



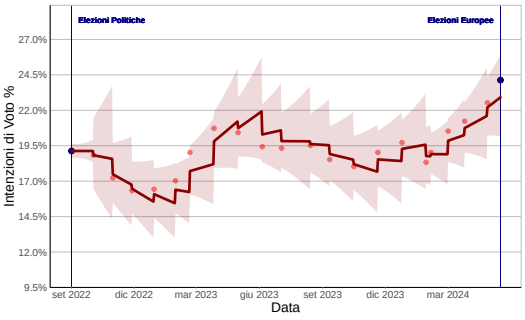
Ipsos - M5S - Intervallo 3



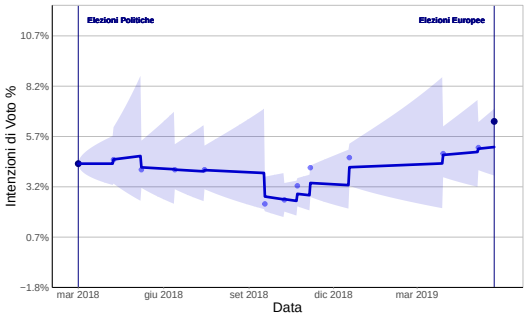
Ipsos - PD - Intervallo 1



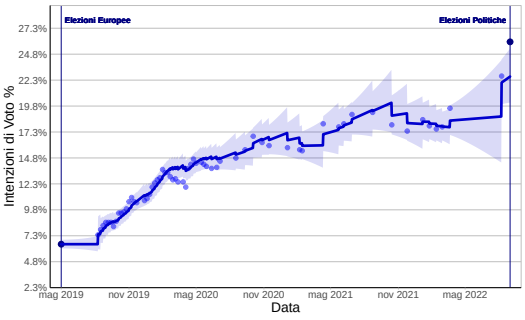
Ipsos - PD - Intervallo 2



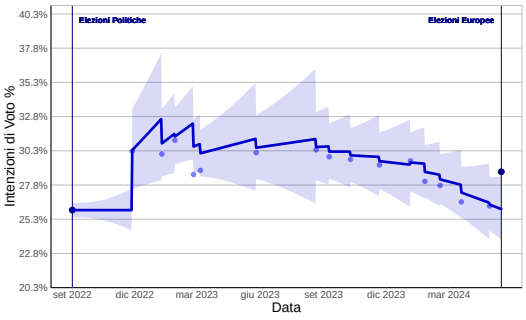
Ipsos - PD - Intervallo 3



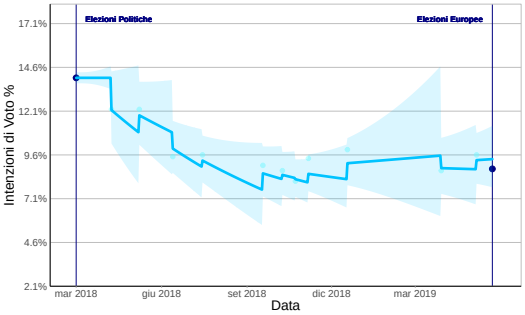
Ixè - FDI - Intervallo 1



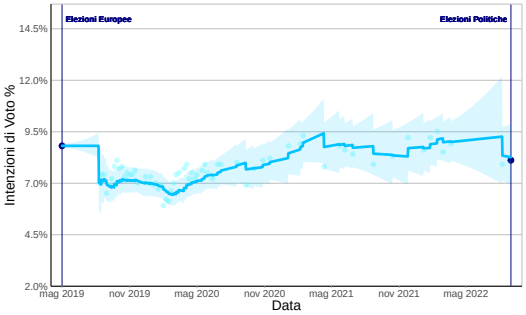
Ixè - FDI - Intervallo 2



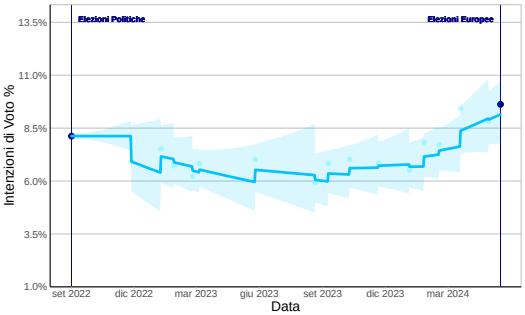
Ixè - FDI - Intervallo 3



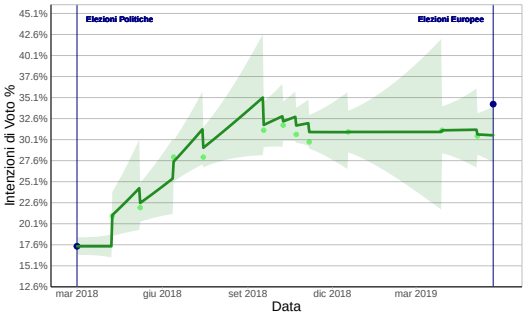
Ixè - FI - Intervallo 1



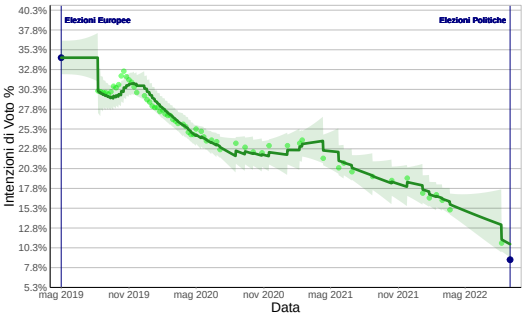
Ixè - FI - Intervallo 2



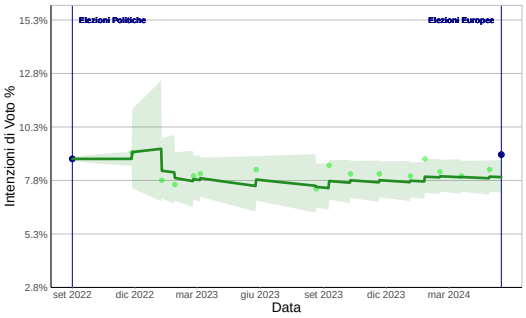
Ixè - FI - Intervallo 3



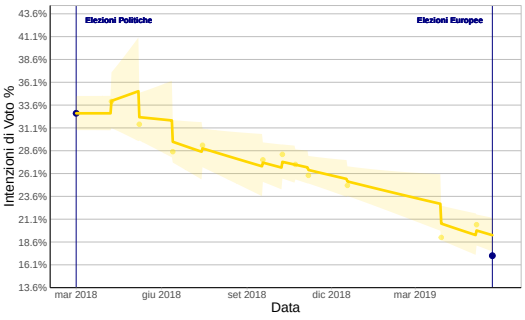
Ixè - Lega - Intervallo 1



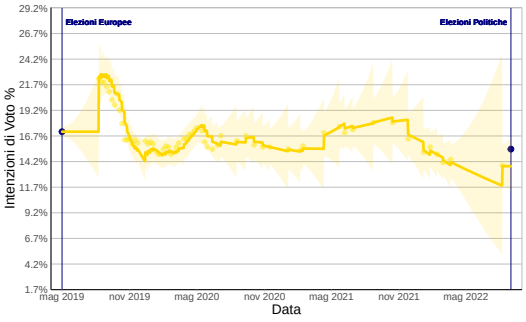
Ixè - Lega - Intervallo 2



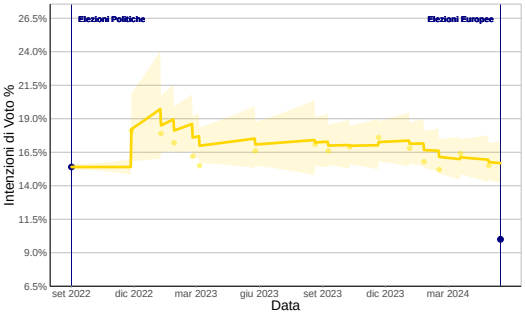
Ixè - Lega - Intervallo 3



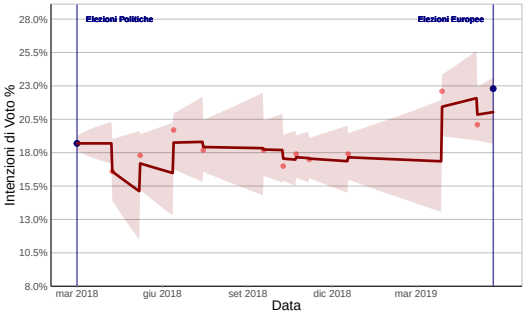
Ixè - M5S - Intervallo 1



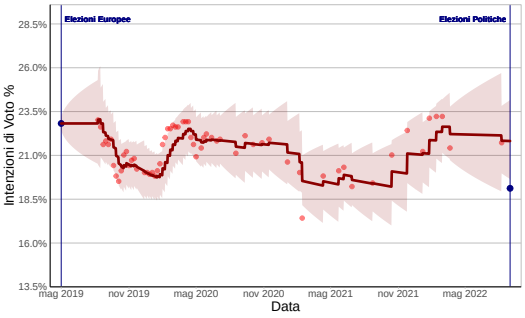
Ixè - M5S - Intervallo 2



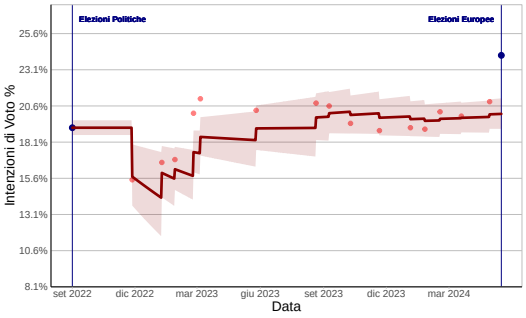
Ixè - M5S - Intervallo 3



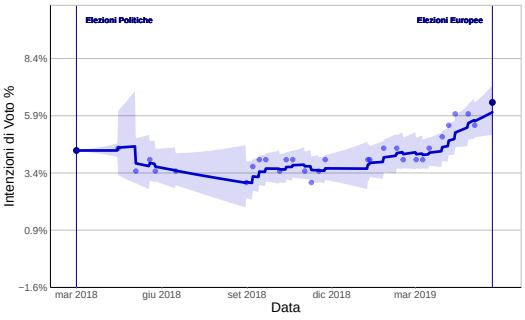
Ixè - PD - Intervallo 1



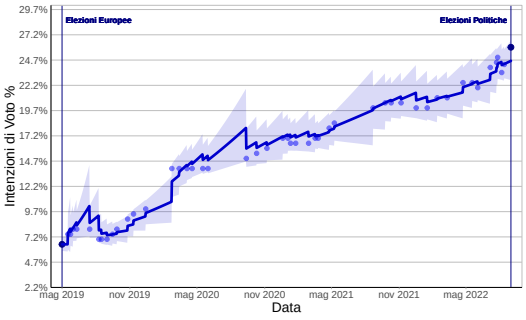
Ixè - PD - Intervallo 2



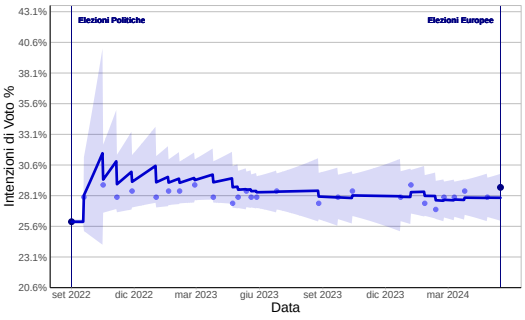
Ixè - PD - Intervallo 3



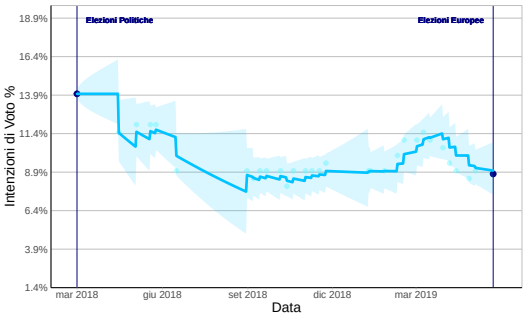
Noto - FDI - Intervallo 1



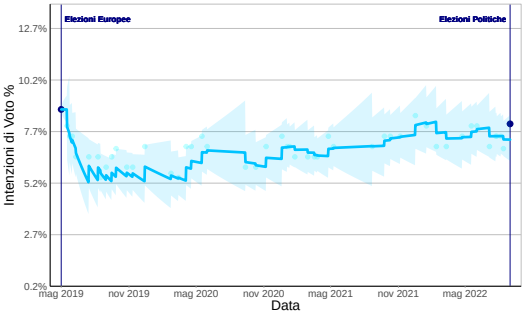
Noto - FDI - Intervallo 2



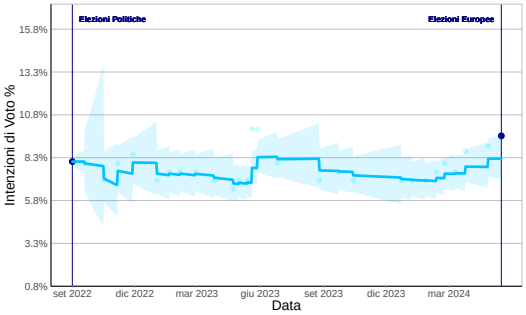
Noto - FdI - Intervallo 3



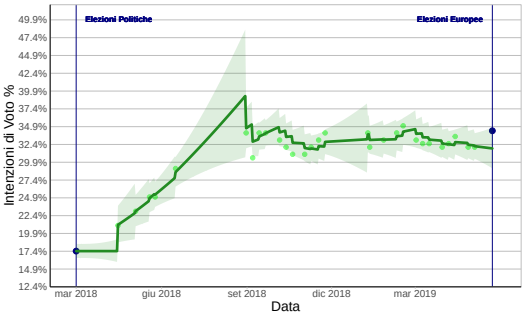
Noto - FI - Intervallo 1



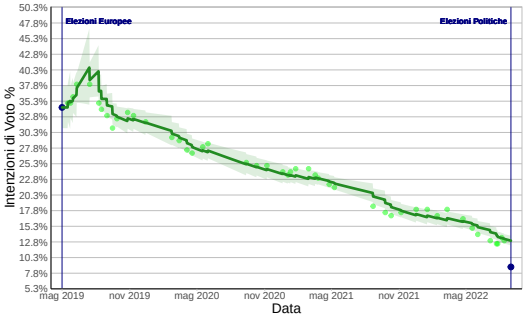
Noto - FI - Intervallo 2



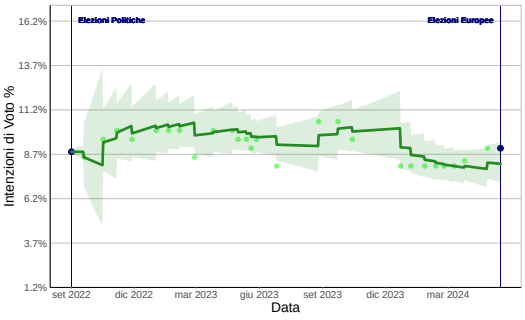
Noto - FI - Intervallo 3



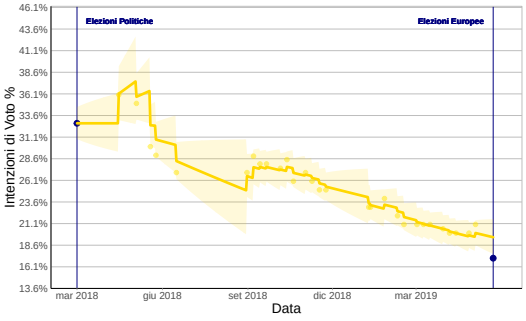
Noto - Lega - Intervallo 1



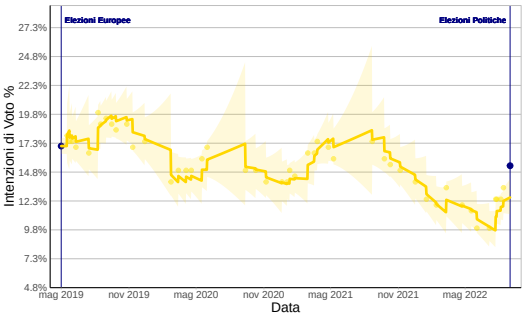
Noto - Lega - Intervallo 2



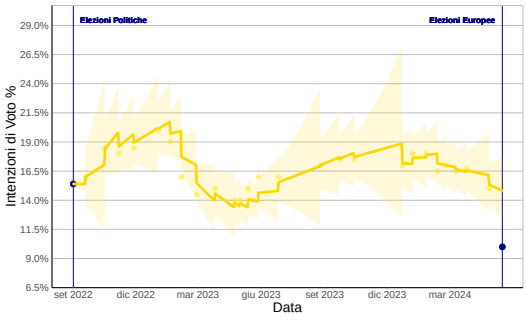
Noto - Lega - Intervallo 3



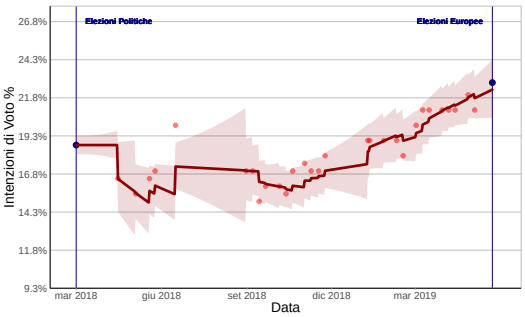
Noto - M5S - Intervallo 1



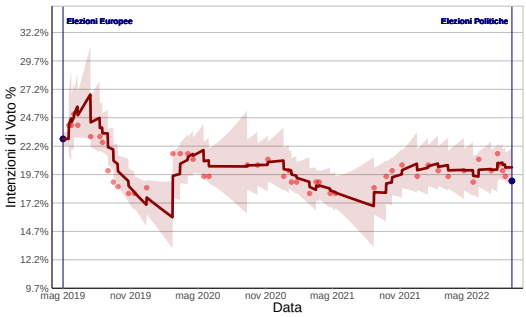
Noto - M5S - Intervallo 2



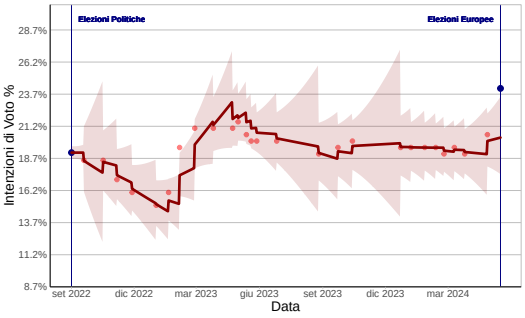
Noto - M5S - Intervallo 3



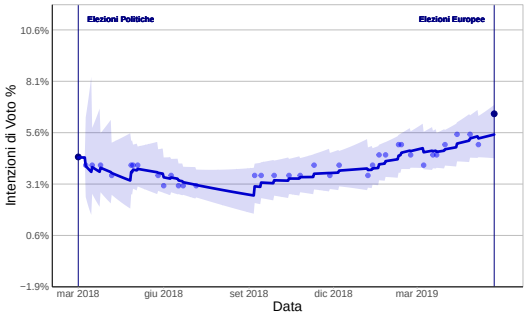
Noto - PD - Intervallo 1



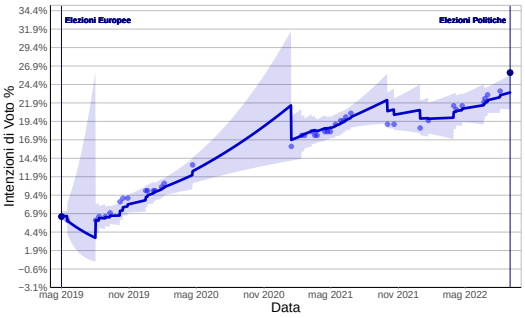
Noto - PD - Intervallo 2



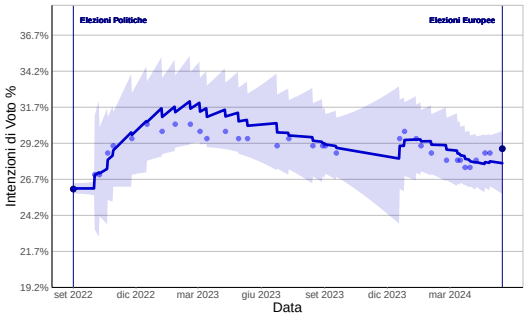
Noto - PD - Intervallo 3



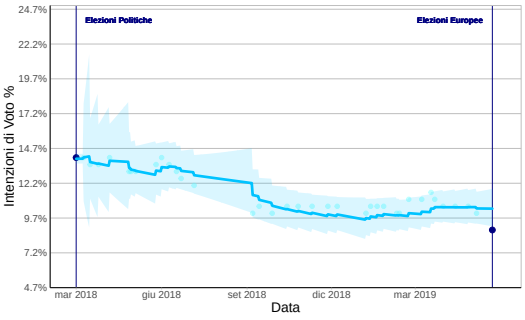
Piepoli - FDI - Intervallo 1



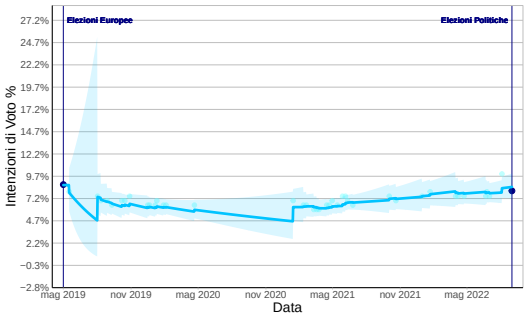
Piepoli - FDI - Intervallo 2



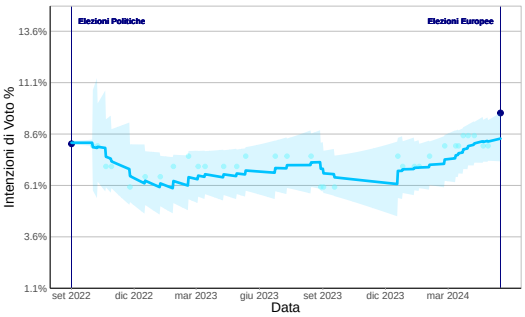
Piepoli - FDI - Intervallo 3



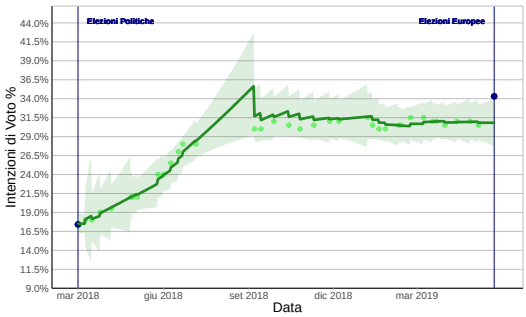
Piepoli - FI - Intervallo 1



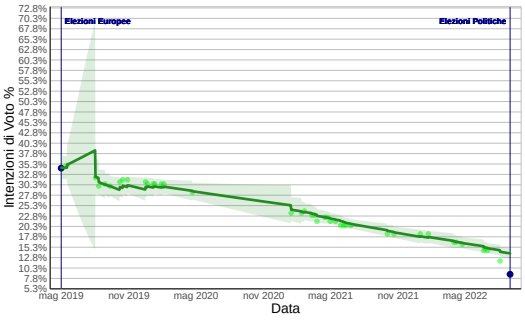
Piepoli - FI - Intervallo 2



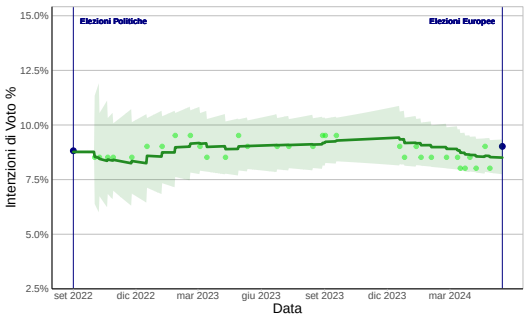
Piepoli - FI - Intervallo 3



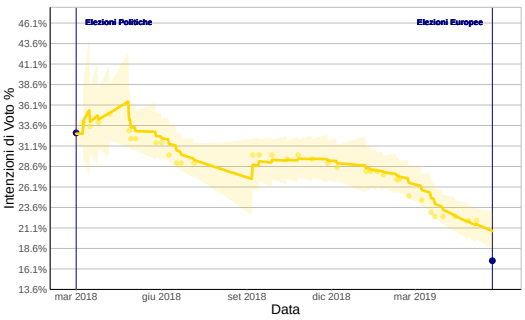
Piepoli - Lega - Intervallo 1



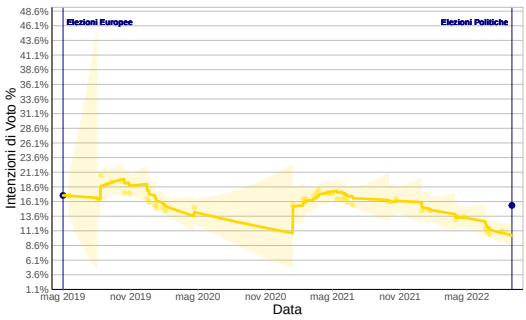
Piepoli - Lega - Intervallo 2



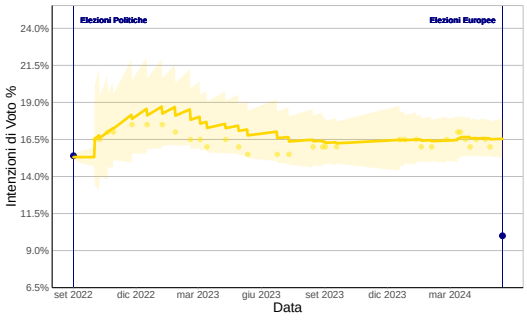
Piepoli - Lega - Intervallo 3



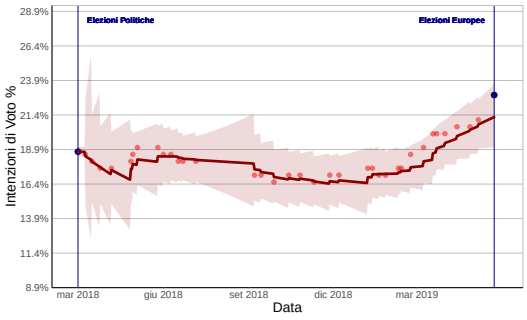
Piepoli - M5S - Intervallo 1



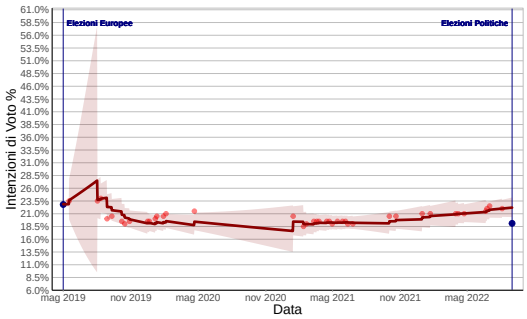
Piepoli - M5S - Intervallo 2



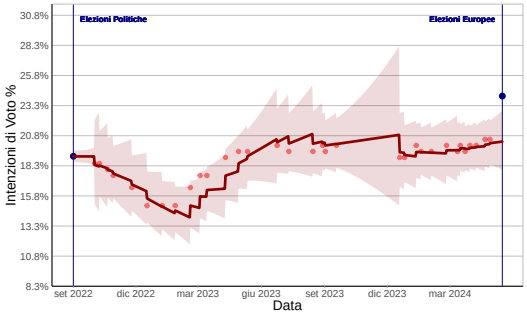
Piepoli - M5S - Intervallo 3



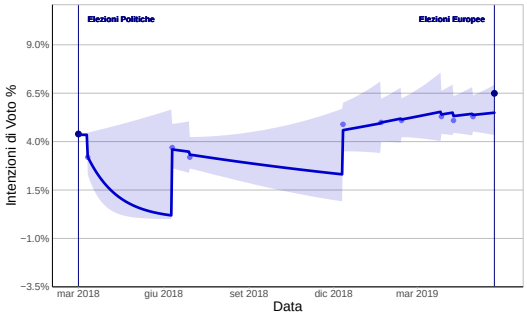
Piepoli - PD - Intervallo 1



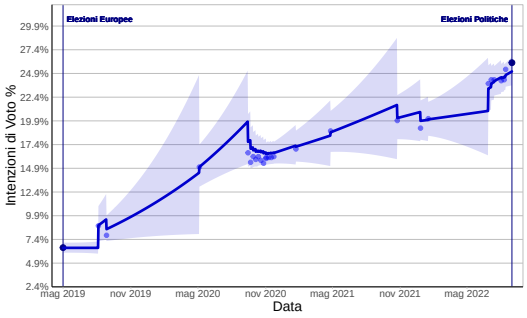
Piepoli - PD - Intervallo 2



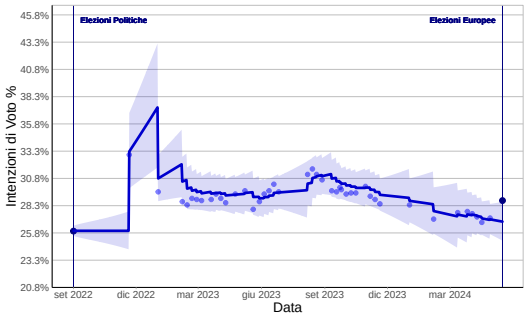
Piepoli - PD - Intervallo 3



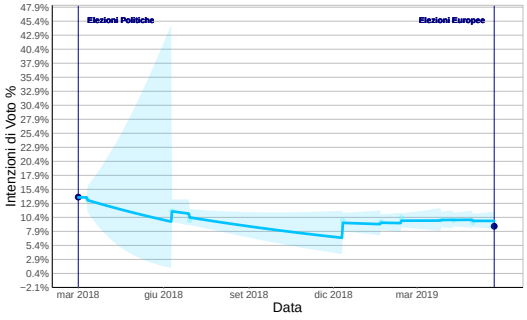
Quorum - FDI - Intervallo 1



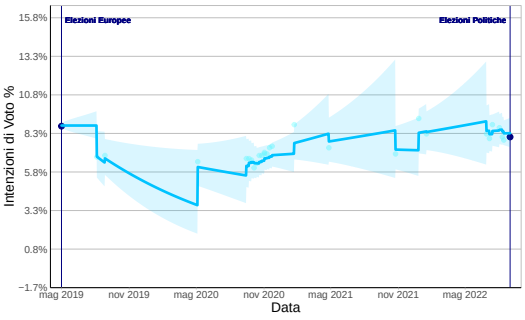
Quorum - FDI - Intervallo 2



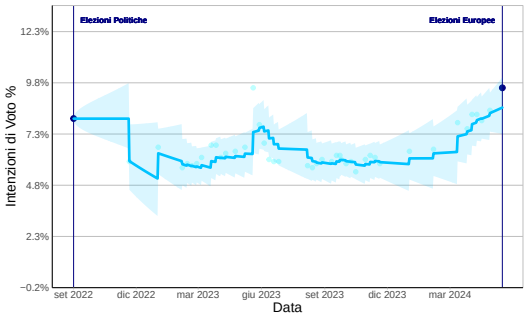
Quorum - FDI - Intervallo 3



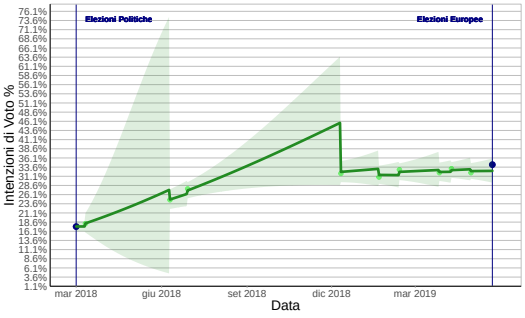
Quorum - FI - Intervallo 1



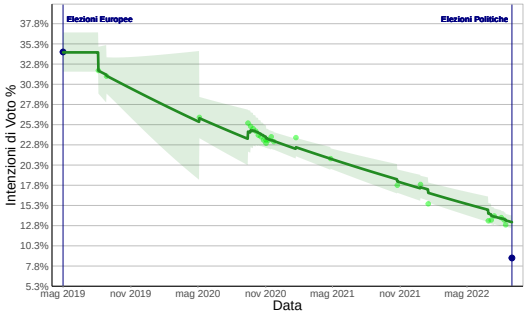
Quorum - FI - Intervallo 2



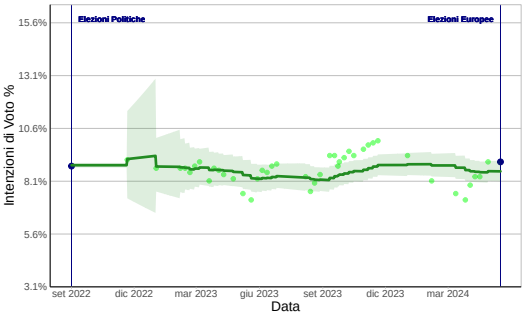
Quorum - FI - Intervallo 3



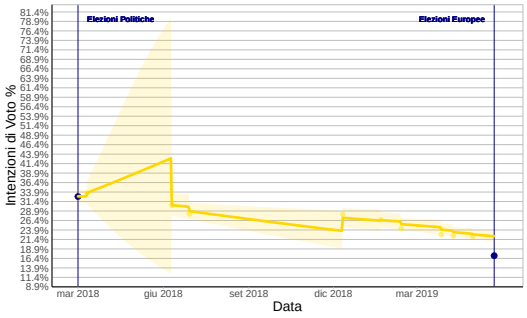
Quorum - Lega - Intervallo 1



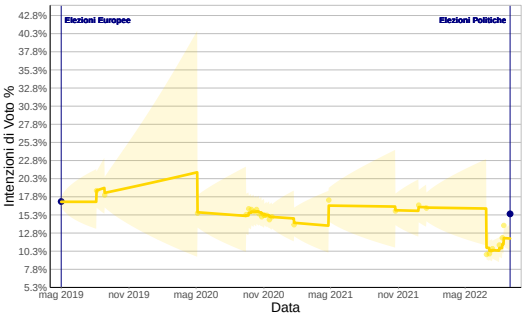
Quorum - Lega - Intervallo 2



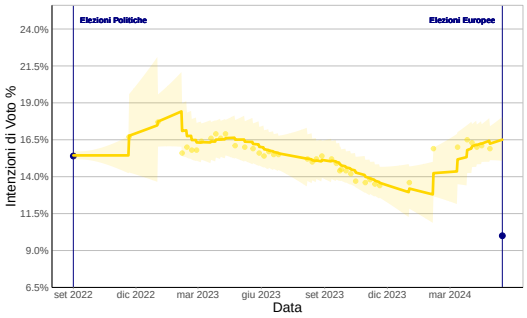
Quorum - Lega - Intervallo 3



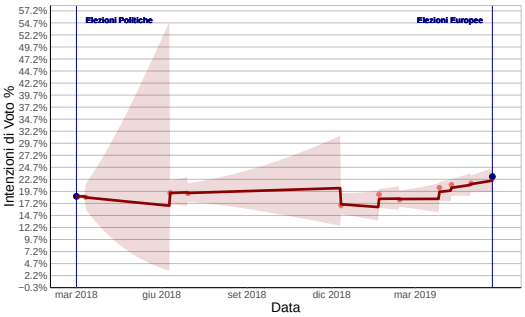
Quorum - M5S - Intervallo 1



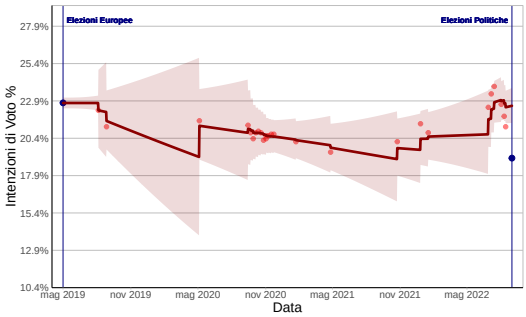
Quorum - M5S - Intervallo 2



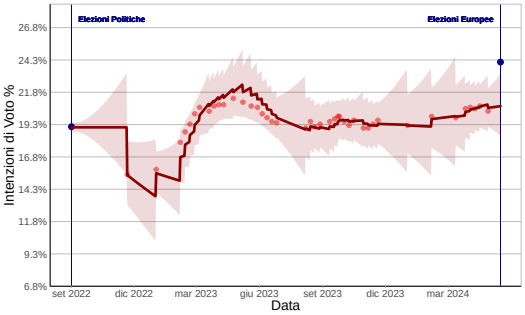
Quorum - M5S - Intervallo 3



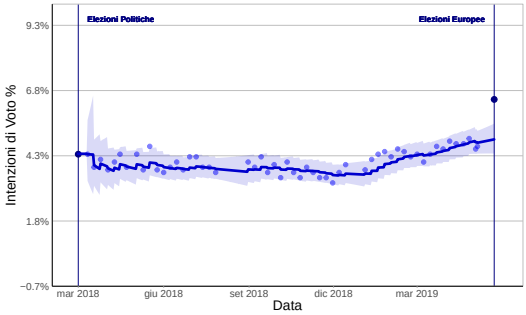
Quorum - PD - Intervallo 1



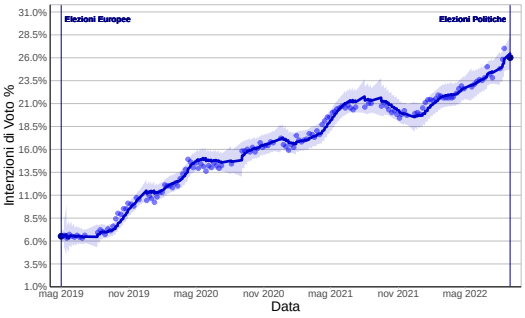
Quorum - PD - Intervallo 2



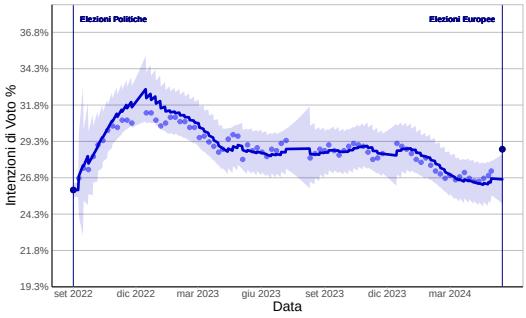
Quorum - PD - Intervallo 3



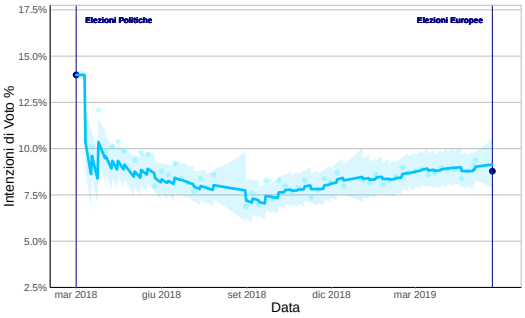
SWG - FDI - Intervallo 1



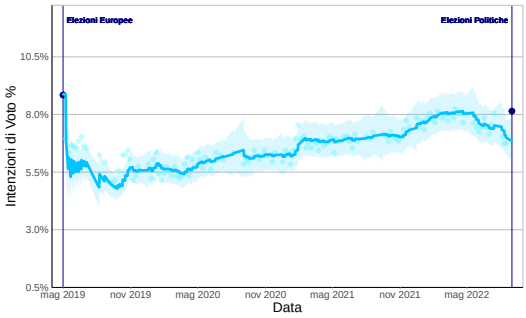
SWG - FDI - Intervallo 2



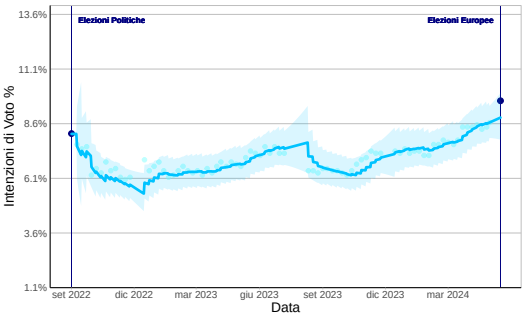
SWG - FDI - Intervallo 3



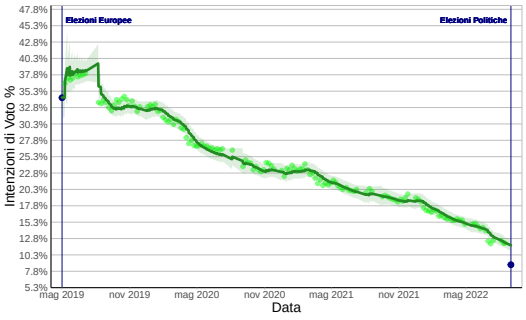
SWG - FI - Intervallo 1



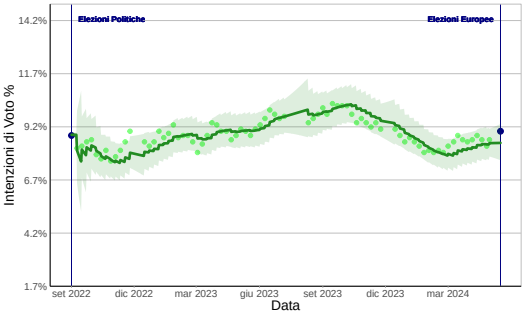
SWG - FI - Intervallo 2



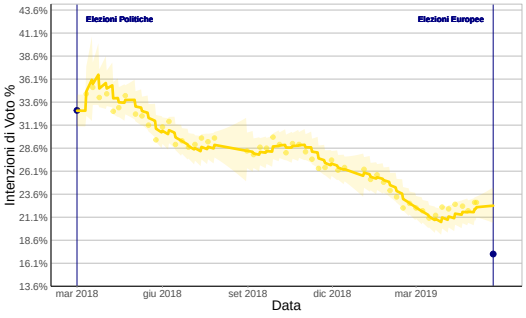
SWG - FI - Intervallo 3



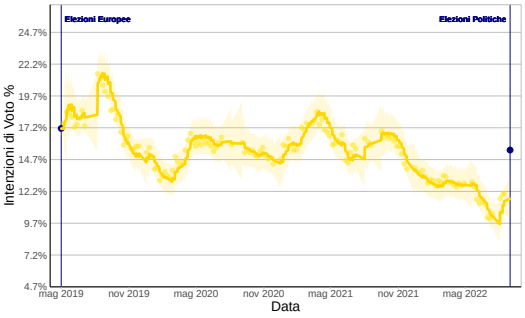
SWG - Lega - Intervallo 2



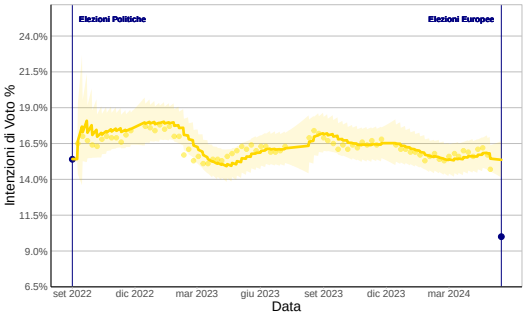
SWG - Lega - Intervallo 3



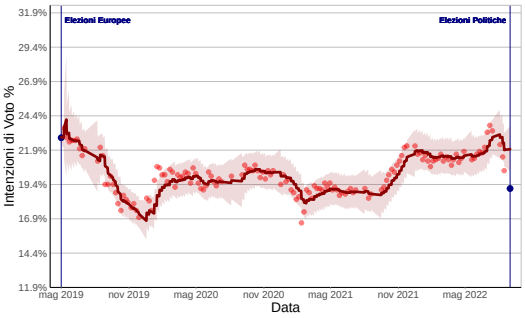
SWG - M5S - Intervallo 1



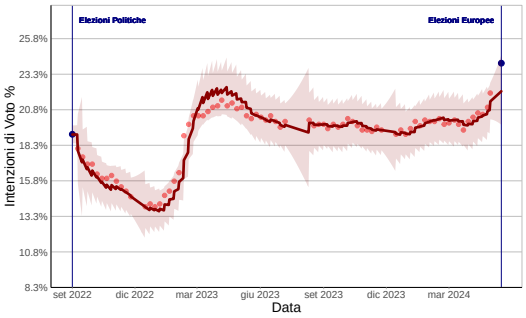
SWG - M5S - Intervallo 2



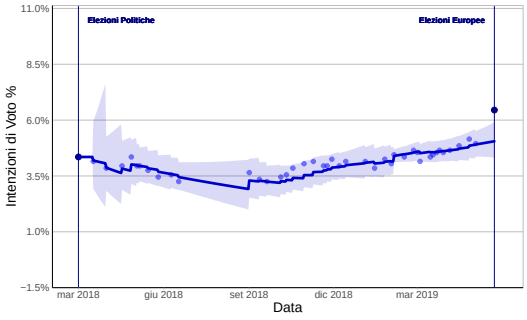
SWG - M5S - Intervallo 3



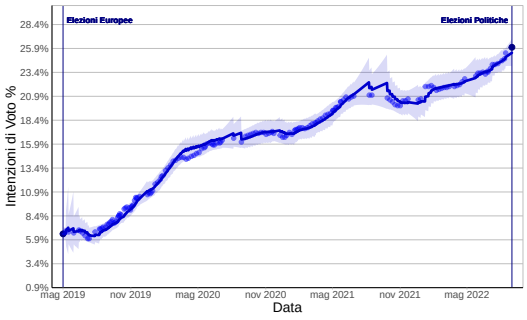
SWG - PD - Intervallo 2



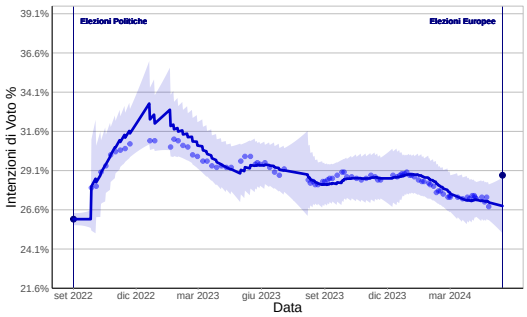
SWG - PD - Intervallo 3



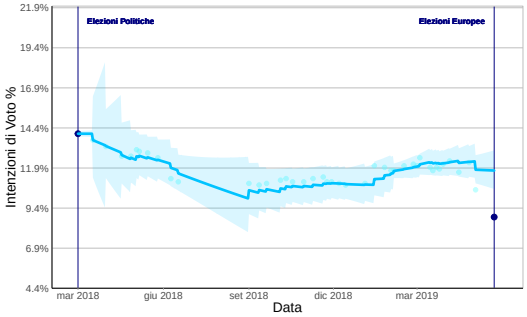
Tecne - FDI - Intervallo 1



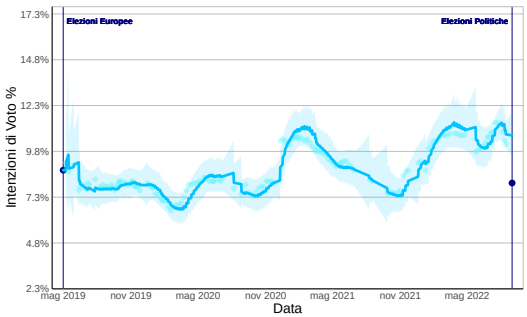
Tecne - FDI - Intervallo 2



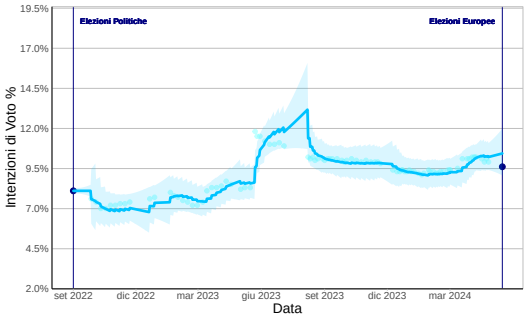
Tecne - FDI - Intervallo 3



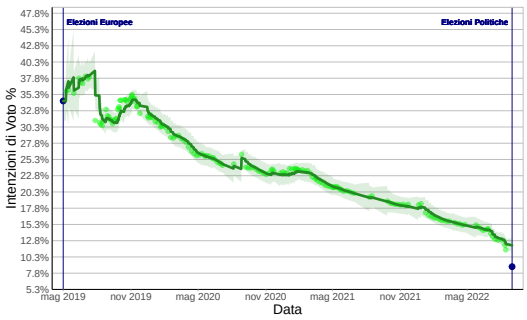
Tecne - FI - Intervallo 1



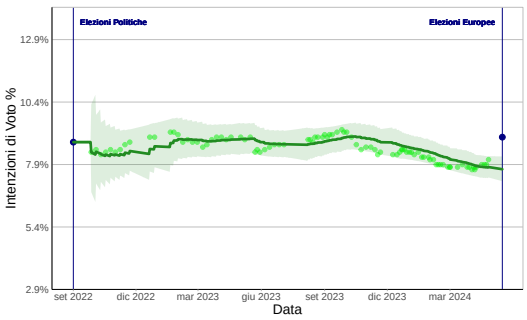
Tecne - FI - Intervallo 2



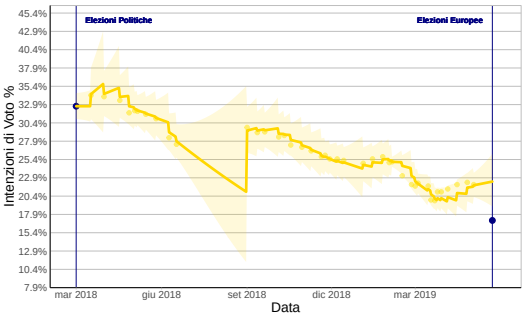
Tecne - FI - Intervallo 3



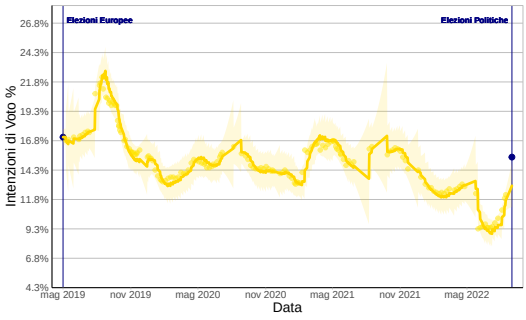
Tecne - Lega - Intervallo 2



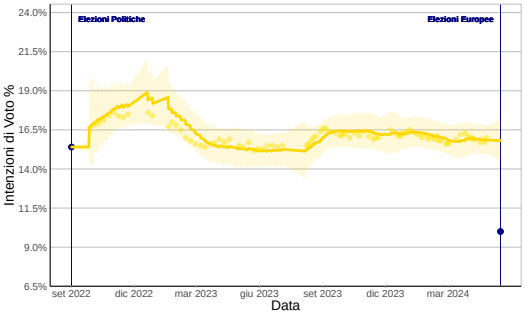
Tecne - Lega - Intervallo 3



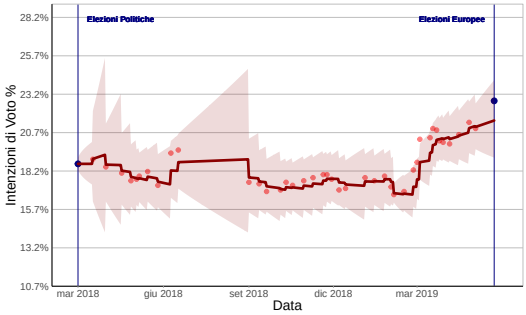
Tecne - M5S - Intervallo 1



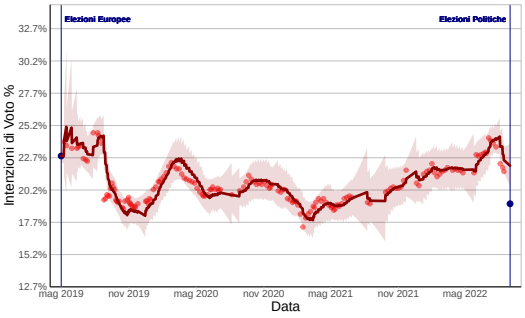
Tecne - M5S - Intervallo 2



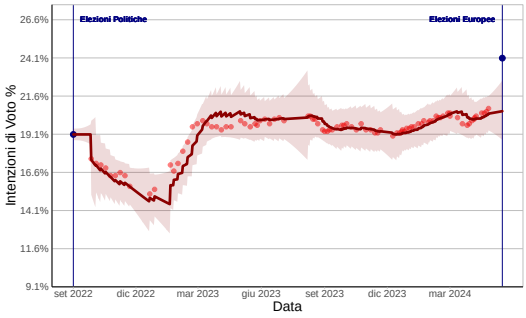
Tecne - M5S - Intervallo 3



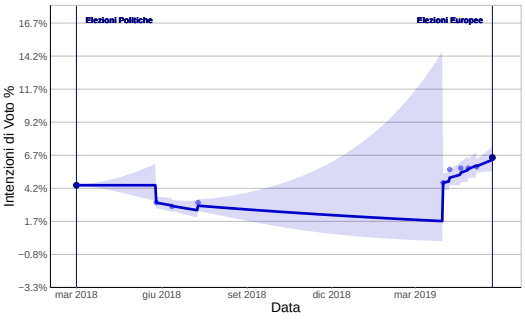
Tecne - PD - Intervallo 1



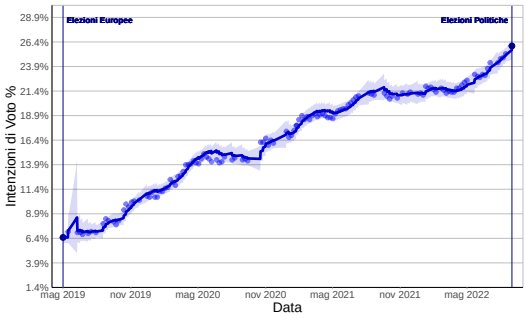
Tecne - PD - Intervallo 2



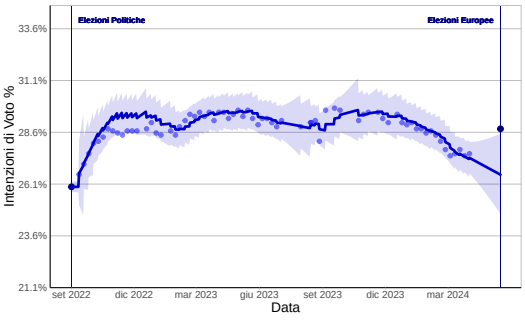
Tecne - PD - Intervallo 3



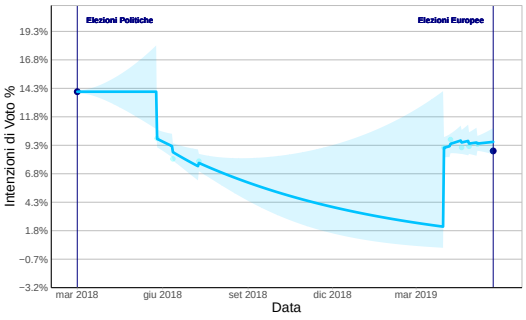
TP - FDI - Intervallo 1



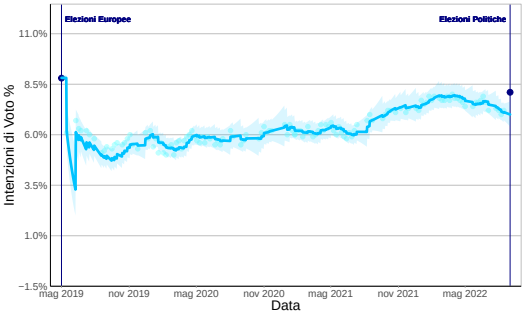
TP - FDI - Intervallo 2



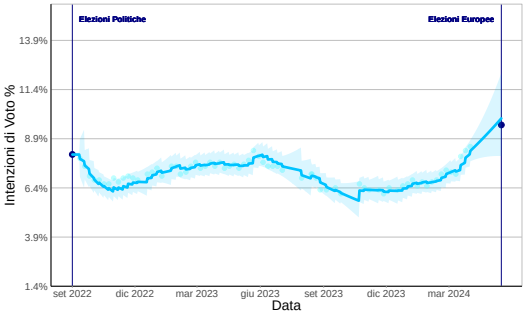
TP - FDI - Intervallo 3



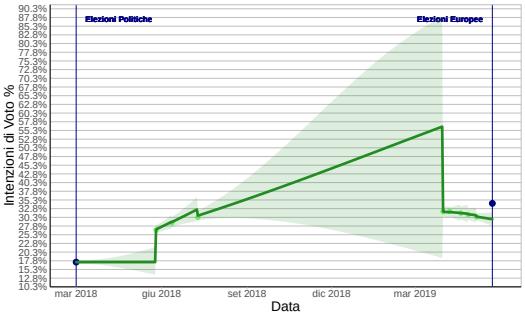
TP - FI - Intervallo 1



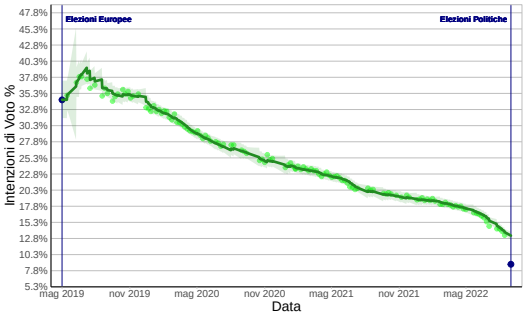
TP - FI - Intervallo 2



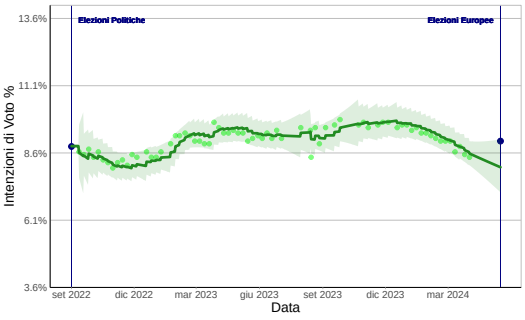
TP - FI - Intervallo 3



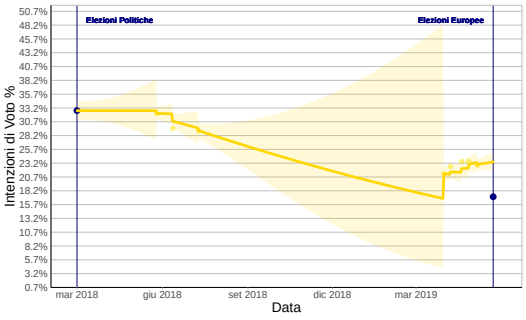
TP - Lega - Intervallo 1



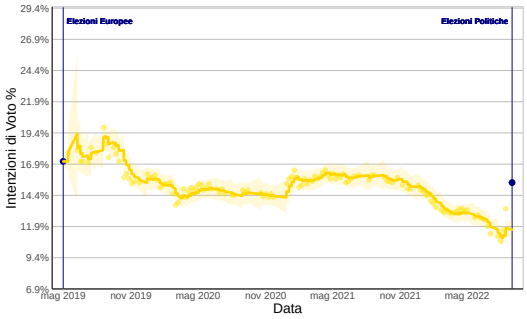
TP - Lega - Intervallo 2



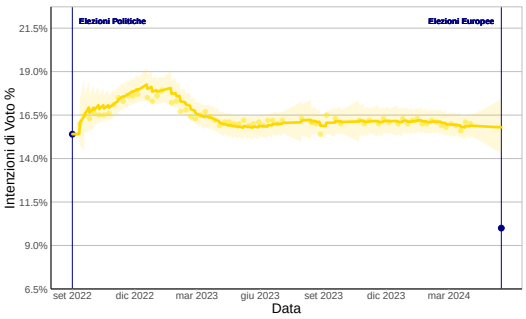
TP - Lega - Intervallo 3



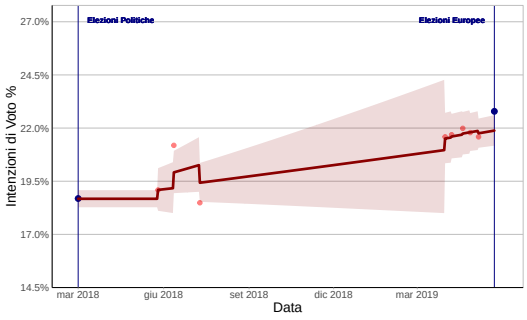
TP - M5S - Intervallo 1



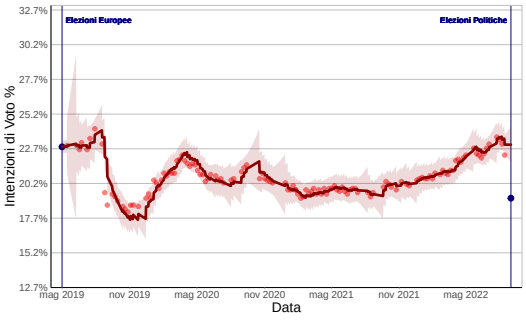
TP - M5S - Intervallo 2



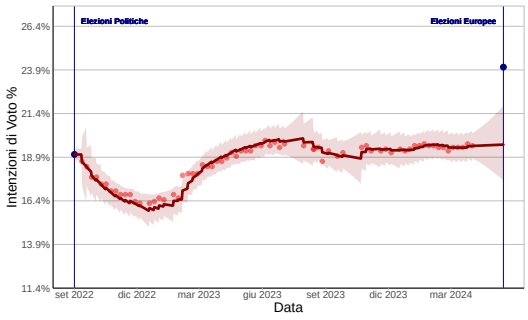
TP - M5S - Intervallo 3



TP - PD - Intervallo 1



TP - PD - Intervallo 2



TP - PD - Intervallo 3

Bibliografia

Bunker, K. (2025). Electoral forecasting in volatile party system settings: Assessing and improving pre-election poll predictions in italy. *Social Science Computer Review* **0**.

Cribari-Neto, F. & Silva, W. (2011). A new heteroskedasticity-consistent covariance matrix estimator for the linear regression model. *AStA Advances in Statistical Analysis* **95**, 129–146.

Durbin, J. & Koopman, S. (2012). *Time Series Analysis by State Space Methods*. Oxford University Press.

Fox, J. & Weisberg, S. (2019). *An R Companion to Applied Regression*. Thousand Oaks CA: Sage, 3rd ed.

Gelman, A. (2021). Failure and success in political polling and election forecasting. *Statistics and Public Policy* **8**, 1–9.

Gelman, A. & Azari, J. (2017). 19 things we learned from the 2016 election. *Statistics and Public Policy* **4**, 1–10.

Gelman, A., Goel, S., Rivers, D. & Rothschild, D. (2016). The mythical swing voter. *Quarterly Journal of Political Science* **11**, 103–130.

Gelman, A. & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press.

Heidemanns, M., Gelman, A. & Morris, G. E. (2020). An Updated Dynamic Bayesian Forecasting Model for the U.S. Presidential Election. *Harvard Data Science Review* **2**. Https://hdsr.mitpress.mit.edu/pub/nw1dzd02.

Helske, J. (2017). Kfas: Exponential family state space models in r. *Journal of Statistical Software* **78**, 1–39.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* **82**, 35–45.

Long, J. S. & Ervin, L. H. (2000). Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician* **54**, 217–224.

Piccolo, D. (2010). *Statistica*. Strumenti / il Mulino. Il Mulino.

Shirani-Mehr, H., Rothschild, D., Goel, S. & Gelman, A. (2018). Disentangling bias and variance in election polls. *Journal of the American Statistical Association* **113**, 607–614.

Silver, N. (2014). Here's proof some pollsters are putting a thumb on the scale.

Sturgis, P., Baker, N., Callegaro, M., Fisher, S., Green, J., Jennings, W., Kuha, J., Lauderdale, B. & Smith, P. (2016). Report of the inquiry into the 2015 british general election opinion polls.

Vandenplas, C., Beullens, K., Loosveldt, G. & Stoop, I. (2018). Response rates in the european social survey: Increasing, decreasing, or a matter of fieldwork efforts? .

Voss, D. S., Gelman, A. & King, G. (1995). Pre-election survey methodology: Details from nine polling organizations, 1988 and 1992. *Public Opinion Quarterly* **59**, 98–132.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* **48**, 817–838.

Whiteley, P. (2016). Why did the polls get it wrong in the 2015 general election? evaluating the inquiry into pre-election polls. *The Political Quarterly* **87**, 437–442.